# Grounded and Transparent Response Generation for Conversational Information-Seeking Systems

by

**Weronika Łajewska**

Thesis submitted in fulfillment of
the requirements for the degree of

PHILOSOPHIAE DOCTOR
(PhD)

University of
Stavanger

Faculty of Science and Technology
Department of Electrical Engineering and Computer Science
2025

University of Stavanger
N-4036 Stavanger
NORWAY
www.uis.no

ISBN: 978-82-8439-381-0
ISSN: 1890-1387

PhD: Thesis UiS no. 868

# Abstract

Information access systems are increasingly shifting toward conversational interactions. Conversational information-seeking (CIS) systems have traditionally focused on passage retrieval, reranking, and query rewriting. However, synthesizing retrieved information into coherent, well-grounded responses remains a significant challenge. This thesis explores response generation in CIS, addressing key issues such as factual grounding, completeness, and transparency.

We investigate key aspects of response generation, including (1) synthesizing the requested information, (2) grounding it in specific facts identified in the passages, (3) estimating response completeness, and (4) revealing the system's limitations. To this end, we propose a modular response generation pipeline, which leverages retrieval-augmented generation and operates on fine-grained information nuggets—minimal, atomic units of relevant information. Our approach ensures factual correctness, facilitates source attribution, and improves response completeness while proactively suggesting follow-up questions.

To further improve response reliability, we introduce techniques for detecting system limitations, including identifying unanswerable questions to mitigate hallucinations. We also study how system transparency regarding the sources, system's confidence, and potential response limitations affects the user experience, demonstrating the importance of enhancing the response with additional information, thereby enabling its critical assessment by users.

By addressing these challenges, our research contributes to the advancement of CIS response generation, fostering more reliable, transparent, and user-centric interactions in information-seeking dialogues. By integrating mechanisms for unanswerable question detection, revealing response completeness, and explicitly communicating response limitations, we aim to mitigate misinformation, foster trust, and empower users to make informed judgments. Our findings provide valuable insights for future research on explainable CIS systems, paving the way for more transparent and effective user-system interactions.

# Acknowledgments

When I announced, 3.5 years ago, that I was leaving a good and stable job to start a PhD, people had many questions. Mostly, they wanted to understand what on earth I needed it for. When I told them I was moving to Norway, they were baffled—why head to the far north, where I'd supposedly live among bears and eat fish five times a day? And when I finally packed up and left, alone, in the middle of winter, they stopped asking questions altogether. I suspect they quietly concluded that something was just ... wrong with me.

Truthfully, I had plenty of doubts myself in those early months. But with every paper I wrote, every conference I attended, and every bold move I made, I became more certain that this was the best decision of my life. The past three years have given me more than I ever expected—intellectual growth, personal development, and a confidence I never knew to be within my reach. The people I've met, the places I've traveled to, and the challenges I've conquered have shaped me in ways I couldn't have imagined. Now, as I reach the finish line, I can say with certainty that I am leaving this journey as a person I was timidly hoping to be for a long time.

And for that, first and foremost, I want to express my deepest gratitude to my supervisor, Krisztian Balog. His guidance, expertise, and unwavering support were invaluable throughout this journey. I arrived in Stavanger as a student, and under his mentorship, I became an independent researcher. His high standards pushed me to aim for excellence, to think bigger, and to refine my scientific curiosity into impactful research. He taught me not just how to conduct research, but how to make it matter—to navigate the art of writing papers, presenting my work, and carving out my place in this competitive field. Thank you for your patience, feedback, and for always encouraging me to grow in so many different ways.

I am also incredibly grateful to my collaborators, especially Damiano Spina and Johanne Trippas from RMIT University in Melbourne, Australia. The three months I spent there were truly transformative—full of exciting research, new challenges, and meaningful connections. We achieved so much in such a short time, but beyond the work, I am most thankful for the way you welcomed me into your team from day one. Your kindness, support, and hospitality made my time in Australia an unforgettable adventure.

A huge thank you also goes to my colleagues at Amazon, particularly Momchil Hardalov, Laura Aina, and Lluís Màrquez from AWS Bedrock Guardrails team in Barcelona. Those five months gave me a firsthand look at the other side of the research world—applied science in industry. It was an eye-opening experience, full of fascinating discussions, intense brainstorming sessions, and, of course, spontaneous happy hours. Thank you to everyone at the Barcelona office for sharing your insights and making my internship such a rewarding experience.

My PhD journey would have been much lonelier (and certainly less fun) without my fellow PhD students at UiS. Sharing this rollercoaster ride with you all has been incredible. Here's to the countless cups of coffee at Bokkafeen, the celebrations, the commiserations, and, yes, the procrastination. Here's to the birthday parties, board game nights, hikes, barbecues, kayaking trips, gym sessions, cabin weekends, Irish pub nights, and everything in between. Thank you for being my support system, my partners in crime, and the best company I could ask for.

To my friends and family back in Poland—despite the distance, I have always felt your love and support. Thank you for the visits, the endless phone calls, and for listening to my rants about every PhD high and low. Thank you for always being there, no matter the miles, the weather, or the time difference.

I owe everything to my incredible parents. Your unwavering support, encouragement, and love have carried me through every challenge. Every success is sweeter because I get to share it with you. Every failure was easier to bear because I knew you were always in my corner. I cannot put into words how much I appreciate you.

And finally, to my wonderful life partner, Milan. You have been my anchor in the toughest moments and my guiding light when I lost my way. Thank you for standing by my side through every challenge, for sharing every victory, and for being part of every adventure. I couldn't have done this without you, and I can't wait for all the adventures still to come.

Research presented in this thesis was supported by the Norwegian Research Center for AI Innovation, NorwAI (Research Council of Norway, project number 309834). Work done in Łajewska *et al.* (2024b) and Łajewska *et al.* (2024a) was supported by the Australian Research Council (grant numbers DE200100064, CE200100005).

# Contents

# Chapter 1

# Introduction

*Science, my lad, is made up of mistakes, but they are mistakes which it is useful to make, because they lead little by little to the truth.*

— **Jules Verne**

Recently, many information access systems have transitioned to a conversational mode, enabling more dynamic and user-friendly interactions. Chatbots now provide real-time assistance for tasks like booking flights, managing purchases, or addressing banking queries. Voice assistants offer informed answers, personalized recommendations, and hands-free convenience while driving or when typing is not possible. Modern conversational assistants are able to handle complex interactions, support decision-making, answer complex questions, and execute specific requests. They can maintain context over extended exchanges, incorporate user feedback, proactively suggest actions, and adapt to ongoing interactions. With these advanced capabilities, conversational systems are becoming a preferred choice for many users and remain a very active area of research.

The shift toward conversational interactions is also transforming search systems, which are the focus of this thesis. Traditional search systems retrieve documents relevant to a user's query and present them as a ranked list (see Figure 1.1). Retrieval can also be performed at a more granular level, by identifying specific passages within documents. However, identifying the passages that are most relevant to a user's query in a document corpus is only a preliminary step; the ultimate goal of a *conversational search system* is to synthesize informa-

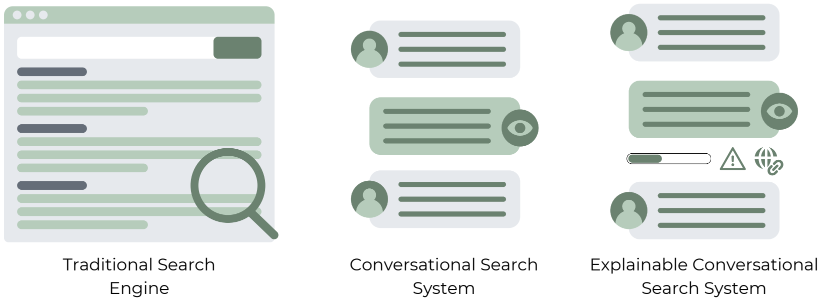

Traditional Search Engine
Conversational Search System
Explainable Conversational Search System

Figure 1.1: Different types of search systems including: (1) a traditional search engine that returns a ranked list of documents, (2) a conversational search system that generates a synthesized answer to the user's query, and (3) an explainable system that enhances conversational responses with additional information, such as source links, warnings about potential limitations, or estimations of response completeness.

tion from these passages into a coherent, informative response–a process known as *conversational response generation* (Ren *et al.*, 2021) (see Figure 1.2). The system's response must effectively distill the most pertinent information into a format that is both clear and easily digestible for the user (Culpepper *et al.*, 2018). In an ideal scenario, when the passages from the top of the ranking answer the question, the task of response generation simplifies to summarization (Owoicho *et al.*, 2022). However, in the realm of conversational information-seeking (CIS) dialogues, involving open-ended questions, indirect answers, and complex queries with partial answers spread across multiple sources (Bolotova-Baranova *et al.*, 2023; Zamani *et al.*, 2023; Gabburo *et al.*, 2024), the assumption that a user's query can be fully answered by summarizing top-retrieved information often falls short of reality.

While synthesizing retrieved information into conversational responses is crucial for enhancing the user experience (Culpepper *et al.*, 2018; Ren *et al.*, 2021), it presents challenges such as ensuring factual correctness (Ji *et al.*, 2023; Koopman and Zuccon, 2023; Tang *et al.*, 2023), source attribution (Rashkin *et al.*, 2021), information verifiability (Liu *et al.*, 2023a), consistency, salience, and coverage (Gienapp *et al.*, 2024). When the answer is not fully contained within top retrieved passages, summarizing them can lead to hallucinations (Tang *et al.*, 2023; Cao *et al.*, 2016; Ji *et al.*, 2023) or introduce bias by covering only one point of view or a partial answer (Gao and Shah, 2020). Consequently, relying solely on a summarization of the top retrieved passages risks providing users with biased, incomplete, or incorrect responses (Tang *et al.*, 2022). Generative language models, while widely adopted for response generation (Zhang *et al.*, 2020b; Lewis *et al.*, 2020), remain vulnerable to these limitations, further underscoring the need for more robust information synthesis methods in the context of answering complex queries.

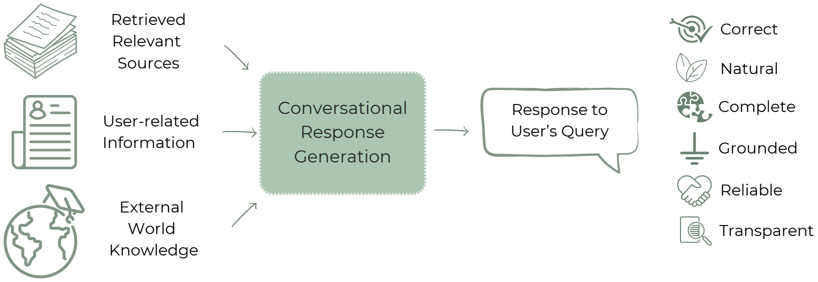
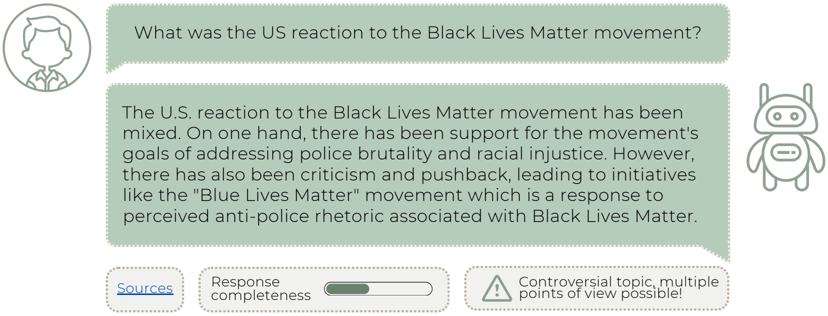


Figure 1.2: Conversational response generation, highlighting the key input components and the expected characteristics of the system output. The tasks of (1) personalizing responses based on user-specific information, and (2) enriching responses with external world knowledge beyond the retrieval corpus are included for completeness, but they fall outside the scope of this thesis and are left for future work.

The increasing reliance on digital information calls for transparent and trustworthy search systems in our daily interactions. In transitioning from traditional search engine result pages to a conversational setting that limits responses to a few sentences, there is a significant concealment of underlying details such as specifics about the sources, the scope of the answer, and the extent to which it is covered. These details are essential for users to assess the scope, novelty, reliability, and topical relevance of the provided information (Xu and Chen, 2006). Since the user is provided only with a short textual response as the final outcome of the generation process, it becomes the responsibility of the conversational system to identify and communicate any potential limitations to its users, ensuring transparency and empowering users to evaluate response quality. While the importance of explainability is broadly recognized for Artificial Intelligence (AI) (Monroe, 2018) and has been extensively studied, for example, for decision support and recommender systems (Nunes and Jannach, 2017; Zhang and Chen, 2020), it has not received due attention for CIS systems.

Recognizing that users are responsible for assessing the completeness, credibility, and accuracy of information provided by the system, it is crucial to equip them with the necessary tools for objective evaluation. We identify three key elements that are essential in this regard: (1) grounding the response in facts retrieved from sources, (2) presenting these sources to users, and (3) ensuring system transparency about the completeness of the provided information (see Figure 1.3). Grounding the response in verifiable sources significantly enhances system reliability and improves the quality of user interactions. Users with limited knowledge of the topic may struggle to filter out inaccurate content, especially as untrained individuals can only differentiate between human- and machine-generated text with near-random accuracy (Clark *et al.*, 2021). Source attribution further aids users in verifying factuality, increasing the transparency of the response generation process. For those unfamiliar with a subject, know-



What was the US reaction to the Black Lives Matter movement?

The U.S. reaction to the Black Lives Matter movement has been mixed. On one hand, there has been support for the movement's goals of addressing police brutality and racial injustice. However, there has also been criticism and pushback, leading to initiatives like the "Blue Lives Matter" movement which is a response to perceived anti-police rhetoric associated with Black Lives Matter.

Figure 1.3: Information-seeking dialogue with a CIS system, including explanations with links to the sources used for response synthesis, the completeness estimation of the generated response, and a warning about potential limitations of the returned answer.

ing the extent of the response's coverage is crucial to deciding how to proceed in their interactions with the system. This transparency helps users navigate the search space and refine their information needs (Azzopardi *et al.*, 2018). In edge cases, system transparency should also extend to handling unanswerable queries, as directly informing users about the system's limitations is far more reliable than providing vague or factually incorrect responses (Koopman and Zuccon, 2023).

Generating responses that meet these criteria would empower users to interpret information more critically and support their decision-making process. The research presented in this thesis explores the complexities of response generation in CIS systems, with a focus on ensuring grounding, completeness, transparency, and factual accuracy.

## 1.1 Research Questions

The main objective motivating this work is building a CIS system capable of generating transparent, factual, and grounded responses that enable users to navigate complex information needs successfully (see Figure 1.4). To advance research in these areas, we need a competitive baseline for both the retrieval component, to collect the sources answering the user's query, and the generation component, to synthesize this information into a natural answer. This motivates our first main research question:

> **RQ1:** What are strong baselines for **(a)** passage retrieval and **(b)** response generation in CIS systems?

Recognizing that high-performing retrieval does not guarantee that the generated responses are useful, we turn our focus to the factors that contribute to response limitations. For example, highly relevant passages retrieved by the

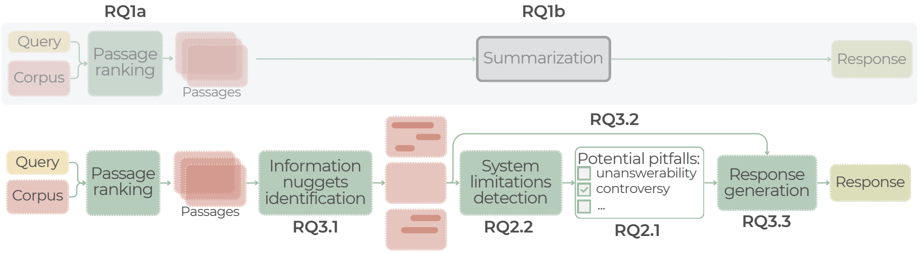


Figure 1.4: Overview of an explainable conversational response generation system compared to the baseline CIS pipeline, along with the research questions addressed in this thesis.

system may contain redundant information leading to low information density. On the other hand, for rare topics, the system may not find relevant information in the corpus, resulting in unaswerability that, if not handled properly, leads to relying entirely on the model's parametric memory and increases the risk of hallucinations. Operating on information units of finer granularity than documents or passages enables us to investigate response coverage, accuracy, and completeness. This leads us to the second main research question, along with specific subquestions:

> **RQ2:** What are the main limitations of CIS systems?
>
> > **RQ2.1:** Which limitations in the responses are detectable by users?
> >
> > **RQ2.2:** How to detect factors contributing to incorrect, incomplete, or biased responses?

Having identified the main challenges of response generation in CIS systems, we turn to addressing them by designing a system capable of generating transparent, grounded, and conversational responses. Providing users with transparent responses that acknowledge the system's limitations is paramount for fostering trust and empowering users to make informed judgments. We aim to generate responses that (1) synthesize the requested information, (2) are grounded in specific facts identified in the passages, (3) articulate the system's confidence, and (4) reveal the system's limitations. To accomplish that, we formulate the following research questions:

> **RQ3:** How to ensure transparent and explainable interactions with responses grounded in attributed sources for users?
>
> > **RQ3.1:** How to identify core information units in the relevant passages that need to be included in the response?
> >
> > **RQ3.2:** How to ensure the grounding of responses in the retrieved sources?

**RQ3.3:** How to generate responses transparent about the system's confidence and limitations?

## 1.2 Main Contributions

This section summarizes the main methodological and empirical contributions of this thesis, as well as the developed datasets that are made available to the research community.

### Methodological

- We design a data collection protocol for high-quality human annotation of information nuggets in CIS dialogues (Chapter 5).
- We develop a baseline approach for predicting query answerability based on the top retrieved passages (Chapter 5).
- We propose a framework for grounded response generation that ensures source attribution and enables automatic manipulation of response completeness in terms of unique pieces of relevant information included in the generated response (Chapter 6).
- We develop a methodology for evaluating the completeness of the response in terms of ground-truth information nuggets covered in the generated text (Chapter 6).

### Empirical

- We investigate the reproducibility of two CIS systems developed for the Conversational Assistance Track at Text REtrieval Conference (TREC CAsT) (Chapter 3).
- We perform a user study investigating the impact of query answerability and response incompleteness on user experience (Chapter 4).
- We conduct a user study investigating effective ways to provide explanations to accompany responses generated by the system (Chapter 7).

### Resources

- We build a high-quality dataset for conversational information seeking containing snippet-level annotations (Chapter 5).
- We build a dataset for question answerability prediction with answerability labels on the sentence, paragraph, and ranking levels (Chapter 5).

## 1.3 Organization of the Thesis

This thesis is organized into two main parts preceded with a background chapter that provides a comprehensive overview of related work in core areas of information access, natural language processing, evaluation techniques, and CIS systems.

Part I presents a baseline CIS system by reproducing two state-of-the-art systems from the TREC CAsT'22 track in Chapter 3. The goal of this chapter is to lay a solid foundation for the exploration of CIS limitations. The impact of query unanswerability and response incompleteness on user experience is explored in Chapter 4.

Part II builds on these insights and introduces a novel dataset annotated with information nuggets that serve as the atomic building blocks of answers. The dataset is extended with answerability labels at multiple levels—sentence, passage, and ranking—to provide a finer granularity assessment for answer verification, and is used to develop a baseline model for detecting query answerability based on the ranking of relevant passages in Chapter 5. Chapter 6 presents a response generation pipeline that ensures grounding of the response in the sources, enables control of the completeness of the response, and generates follow-up questions that are both answerable and relevant, thus enhancing user engagement and conversational continuity. Chapter 7 discusses the transparency of CIS responses and explores different strategies of presenting explanations about the sources, the system's confidence in the response, and any potential response limitations to the user.

The thesis concludes by revisiting the key findings and discussing their implications for the future of transparent, reliable, and explainable conversational information-seeking systems in Chapter 8. It synthesizes the contributions of the work by revisiting the research questions and showing how they have been addressed through the proposed methods. In this chapter, the limitations of the current work are also acknowledged, and avenues for future research are suggested. It ends with a reflection on the broader impact of this research in the fields of information retrieval and human-computer interaction.

## 1.4 Origins of the Material

The content of this thesis is based on a number of papers. Some of these have been published, while others are under review at the time of writing this dissertation.

### Introduction

**P1.** Łajewska (2024): *Grounded and Transparent Response Generation for Conversational Information-Seeking Systems*, WSDM '24 [doctoral consortium]

**Part I: Understanding Limitations**

**P2.** Łajewska *et al.* (2022): *The University of Stavanger (IAI) at the TREC 2022 Conversational Assistance Track*, TREC '22 [benchmark paper]

**P3.** Łajewska and Balog (2023a): *From Baseline to Top Performer: A Reproducibility Study of Approaches*, ECIR '23 [reproducibility paper]

**P4.** Łajewska *et al.* (2024a): *Can Users Detect Biases or Factual Errors in Generated Responses in Conversational Information-Seeking?*, SIGIR-AP '24 [full paper]

**Part II: Addressing Limitations**

**P5.** Łajewska and Balog (2023b): *Towards Filling the Gap in Conversational Search: From Passage Retrieval to Conversational Response Generation*, CIKM '23 🏆 [resource paper]

**P6.** Łajewska and Balog (2024a): *Towards Reliable and Factual Response Generation: Detecting Unanswerable Questions in Information-seeking Conversations*, ECIR '24 🏆 [short paper]

**P7.** Łajewska and Balog (2025): *GINGER: Grounded Information Nugget-Based Generation of Responses*, SIGIR '25 [short paper]

**P8.** Łajewska and Balog (2024b): *The University of Stavanger (IAI) at the TREC 2024 Retrieval-Augmented Generation Track*, TREC '24 [benchmark paper]

**P9.** Łajewska *et al.* (2024b): *Explainability for Transparent Conversational Information-Seeking*, SIGIR '24 [full paper]

**P10.** Łajewska and Balog: *X-GINGER: Explainable and Grounded Conversational Response Generation* [journal paper, submitted]

The following papers are not directly related to this thesis but brought insights to working with conversational systems, user-related data and large language models.

**P11.** Kostric *et al.* (2022): *DAGFiNN: A Conversational Conference Assistant*, RecSys '22 [demonstration paper]

**P12.** Skjæveland *et al.* (2024): *An Ecosystem for Personal Knowledge Graphs: A Survey and Research Roadmap*, AI Open, 2024 [journal paper]

**P13.** Sekulić *et al.* (2024): *Estimating the Usefulness of Clarifying Questions and Answers for Conversational Search*, ECIR '24 [short paper]

**P14.** Bernard *et al.* (2024): *PKG API: A Tool for Personal Knowledge Graph Management*, WWW '24 [short paper]

**P15.** Łajewska *et al.* (2025): *Understanding and Improving Information Preservation in Prompt Compression for LLMs* [full paper, under review]

# Chapter 2

# Background

*What is now proved was once only imagined.*

— **William Blake**

This chapter provides the theoretical and conceptual foundation necessary to understand the motivation and implications of the work carried out in this thesis. It explores the broad landscape of information access systems and relevant Natural Language Processing (NLP) tasks (see Figure 2.1), with a particular focus on conversational search—the central theme of this research.

Information Access is concerned with the ability to efficiently identify, retrieve, and use relevant information. Information access systems are algorithmic frameworks that bridge the gap between a collection of items (traditionally, documents) and a user's information need. The fundamental challenge is: given an item collection and a user's information need, how can the system present the most relevant items to meet that need? This information need, often a latent and unobserved concept, can be inferred from explicit user inputs like keyword queries or questions, as well as implicit data, such as previously consumed content or user behavior. As the user interacts with the system, their expression of need evolves (Bates, 1989). The success of an information access system is ultimately measured by the user's satisfaction with the results. These systems span several research areas, including information retrieval (IR), information filtering, recommender systems, and specific applications of NLP (Ekstrand *et al.*, 2022).

The first section of this chapter (Section 2.1) traces the evolution of document retrieval approaches, from classical document ranking techniques for key-

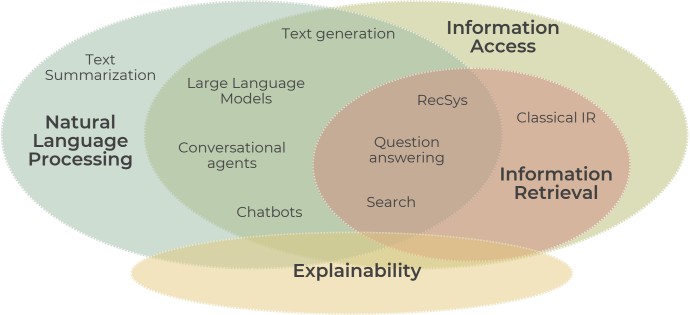


Figure 2.1: An overview of the fields of Information Access, Information Retrieval, and Natural Language Processing.

word search to neural IR models for passage reranking, and concludes with a discussion of common evaluation techniques, both human and automatic, in IR. It also covers the foundations of text representation and natural language modeling before we take a deep dive into their role in making users' interactions with the system more natural. Section 2.2 focuses on conversational search systems, introducing the concept of search as a conversational process. It examines conversational information-seeking dialogues, the task of generating conversational responses—particularly through retrieval-augmented generation—and addresses the challenge of evaluating natural language responses. The final section (Section 2.3) shifts to the open problem of explainability in Artificial Intelligence (AI) systems, exploring the development and evaluation of explainable components in search systems.

## 2.1 Information Retrieval

Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information (Salton, 1968). Document retrieval is a core task in IR, where the goal is to identify a subset of documents from a large document collection (a set of unordered text documents) that satisfy a user's information need defined using a query (Zhai and Massung, 2016). Traditional search systems provide a ranked list of documents in response to a query. Topic or domain is not restricted in these systems. In the web context, short keyword queries that a user types into a search box are merely the external manifestations of an information need, which is the motivation that compelled the user to seek information in the first place. Belkin (1980) calls this an "anomalous state of knowledge," where searchers perceive gaps in their cognitive states with respect to some task or problem. Queries are not synonymous with information needs. The same information need might give rise to different manifestations with different systems: for example, a few keywords are typed into the search box of a web search engine, but a fluent, well-formed natural language question is spoken to a voice assistant (Yates *et al.*, 2021).

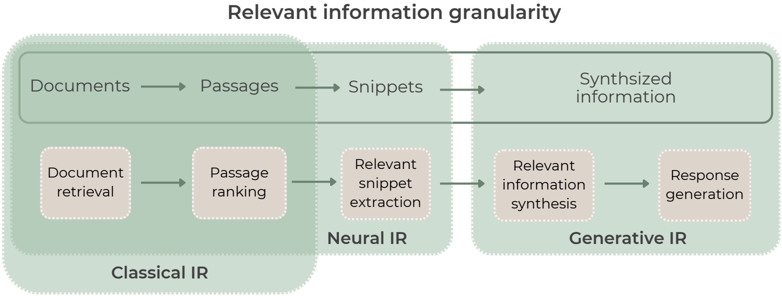


Figure 2.2: Evolution of methods used for document retrieval.

Traditional IR focuses on the representation of texts and queries, and comparison of these representations to identify documents that are relevant to the query (Belkin, 1995). The relevant document contains the information that a person was looking for when they submitted a query to the search engine. The first step of the IR process is offline indexing which includes defining the search collection (corpus with documents), taking into account metadata, and representing items in such a way that they can be connected with information needs. In the next step, the system performs online querying that focuses on understanding the query, putting it in the broader context of a specific user and the history of interactions, and representing it in such a way that it can be matched with the items in the collection. The retrieval model defines how a relevance score between a document and a query is computed using their respective representations; it estimates the utility of the document to an information need using a scoring function. Based on the relevance scores, the system retrieves items that match the need. The last step renders the retrieved items to be presented to the user.

In traditional IR, the system indexes whole documents (or passages) and returns a ranking of the most relevant items retrieved. However, more complex queries may require operating on shorter text, such as text snippets, also referred to as information nuggets. More advanced systems not only retrieve relevant snippets of information but also synthesize them into a natural, concise, and coherent response (see Figure 2.2). Retrieving shorter relevant information nuggets and synthesizing them into a final response simplifies the interaction with the system for the user and reduces the cognitive burden of extracting the pieces of information that actually address the user's information needs from search results. This scenario is discussed in more detail in Section 2.2.

### 2.1.1 Relevance

The concept of relevance is fundamental for the functioning and evaluation of IR systems (Borlund, 2003). Relevance is subjective, meaning it can be perceived and assessed differently by different users. It is also dynamic, as the same user's perception of relevance may shift throughout a session (Borlund, 2003).

Topical relevance—whether a document pertains to the subject at hand—differs from user relevance, which is subjective and shaped by the individual's abstract information need. Ultimately, only the user who formulates the information need can determine which documents are relevant, as evaluating relevance is a complex cognitive process. Therefore, each person is the ultimate arbiter of relevance for their own information need. Relevance is not an absolute truth or an inherent property of a text that can be "unlocked" by an assessor—it varies based on personal interpretation (Yates *et al.*, 2021).

Relevance judgments are human-provided relevance annotations on query-document pairs and they serve two key purposes: they are used both to train ranking models in supervised settings and to evaluate the effectiveness of those models. Relevance judgments reflect a particular individual's opinion, making them inherently subjective. Assessor agreement on relevance judgments is typically low, with a commonly cited overlap of just 60% (Voorhees, 1998), with overlap defined as the size of the intersection of the relevant document sets divided by the size of the union of the relevant document sets. This low agreement arises because assessors often interpret information needs based on external representations like topic descriptions, which may not fully capture the original user's cognitive state. As a result, assessors may have differing interpretations of what constitutes relevance. Although the specific scores of ranking systems may vary depending on which assessor's judgments are used, the relative ranking of different systems tends to remain stable despite these variations (Voorhees, 1998).

### 2.1.2 Text Representation

Text representations form the foundation of IR, enabling machines to represent textual data in a manner that captures semantic and syntactic properties. Word representation learning is typically an unsupervised or self-supervised procedure that does not require manual training data annotation. Instead, raw texts, available at scale, can be reliably used to compute word statistics that inform the creation of semantic representations. These representations have evolved from simplistic one-hot encodings to sophisticated contextualized embeddings.

One of the earliest methods for representing words, one-hot encoding, maps each word to a distinct fixed-size binary vector. In this representation, a word is represented as an n-dimensional vector (where n is the vocabulary size), with all values set to zero except for one, indicating the word's index in the vocabulary. While unique, one-hot representations do not indicate semantic similarity between words and are computationally expensive, as their size grows proportionally with the vocabulary.

A significant advancement was the shift to vector space representations, where words, documents, or other semantic units are mapped to points in a continuous, multidimensional semantic space (Salton *et al.*, 1975). Unlike discrete one-hot representations, vector space models capture semantic similarity by representing the relationships between words as distances in space. This representation laid the groundwork for understanding and modeling word meaning through proximity.

The distributional hypothesis, stating that words occurring in similar contexts have similar meanings (Harris, 1954; Firth, 1957), is a foundation for early word vector space models. Early approaches, referred to as count-based methods, used word occurrence and co-occurrence statistics to create word vectors. However, raw frequency counts proved to be an unreliable measure of association due to the over-representation of common words bearing little semantic meaning, e.g. "the." To address this, methods like positive pointwise mutual information (PPMI) were introduced to normalize co-occurrence frequencies, emphasizing meaningful associations between words by checking whether two words co-occur more than they occur independently (Church and Hanks, 1990). Despite their utility, these methods generate high-dimensional sparse representations, that require dimensionality reduction techniques like singular value decomposition (SVD) (Turney, 2005).

With the rapid development of neural networks, they have been directly applied to learning dense and low-dimensional word representation without having to resort to an additional dimensionality reduction step. Word embeddings were popularized by Word2Vec (Mikolov *et al.*, 2013), which employs a simple feedforward neural network trained using a language modeling objective. Two models, Continuous Bag-of-Words (CBOW) and Skip-Gram, were proposed. CBOW predicts a word based on its context, while Skip-Gram predicts context words given a target word. These embeddings provided richer representations while eliminating the need for separate dimensionality reduction steps.

Traditional embeddings like Word2Vec are static, meaning that a word's representation remains unchanged regardless of its context. Contextualized embeddings address this limitation by dynamically adapting word representations based on their context. These embeddings capture both semantic and syntactic nuances of words, enabling more sophisticated language understanding (Hewitt and Manning, 2019). Built using advanced language models, contextualized embeddings leverage the predictive task of modeling words within a sequence, ensuring that both syntactic and semantic roles are encoded in the representation.

### 2.1.3 Language Models

A language model (LM) is a probabilistic framework, originally designed to distinguish grammatical from ungrammatical sequences in a given language (Chomsky, 1957). Traditional statistical language modeling relies on the $n$-th order Markov assumption, estimating the probability of a sequence based on n-gram frequencies derived from large text corpora. However, this approach faces significant limitations due to data sparsity, as the number of possible n-grams grows exponentially with vocabulary size and sequence length, a challenge often referred to as the *curse of dimensionality.*

Neural language models address the limitations of count-based methods by replacing discrete one-hot word representations with continuous distributed representations. In this approach, each word is represented as a vector in a continuous space, capturing semantic and syntactic relationships. Instead of memorizing specific word sequences, neural LMs learn to generalize patterns, enabling them to assign probabilities to unseen sequences based on their similarity to observed patterns. This shift allows smoother and more robust modeling of language.

Modern LMs go beyond simple next-word prediction, encoding complex syntactic and semantic information. They are broadly used for both natural language understanding and generation tasks, and their ability to learn from raw text in an unsupervised manner has removed the bottleneck of manual knowledge acquisition. Large-scale LMs are often pre-trained on extensive corpora with a language modeling objective and then fine-tuned for specific tasks (Radford *et al.*, 2019). This pre-training process builds contextualized embeddings that capture deep semantic relationships and is now widely used for downstream applications such as summarization and question-answering.

The evolution of contextualized embeddings began with ELMo (Peters *et al.*, 2018), which utilized recurrent neural networks (RNNs). However, RNN-based models were quickly surpassed by Transformer-based architectures (Vaswani *et al.*, 2017), which offer several key advantages: (1) parallel processing of inputs and (2) the self-attention mechanism, enabling the model to focus on relevant words across an entire sequence, even if they are far apart. Generative Pre-trained Transformer (GPT) (Radford *et al.*, 2018) was one of the first attempts at representation learning with Transformers. However, both ELMo and GPT were based on unidirectional language modeling, limiting their ability to fully understand a word's meaning by considering only the preceding token. This constraint led to the development of Bidirectional Encoder Representations from Transformers (BERT) (Devlin *et al.*, 2019) that leverages the full Transformer encoder architecture to jointly condition on both left and right contexts, enabling richer and more nuanced text representations. BERT's bidirectional nature is achieved through a masked language modeling objective, where certain tokens in an input sequence are masked, and the model is trained to predict these tokens using context from both sides. Additionally, BERT introduced a next-sentence prediction task, training the model to determine whether one sentence logically follows another. This dual training objective allows BERT to encode both word-level and sentence-level relationships, significantly enhancing its ability to understand and generate coherent text.

These early pre-trained context-aware word representations, such as ELMo and BERT, have proven to be highly effective as general-purpose semantic features for a wide range of NLP tasks. The subsequent wave of language models was largely driven by scaling existing solutions in terms of model size or data volume, a strategy shown to significantly enhance performance on downstream tasks (Kaplan *et al.*, 2020). Model parameters have grown dramatically, from 101M in BERT and 120M in GPT-1 to 70B in LLaMA2 (Touvron *et al.*, 2023), 540B in PaLM (Chowdhery *et al.*, 2023), and an astounding 1.76T in GPT-4 (OpenAI *et al.*, 2024). These large pre-trained language models, commonly referred to as Large Language Models (LLMs), exhibit remarkable capabilities not observed to the same extent in smaller models, despite having similar architectures and training objectives. A prominent example is in-context learning (Brown *et al.*, 2020), where the model generates text based on a provided context or prompt, enabling it to produce more coherent and contextually relevant responses. This makes LLMs particularly well-suited for interactive and conversational applications. Another key advancement is Reinforcement Learning from Human Feedback (Christiano *et al.*, 2017), a fine-tuning technique that uses human-generated responses as rewards, allowing the model to iteratively learn from its mistakes and improve performance over time.

Advancements in language modeling, from statistical approaches to sophisticated neural architectures, have dramatically improved text representation quality. These improvements have cascaded into better performance across a wide range of natural language understanding and generation tasks, solidifying language models as the backbone of modern NLP applications including conversational information access, multi-document abstractive summarization, and multi-hop question answering.

### 2.1.4 Ranking Models

The goal of text ranking is to produce an ordered list of texts from a corpus in response to a user query for a specific task. IR systems typically operate by ranking items such as documents or passages. Historically, the most common ranking approach has been to sort items by decreasing relevance according to the Probability Ranking Principle (PRP) (Robertson, 1977). The PRP states that documents should be ranked in descending order of their estimated probability of relevance to the user's information need. Ranking remains a fundamental component in most information access approaches, serving as the primary basis for evaluating the effectiveness of retrieval systems.

One classical approach to text ranking is the vector space model, where documents and queries are represented as vectors in a high-dimensional space. The core idea is that if a document $d_1$ is more similar to the query than another document $d_2$, $d_1$ is considered more relevant. In this model, each dimension corresponds to a term in a vocabulary, and relevance is determined by calculating similarity measures such as the dot product or cosine similarity between the vectors representing the query and document (Salton *et al.*, 1975). Common text representations include the term frequency-inverse document frequency (TF-IDF), which assigns greater weight to terms that appear in fewer documents, as such terms are more likely to be indicative of relevance.

In contrast, probabilistic retrieval models treat queries and documents as random variables and estimate the likelihood that a document is relevant to a query. The score of a document is defined with respect to a query as the probability that this random variable is equal to 1 given a particular document and query (Lafferty and Zhai, 2003). One prominent probabilistic model is the query likelihood model, which quantifies how likely it is that a given query would be generated by sampling words from a specific document (Ponte and Croft, 1998). Among the most effective retrieval models derived from this framework is BM25, which adjusts for term frequency and document length, making it a widely-used method (Robertson and Zaragoza, 2009).[1]

The joint characteristic of both vector space and probabilistic models is their reliance on sparse representations, where each vector has a length corresponding to the vocabulary size, with components encoding discrete values indicating the presence or absence of a term in a document. These sparse representations, based on the bag-of-words (BOW) model, ignore word positions and treat texts as unordered sets of terms. While sparse representations can be efficiently indexed using inverted indices, they suffer from a key limitation: the vocabulary mismatch problem. This arises because traditional models rely on exact term matching between the query and document, meaning that relevant documents might not be retrieved if they do not contain the specific query terms, even if they address the same underlying information need.

### 2.1.5 Neural IR

BM25 and other term-weighting schemes briefly discussed in the previous section are typically unsupervised, relying on statistical properties of terms within texts to estimate relevance. In contrast, learning-to-rank (LTR) approaches leverage supervised machine learning to create ranking models based on manually-engineered features. These features are often derived from statistical properties of terms and intrinsic text characteristics (Liu, 2010). However, LTR models require extensive feature engineering, which is labor-intensive and limits scalability. The emergence of deep learning has enabled the use of continuous vector representations (see Section 2.1.2), removing the need for such hand-crafted features while addressing challenges associated with exact term matching (Yates *et al.*, 2021).

Neural ranking models can be broadly classified into two types. Representation-based models independently learn dense vector representations of queries and documents and compare them using similarity metrics like cosine similarity or inner products (Nalisnick *et al.*, 2016; Huang *et al.*, 2013). Interaction-based models, on the other hand, focus on capturing term-level interactions between queries and documents by generating a similarity matrix to determine relevance (Guo *et al.*, 2016; McDonald *et al.*, 2018). Both approaches typically use neural models, such as convolutional or recurrent neural networks, to extract relevance signals and are trained end-to-end using relevance judgments. The deep learning shift accelerated with the introduction of BERT (Devlin *et al.*, 2019) (see Section 2.1.3), which was first applied to the MS MARCO passage ranking task (Nogueira and Cho, 2019), and started a new era of transformer-based ranking models.

One of the most notable applications of BERT in ranking is the monoBERT model, which formulates text ranking as a binary classification problem where the model predicts the probability that a text is relevant to the user's query.

[1] Even though BM25 originates from the probabilistic retrieval framework, it functions as a term-weighting scheme, interpreting term contributions to relevance, and operates similarly to the vector space model using inner products on sparse bag-of-words vectors.

This a direct realization of PRP where documents are ranked based on their estimated probability of belonging to the relevant class (Nogueira and Cho, 2019). MonoBERT is a pointwise approach, meaning each document is evaluated independently during training (Liu, 2010) and it assumes binary relevance. In contrast, duoBERT adopts a pairwise approach, comparing two candidate documents at a time to estimate which is more relevant to the query (Nogueira *et al.*, 2019). Pairwise ranking has the advantage of harnessing signals present in other candidate texts to decide if a text is relevant to a given query, implementing the notion of graded relevance judgments. Although pairwise approaches can yield more nuanced rankings, they are computationally expensive due to their quadratic growth in the number of comparisons as the size of candidates set increases (Nogueira *et al.*, 2019).

Modern ranking pipelines often employ multi-stage retrieval, where an initial sparse retrieval method (e.g., BM25) identifies a candidate set of documents, followed by neural rerankers like monoBERT to refine the ranking. However, keyword-based retrieval methods can miss relevant documents due to the vocabulary mismatch problem, where the query and relevant documents use different terminology. Dense retrieval models address this challenge by learning dense representations of queries and documents, transforming text into fixed-length vectors. These vectors are optimized so that the similarity between relevant query-document pairs is maximized, while non-relevant pairs are minimized, based on a given similarity metric. In dense retrieval, document representations are precomputed offline to facilitate low-latency searches. At query time, query and document vectors are compared using simple operations like inner product, enabling efficient nearest neighbor search. To scale this approach, techniques such as approximate nearest neighbor (ANN) algorithms are often employed, making dense retrieval a practical solution for large collections (Xiong *et al.*, 2020).

One of the first approaches in dense retrieval is DPR proposed for open-ended QA (Karpukhin *et al.*, 2020). DPR leverages the BERT pretrained model alongside a dual-encoder architecture and focuses on an optimized training scheme using a relatively small number of question-passage pairs. The embeddings are trained to maximize the inner product between question and relevant passage vectors, with the objective of comparing all pairs of questions and passages in a batch. Around the same time, ANCE (Approximate Nearest Neighbor Negative Contrastive Estimation) was introduced (Xiong *et al.*, 2020). Like DPR, ANCE employs a bi-encoder design, but it uses RoBERTa's `[CLS]` token representation as the encoder and applies the same encoder to both queries and documents. ANCE introduces hard negative selection via ANN search on an index built from encoder-generated representations. To mitigate the computational burden of maintaining a fully up-to-date index, the ANN index is updated asynchronously during training.

Recent advancements in ranking techniques have taken advantage of LLMs. RankGPT demonstrates the successful application of generative LLMs, such as ChatGPT and GPT-4, for relevance ranking (Sun *et al.*, 2023). RankVicuna is the first fully open-source LLM capable of performing high-quality listwise reranking in a zero-shot setting (Pradeep *et al.*, 2023). It uses RankGPT as a teacher model to generate high-quality ranked lists through prompt decoding. Another innovative approach, the Pairwise Ranking Prompting (PRP) paradigm, simplifies ranking tasks for LLMs by including only the query and a pair of documents in the prompt (Qin *et al.*, 2024). This approach reduces task complexity for LLMs and resolves calibration issues. PRP relies on straightforward prompt design and works with both generative and scoring LLM APIs. PRP has achieved state-of-the-art ranking performance using moderate-sized, open-source LLMs on standard benchmark datasets.

### 2.1.6 Evaluation

Evaluation in IR presents significant challenges due to the vast size of corpora and the inherent ambiguity in queries representing user information needs. The complexity is compounded by the fact that IR systems are highly interactive, with users playing a central role in the process. There are two primary approaches to capturing the human element in evaluation: system-oriented and user-oriented evaluations (Gao *et al.*, 2023a). System-oriented evaluation focuses on fixed datasets that capture user requests and preferences, which can be shared and reused. This approach allows researchers to develop models that best match the preferences captured in the dataset, supporting both component-based and end-to-end evaluations. In contrast, user-oriented evaluation observes real users interacting with the system, either in controlled lab environments or field studies. These studies may involve surveys or diary entries to gain deeper insights into user behavior. For deployed systems, user log data can be analyzed to capture large-scale interactions, though interpreting this data to improve system evaluation poses its own challenges (Dumais *et al.*, 2014). Both approaches are complementary, with each informing the other in improving IR systems (Gao *et al.*, 2023a).

#### System-Oriented Evaluation

A test collection for evaluating IR models includes a text corpus, a set of information needs (topics), and relevance judgments (qrels). The Text Retrieval Conference (TREC) series organized by the National Institute of Standards and Technology (NIST), facilitates large-scale, community-wide evaluations of IR methods, enabling collaboration between academia, industry, and government. TREC workshops, initiated in 1992, have shaped IR research, offering standardized test collections and fostering technology transfer to real-world applications. System evaluation at TREC follows the Cranfield paradigm which is a system-oriented methodology that treats search as an optimization problem and uses quantitative ranking metrics, such as precision, recall, and reciprocal rank, which are based on relevance judgments (Sanderson, 2010). Discounted cumulative gain (DCG) is a more advanced metric that accounts for graded relevance and the diminishing likelihood that users will examine lower-ranked documents (Kekäläinen and Järvelin, 2002; Järvelin and Kekäläinen, 2002). Following this system-oriented evaluation approach, TREC's "top-k pooling" method gathers top-ranked results from participants for each topic to generate relevance judgments. TREC's success has inspired similar initiatives worldwide, such as CLEF in Europe and NTCIR in Asia.

Due to the time and cost involved in document-based relevance judgments, alternative approaches like nugget-based evaluation have been explored. The main task of TREC'06 Question Answering Track consists of series of natural language questions and the goal is to return answers, rather than documents containing answers. System answers are evaluated by identifying "nuggets" defined as "minimal, atomic units of relevant information" representing a conceptual entity (Pavlu *et al.*, 2012). Nuggets are classified by the assessor as either vital (representing concepts that must be present in a good answer) or okay (containing interesting information, but are not essential) with an F-score defined as a weighted harmonic mean between nugget precision and nugget recall, that is calculated solely on vital nuggets. Nugget pyramids extend this idea by collecting judgments from multiple assessors for finer granularity (Dang and Lin, 2007). Similarly, Semantic Content Units (SCU), defined as concise sentences containing a single fact, are used in summary evaluation to assess factual accuracy and coverage. The Pyramid method involves extracting SCUs from reference summaries and checking system summaries for SCU coverage to generate an overall score (Nenkova and Passonneau, 2004; Shapira *et al.*, 2019).

The idea of using large language models (LLMs) as "judges" for relevance is getting a lot of attention, with recent studies showing potential but also limitations (Faggioli *et al.*, 2023; Dietz, 2024; MacAvaney and Soldaini, 2023; Balog *et al.*, 2025). There is a spectrum of collaboration between humans and LLMs in this task, from fully human-driven judgments to fully automated ones, with various mixed approaches in between. LLMs can enhance relevance judgments by offering explanations, scalability, and consistency, underlining the great potential of deploying them as a complement to human assessors in certain judgments task. Faggioli *et al.* (2023) show promising results, with LLM judgments correlating reasonably well with human assessments, though LLMs sometimes miss subtle details that humans catch, especially in borderline scenarios. Sensitivity, or the ability to detect meaningful differences between retrieval systems, is also generally lower in LLM judgments than in human ones. However, LLMs prove their effectiveness in evaluating open-domain questions. Well-instructed LLMs can distinguish between relevance and utility, and are highly receptive to newly generated counterfactual passages (Zhang *et al.*, 2024). The EXAM++ Answerability Metric further illustrates the potential of LLMs for automatically assessing the information content of a system's response, without resorting to expensive human judgments (Farzi and Dietz, 2024a).

### User-Oriented Evaluation

The Cranfield paradigm for evaluating information retrieval and text ranking systems raises concerns about whether test collections truly reflect real-world information needs. A key question is whether system improvements, as measured by Cranfield, actually benefit users. The ultimate goal is user satisfaction, as users seek information to accomplish tasks like making a purchase, writing a report, or finding a job. However, improvements in ranking models do not necessarily translate into better task performance for users (Hersh *et al.*, 2000). While Cranfield evaluations provide useful insights into model effectiveness, they miss the complete picture and should be complemented by human evaluations to ensure real-world usefulness.

Interactive evaluations, which place humans in the loop, are crucial for understanding user behavior in information-seeking dialogues (Kelly, 2007). User-focused criteria in interactive IR include credibility, cognitive load, engagement, satisfaction, information gain, effort, and task success (Anand *et al.*, 2019). From a retrieval perspective, factors like utility and completeness are important, while dialogue evaluations consider information gain about the user and error recovery. Usability measures, on the other hand, rely on evaluative feedback from users, gaining insights into their attitudes, feelings, and overall experience with the system (Kelly, 2007).

Human evaluation is often performed using user studies or crowdsourcing. The two approaches differ significantly in their objectives, participant selection, and methodology. A user study is typically focused on in-depth analysis of user behavior, preferences, or usability, with carefully selected participants who represent the target demographic or user base. It is often used to test specific features or hypotheses within a controlled environment. Factorial designs are commonly used to study the effects of multiple variables on outcome measures, with each variable, or factor, having discrete levels. Studies may use between- or within-subjects designs, or a mix of both, to examine the impact of these variables. Techniques like rotation, counterbalancing, and randomization help control for order effects. Pilot tests are essential to refine the study setup, instruments, and protocols, ensuring the reliability and validity of the main experiment (Kelly, 2007).

In contrast, crowdsourcing gathers input, data, or services from a large, diverse group of people through online platforms. Crowdsourcing participants are generally less targeted, as the goal is to scale data collection or problem-solving across many individuals quickly and efficiently. Crowdsourcing harnesses the intelligence of large, diverse groups to complete tasks that computers struggle with, but it also introduces challenges in quality control due to the varied skills and motivations of crowd workers (Daniel *et al.*, 2019). Platforms like Amazon Mechanical Turk have proven effective for quickly and affordably evaluating search relevance by employing quality control methods such as training phases and inserting gold-standard data (Le *et al.*, 2010). Crowdsourcing combines scalability with the power of human judgment to solve complex tasks, but task quality depends heavily on factors like worker characteristics (Kazai *et al.*, 2011), platform design (Vakharia and Lease, 2013), task setup (Eickhoff, 2018), and quality measures employed (Allahbakhsh *et al.*, 2013). The design of manual annotation, including the number and assignment of annotators, significantly impacts the reliability of data for statistical analysis (Steen and Markert, 2021).

## 2.2 Conversational Information Seeking

Conversational AI has traditionally been categorized into task-oriented and non-task-oriented systems (Chen *et al.*, 2017). More contemporary distinctions expand this view, identifying task-oriented systems, social chats, and interactive question-answering (QA) systems as distinct categories (Gao *et al.*, 2019). Among these, conversational information access represents a specialized subset of conversational AI systems that focus on task-oriented exchanges. These systems support diverse user goals, such as search, recommendation, and exploratory information gathering, often requiring multi-step interactions across multiple modalities. They blend characteristics of task-oriented and interactive QA systems, leveraging both short-term and long-term user information to address complex information-seeking tasks (Gao *et al.*, 2023a).

At its core, a conversational information access process can be viewed as a task-oriented dialogue, where the primary task is information seeking (Gao *et al.*, 2023a). Information-seeking tasks are generally classified into two categories: information lookup and exploratory search (Marchionini, 2006). Lookup tasks involve retrieving factoid answers or addressing specific questions for which modern search engines and database systems are well-suited. In contrast, exploratory tasks are more complex, requiring long-term iterative interactions. Exploratory searches demand an intensive sensemaking process (Pirolli and Card, 2015), where users synthesize information from multiple sources to form a conceptual understanding. Examples include learning searches, which require synthesizing information to gain new knowledge, and investigating searches, such as travel planning or academic research, where users iterate through information to form personal perspectives (Gao *et al.*, 2023a).

Conversation, as an interactive process for exchanging information, involves a sequence of interactions between two or more participants—humans or machines. Information-seeking conversations specifically aim to satisfy the information needs of one or more participants. Conversational Information-Seeking (CIS) systems enable users to navigate unfamiliar domains, address complex information needs, ask follow-up questions, and provide feedback through natural language dialogues (Zamani *et al.*, 2023). These systems address a wide range of queries, from straightforward factoid questions to open-ended inquiries requiring the exploration of diverse viewpoints. Exploratory search, a common scenario in CIS interactions, often involves users with little prior knowledge about the topic, making them more susceptible to misinformation or biases due to their limited ability to verify the system's responses (Schneider *et al.*, 2023). By facilitating iterative, multi-turn dialogues, CIS systems address the inherent complexities of information-seeking tasks, bridging the gap between user intent and the vast landscape of available information.

### 2.2.1 Conversational Search Systems

A CIS system is designed to address the information needs of one or more users through interactive, dynamic conversations. These systems provide responses that are concise, fluent, stateful, mixed-initiative, context-aware, and personalized (Zamani *et al.*, 2023). Conversational search focuses on retrieving and synthesizing information through iterative, context-driven dialogues. Unlike question answering, which often provides direct and specific responses, conversational search deals with broader and more complex information-seeking dialogues. This includes addressing open-ended questions requiring descriptive answers, indirect responses that involve inference or contextual knowledge, and queries with partial answers distributed across multiple passages (Bolotova-Baranova *et al.*, 2023). This complexity means that conversational search systems must go beyond simply returning snippets from the top-retrieved passages; they need to synthesize information from multiple sources to construct comprehensive and coherent responses.

Conversational search systems enable a *mixed-initiative* interaction, where both the user and the system contribute to guiding the dialogue. The system's actions are informed by a dynamic understanding of the user's current needs, taking into account both the immediate conversational context and long-term knowledge about the user (Radlinski and Craswell, 2017). To effectively support these interactions, the system helps users articulate and refine their information needs, including uncovering latent preferences (so-called *user revealment).* At the same time, it transparently communicates its capabilities and limitations, helping users form realistic expectations of what it can and cannot do (*system revealment*). The system also maintains *memory* of past interactions, allowing users to reference previous statements and ensuring consistency unless explicitly contradicted. Furthermore, it reasons about the utility of sets of complementary items, optimizing its retrieval process to provide more comprehensive support (*set retrieval*) (Radlinski and Craswell, 2017). By integrating these features, conversational search systems are able to manage iterative, adaptive, and context-sensitive interactions. This makes them particularly well-suited to addressing complex, exploratory information-seeking tasks.

### Mixed Initiative

Mixed-initiative interaction is a flexible interaction strategy in conversational search systems where both the user and the system contribute to the conversation independently. This approach allows the system to take control either at the dialogue level, by asking clarifying questions or requesting elaboration, or at the task level, by suggesting alternative courses of action (Horvitz, 1999). By enabling both participants to initiate and guide the dialogue, mixed-initiative systems achieve a more human-like and dynamic interaction. The primary objective of incorporating mixed-initiative strategies is to improve search effectiveness by allowing the system to take the initiative when needed (Radlinski and Craswell, 2017).

At different points in a conversation, the balance of initiative may shift between the user and the system. For instance, the system might take the lead to clarify ambiguous user requests, elicit additional information, or prevent errors, while also allowing the user to drive the dialogue at other times (Radlinski and

Craswell, 2017). Clarifying questions help resolve ambiguities in user inputs, minimize errors, and refine the understanding of user needs. Advanced systems enhance these capabilities further by incorporating context tracking, enabling them to continuously monitor the topic of conversation and ask follow-up questions (Zamani *et al.*, 2023). This interplay enhances the system's ability to provide accurate and relevant responses.

Another dimension of mixed-initiative interaction is system revealment, which emphasizes transparency, reliability, and trustworthiness by disclosing the system's capabilities and the underlying corpus (Radlinski and Craswell, 2017; Azzopardi *et al.*, 2018). Through system revealment, users gain a better understanding of the scope and limitations of the system, which helps manage their expectations. In conversational search, this involves informing users about the characteristics of the available search space and assisting them in articulating their information needs (Radlinski and Craswell, 2017). For example, when a user's information need is unclear and the system is unable to locate an answer, it might ask clarifying questions. Similarly, when the system understands the query but the desired information is unavailable in its collection, it must handle unanswerable questions appropriately, maintaining user trust and engagement.

### TREC Conversational Assistance Track (CAsT)

Launched in 2019, the TREC Conversational Assistance Track (CAsT) has played a pivotal role in advancing research on CIS systems by providing large-scale reusable test collections (Dalton *et al.*, 2019, 2020). Unlike generative AI approaches, TREC CAsT emphasizes answers grounded in specific passages, focusing on conversational passage retrieval as a key problem (Pradeep *et al.*, 2021; Vakulenko *et al.*, 2021c; Luan *et al.*, 2021). These collections include MS MARCO (Campos *et al.*, 2016), Wikipedia (Petroni *et al.*, 2021), TREC CAR (Dietz *et al.*, 2018), and the Washington Post v4.[2]

One of the main characteristics of TREC CAsT tasks is their emphasis on context. In the 2019 edition of the task (Dalton *et al.*, 2019), user utterances are contextualized solely by references to previous user utterances. Since 2020, the task scope expands significantly by incorporating references to prior system responses, requiring systems to integrate a broader range of contextual information (Dalton *et al.*, 2020). TREC CAsT'21 is characterized by the increased dependence on previous system responses, as well as simple forms of user revealment, reformulation, and explicit feedback introduced in users' utterances (Dalton *et al.*, 2021). Starting in 2020, the task scope expanded significantly by incorporating references to prior system responses, requiring systems to integrate a broader range of contextual information (Dalton *et al.*, 2020). By 2021, the task introduced even greater dependence on previous system responses, along with simple forms of user revealment, reformulation, and explicit feedback embedded in user utterances (Dalton *et al.*, 2021). TREC CAsT collections have

[2] https://trec.nist.gov/data/wapost/

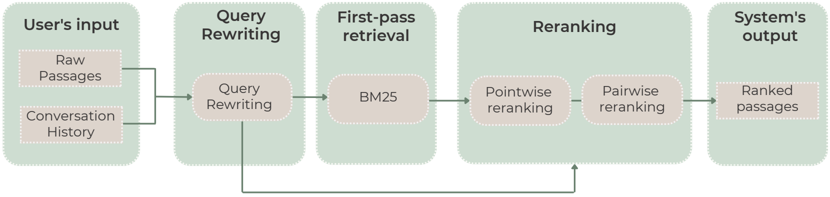


Figure 2.3: Architecture of a two-step passage ranking pipeline with a query rewriting module.

provided a crucial benchmark for enabling systems to handle increasingly complex and dynamic interactions effectively.

### Standard CIS Pipeline

By the time of TREC CAsT'21, a standard architecture for conversational search systems had emerged, characterized by a two-step passage ranking pipeline (see Figure 2.3). The first step typically involves passage retrieval using an unsupervised sparse retrieval model, such as BM25. This initial retrieval is followed by reranking, employing a neural model trained on passage retrieval tasks, such as T5 fine-tuned on MS MARCO (Craswell *et al.*, 2020). A key element of this pipeline is a query rewriting module that ensures the conversational queries to be de-contextualized, making them independent of previous dialogue turns.

Query rewriting is central to managing conversational phenomena such as omission, coreference (Dalton *et al.*, 2019), zero anaphora, topic changes, and topic returns (Voskarides *et al.*, 2020). Approaches to query rewriting can be grouped into three categories: unsupervised methods, supervised feature-based methods, and supervised neural methods. Unsupervised query rewriting methods expand the original query using terms from the conversation history. This can be achieved using similarity metrics, such as BM25 scores (Yilmaz *et al.*, 2019), cosine similarity (Voskarides *et al.*, 2019), or frequency-based signals (Lin *et al.*, 2021). Supervised feature-based approaches rely on linguistic features, including dependency parsing, coreference resolution, named entity recognition, and part-of-speech tagging (Mele *et al.*, 2020). Supervised neural methods, on the other hand, leverage large pre-trained language models like GPT-2 (Vakulenko *et al.*, 2021a) or T5 (Yan *et al.*, 2021), fine-tuned on conversational datasets such as CANARD (Vakulenko *et al.*, 2021a) or QReCC (Yan *et al.*, 2021). These neural models can generate reformulated queries, further enriched with conversation history terms (Vakulenko *et al.*, 2021a), paraphrases (Ju *et al.*, 2021), or sentences from semantically related documents (Chang *et al.*, 2020).

The architectural choices in CIS systems exhibit significant diversity. While the most common pipeline involves retrieval followed by reranking and query rewriting, alternative approaches have also been explored. For instance, some systems adopt few-shot conversational dense retrieval, which scores documents using the dot product of contextualized embeddings of user utterances and collection of documents (Yu *et al.*, 2021). Another technique, document expansion, addresses vocabulary mismatches between user queries and document content, where sequence-to-sequence models are employed to enhance document content by boosting term statistics and generating additional terms. Techniques like Doc2Query generate new queries tailored to a specific document, making it more likely to align with diverse query formulations (Gospodinov *et al.*, 2023). Together, these advancements in query rewriting, passage retrieval, reranking, and document expansion form the backbone of modern conversational search system pipelines, enabling them to find documents addressing complex information needs. However, to make the system truly conversational, an additional step is needed to synthesize the retrieved information into a natural language response.

### 2.2.2 Response Generation

Traditional search engines focus on delivering ranked lists of documents, leaving users to actively go through the results to find specific answers (White, 2014). In contrast, conversational response generation aims to synthesize information from retrieved passages into a single, concise, and coherent response that captures the most relevant details in an easily consumable form. This process requires responses to be factual, grounded in credible sources, complete, transparent, coherent, and free from hallucinations (Sakai, 2023; Gienapp *et al.*, 2024).

The growing interest in conversational response generation spans various domains, including task-oriented dialogue systems (Budzianowski *et al.*, 2018; Lippe *et al.*, 2020; Pei *et al.*, 2020; Dubiel *et al.*, 2020), question answering (Baheti *et al.*, 2020), and open-domain chatbots (Xing *et al.*, 2017; Dziri *et al.*, 2019; Tian *et al.*, 2019). Its application to conversational information-seeking gained momentum in the 2022 edition of TREC CAsT, which introduced a subtask focused on generating responses from retrieved results (Owoicho *et al.*, 2022). The approach proposed by Ren *et al.* (2021) divides this task into three stages: (optional) query rewriting, identifying supporting sentences from search engine result pages, and summarizing them into concise conversational responses.

Generative language models, such as GPT-2 (Zhang *et al.*, 2020b), have become instrumental in response generation. However, aggregating supporting facts through abstractive summarization (Ferreira *et al.*, 2022; ter Hoeve *et al.*, 2022) introduces risks of factual errors (Tang *et al.*, 2023) and hallucinations (Cao *et al.*, 2016). Existing search engine responses often produce outputs of high fluency and perceived utility that frequently contain unsupported statements or inaccurate citations (Liu *et al.*, 2023a). Even with the recent development of LLMs, abstractive summaries still suffer from faithfulness errors related to generating information that is not present in the original text (Ladhak *et al.*, 2022) and factual errors (Tang *et al.*, 2023; Falke *et al.*, 2019; Tang *et al.*, 2022). In general, factual accuracy of current generative language models is often lacking, making them particularly prone to hallucinations (Ji *et al.*, 2023; Koopman and Zuccon, 2023; Tang *et al.*, 2023). Despite advancements, CIS systems remain vulnerable to these limitations, emphasizing the need for research to improve system transparency, explainability, and reliability.

Challenges also arise from the nature of retrieved documents. In ad hoc retrieval, a document is deemed relevant if it contains at least one useful piece of information (Pavlu *et al.*, 2012), even if most of its content is unrelated or only vaguely related to the query. Research shows that unrelated text can severely hinder retrieval-augmented generation (RAG) systems (Cuconasu *et al.*, 2024), while evidence in the prompt, even if accurate, may degrade response accuracy for complex queries (Koopman and Zuccon, 2023). Additionally, language models often struggle to utilize long contexts effectively, with performance dropping significantly when relevant information is embedded in the middle of lengthy inputs (Koopman and Zuccon, 2023). These findings challenge the efficacy of basic retrieve-then-generate pipelines and highlight the importance of addressing irrelevant information in the generation process.

### Response Requirements

In CIS systems, conversational responses must meet key requirements to ensure the effectiveness and reliability of the system. A critical aspect is response grounding, which ensures response faithfulness to reliable sources (Gienapp *et al.*, 2024). Two primary dimensions for grounding are source attribution and verifiability. Source attribution evaluates the accuracy with which a generated response uses cited documents, facilitating easier verification of claims (Rashkin *et al.*, 2021). Verifiability requires every generated statement to be supported by citations, ensuring that responses are based on evidence that users can investigate (Liu *et al.*, 2023a). However, systems with high citation precision may produce responses lacking fluency, while more fluent responses risk misleading users due to unsupported claims.

The correctness of information included in the generated responses is tightly bound to source attribution. Evaluating factual correctness requires background knowledge and general familiarity with the topics. Therefore, the actual correctness of the responses does not always align with the perceived one.

Completeness is another vital response characteristic, especially for addressing non-factoid questions. Responses must cover the question's aspects comprehensively, including explanations, examples, and diverse opinions where applicable. Users consistently rate answers higher when they are not only correct but also complete, demonstrating that completeness contributes significantly to perceived quality (Bolotova *et al.*, 2020). The usefulness of responses depends also on their relevance and comprehensiveness. Answers should be detailed but concise, free of inconsistencies, and presented in a readable and unbiased manner. Features like serendipity—offering unexpected yet valuable information—and references to additional sources further enhance the perceived utility (Cambazoglu *et al.*, 2021). All these elements collectively define user satisfaction with CIS systems.

### Retrieval-Augmented Generation (RAG)

Recently proposed Retrieval-Augmented Generation (RAG) systems aim to address the problem of hallucinations and facilitate grounding of the generated responses, which is particularly important in knowledge-intensive conversational information-seeking tasks. They combine retrieval and generative processes to produce more factually correct and diverse outputs (Lewis *et al.*, 2020). RAG systems follow different approaches, each enhancing the integration of retrieval and generation. Naive RAG consists of a straightforward pipeline: indexing, retrieval, and generation. Advanced RAG builds on this by incorporating pre- and post-retrieval optimization strategies, retaining a sequential structure. Modular RAG further improves with flexible architectures, introducing functional modules and allowing iterative and adaptive retrieval. Iterative retrieval alternates between retrieval and generation to refine context, recursive retrieval decomposes complex queries into subproblems, and adaptive retrieval dynamically determines the necessity of external knowledge retrieval and when to terminate the process (Gao *et al.*, 2023b).

Generative processes in RAG systems are typically conditioned on the retrieved material. Evidence can be incorporated into prompts (Izacard and Grave, 2021; Shi *et al.*, 2024; Ram *et al.*, 2023) or attended to during inference through a learned textual knowledge retriever (Guu *et al.*, 2020). Retrieval-augmented generation models also integrate parametric and non-parametric memory to balance the retrieval and generation processes (Lewis *et al.*, 2020; Huang and Huang, 2024). While these approaches enhance the alignment of generated responses with retrieved evidence, challenges such as factual accuracy and the influence of irrelevant content persist.

Context curation plays a crucial role in RAG systems, as directly feeding all retrieved information to a language model (LLM) can lead to degraded performance. Redundant or lengthy contexts may result in issues like the "lost in the middle" problem, where LLMs struggle to utilize information effectively when it appears in the middle of a long text (Liu *et al.*, 2024). To address this, retrieved content often undergoes further refinement. Techniques such as prompt compression use small language models to remove unimportant tokens, producing a streamlined input that remains comprehensible to LLMs while reducing unnecessary complexity (Jiang *et al.*, 2024). Other approaches, like training information extractors or condensers through contrastive learning, focus on isolating and retaining essential information while discarding irrelevant details (Yang *et al.*, 2023; Xu *et al.*, 2023).

In addition to these RAG approaches, multi-stage frameworks have been developed to address the complexity of large-scale systems. For example, PisticRAG employs distinct stages, including matching, pre-ranking, ranking, reasoning, and aggregating, to refine the retrieval process and align it with LLM capabilities (Bai *et al.*, 2024). Similarly, the CompAct framework actively compresses retrieved documents into compact contexts, iteratively refining inputs until sufficient information is gathered to generate a complete response (Yoon *et al.*, 2024). Another enhancement leverages meta-knowledge to cluster documents into metadata-based sets of synthetic questions and answers, guiding query augmentation and retrieval for more personalized and effective responses (Mombaerts *et al.*, 2024). While promising, RAG systems face limitations in terms of transparency, often failing to identify low-confidence responses or address flaws arising from the incompleteness of retrieved content or biased generation processes.

### 2.2.3 CIS Response Evaluation

The evaluation of conversational responses involves addressing multiple dimensions that capture different aspects of response quality. While response *usefulness* remains the most commonly assessed dimension in human evaluation, its subjective nature makes it heavily dependent on the user's specific information needs. A utility model proposed by Gienapp *et al.* (2024) identifies coherence, coverage, consistency, correctness, and clarity as critical dimensions for evaluation, each encompassing distinct sub-dimensions. *Coherence* involves logical and stylistic organization of the response, shaping the manner in which information is conveyed. *Coverage* evaluates the breadth and depth of information, ensuring that the response addresses the user's needs comprehensively. It is closely related to response *completeness*, which captures the extent to which the question is answered. Statement-level *consistency* examines the relationship between generated content and its supporting sources (Rashkin *et al.*, 2021). *Correctness* emphasizes factual and topical accuracy, grounded in verifiability rather than absolute truth, and aligns with the broader concept of *faithfulness*, which measures the accuracy of generated responses against their sources (Falke *et al.*, 2019), ensuring the generation process with no misinformation (Es *et al.*, 2024). Proposed faithfulness evaluation metrics are based on entailment of the generated output in the source article (Falke *et al.*, 2019), BERT-based faithfulness classification (Kryscinski *et al.*, 2020), or factuality detection model (DAE) trained on word-, dependency- and sentence-level faithfulness annotations (Goyal and Durrett, 2021). The ultimate goal is increasing the faithfulness of the generated summary, while not increasing the extractiveness which is referred to as effective faithfulness (Ladhak *et al.*, 2022). The last dimension, *clarity*, focuses on the salience of information, requiring responses to prioritize essential content. Techniques like the Atomic Content Unit annotation protocol further enhance objectivity in assessing clarity by isolating individual facts for evaluation (Liu *et al.*, 2023b). Clarity is also closely linked to *relevance*, defined as how well the response addresses the subject of the question. Relevance has been identified as a key factor in answer utility, showing a strong correlation with how useful users perceive answers to be—especially in the context of non-factoid questions (Cambazoglu *et al.*, 2021).

#### Automatic Nugget-based Evaluation

The standard methodology for reproducible evaluation in traditional information retrieval research and development is offline evaluation (see Section 2.1.6).

This involves a document collection, user-reflective topics, and test systems (Sanderson, 2010). In generative IR, while this basic procedure remains relevant, an additional layer of complexity arises from the need to evaluate the generated response text alongside the ranked lists of retrieved document identifiers (Gienapp *et al.*, 2024). Therefore, the evaluation of CIS responses encompasses two primary aspects: retrieval-based and generation-based evaluation. Retrieval-based evaluation focuses on the system's effectiveness in retrieving relevant information to support generation tasks (see Section 2.1.6). In contrast, generation-based evaluation assesses the quality of the text produced by large language models, using measures such as linguistic quality and fluency with BLEU (Papineni *et al.*, 2001) and ROUGE-L (Lin, 2004), and overlap with ground-truth responses through metrics like Exact Match (Huang and Huang, 2024).

Responses provided by the system can be annotated holistically or segmented into smaller units for more granular analysis (Sakai, 2023; Gienapp *et al.*, 2024). Responses are often conceptualized as sequences of atomic statements, each optionally linked to sources of evidence. These statements are treated as discrete "atomic/semantic content units" (Nenkova *et al.*, 2007; Liu *et al.*, 2023b) or "information nuggets" (Pavlu *et al.*, 2012; Sakai, 2023), representing minimal units of relevant information that address user information needs (Gienapp *et al.*, 2024) (similarly to what has been proposed for nugget-based relevance evaluation). Nugget-based evaluation of responses is rooted in the TREC'03 QA track (Voorhees, 2004), where assessors identified relevant information nuggets across submissions and marked their presence in system outputs. This methodology has been adapted for modern RAG systems and validated during the TREC RAG'24 track, which introduced the AutoNuggetizer evaluation framework comprising two steps: nugget creation and nugget assignment (Pradeep *et al.*, 2024). In nugget creation, nuggets are formulated based on relevant documents and classified as either "vital" or "okay" (Voorhees, 2004). Traditionally, human assessors performed this task, but AutoNuggetizer automates this step by iterative LLM prompting. The second step, nugget assignment, involves assessing whether a system's response contains specific nuggets from the answer key. In the AutoNuggetizer framework, this step has also been automated using LLMs to match nuggets with system responses. Together, these processes ensure a fast and scalable evaluation of CIS system outputs.

### 2.2.4 CIS System's Limitations and Challenges

Despite rapid advancements and growing interest, CIS systems face numerous challenges and limitations. Among these are issues of unanswerability, where systems fail to fully address user queries or generate hallucinated content when no relevant context is found. While retrieval-augmented generation is a promising direction in mitigating hallucinations and extending knowledge beyond training data (Shuster *et al.*, 2021), multiple challenges persist. The conventional notion of relevance in retrieval is insufficient, as relevant passages may lack complete answers or need additional reasoning over multiple sources. Measures like answerability and completeness, although critical, remain difficult to evaluate automatically (Chen *et al.*, 2021; Pavlu *et al.*, 2012).

Out-of-scope queries further complicate CIS systems, as they often hinge on assumptions or conditions that make them unanswerable. Techniques like local entailment and semantic extraction aim to detect these mismatches between the semantic representations of the question with the context (Hu *et al.*, 2019; Huang *et al.*, 2019a), yet the field still lacks robust frameworks for addressing such queries comprehensively. Temporal considerations also pose significant challenges, as time-sensitive queries require parsing temporal information and understanding time-evolving facts (Allein *et al.*, 2021; Yang *et al.*, 2024). Temporal IR methods attempt to integrate document relevance with temporal relevance (Campos *et al.*, 2015), but assessing temporal validity is often challenging.

Bias in queries and the lack of viewpoint diversification introduce further limitations. CIS systems frequently retrieve information reflecting narrow perspectives, risking reinforcement of existing beliefs (Azzopardi, 2021). Approaches like search result diversification, stance detection, and fairness metrics aim to address these issues (Gao and Shah, 2020; Draws *et al.*, 2021a), but synthesis of this information into balanced responses remains an open challenge. A comprehensive view of the subject can be achieved by covering as many aspects as possible in the diversified set of results (Gao and Shah, 2020). However, in a conversational setup, we aim for the responses to be concise, resulting in a trade-off between conciseness and completeness of the generated answer.

Limiting the responses to a few sentences can result in a significant concealment of underlying details such as the ranking of results and specifics about the sources. These details are essential for users to assess the scope, novelty, reliability, and topical relevance of the provided information (Xu and Chen, 2006). Since the user is provided only with a short textual response as the final outcome of the generation process, it becomes the responsibility of the conversational system to identify and communicate any potential limitations to its users, ensuring transparency and empowering users to evaluate response quality. While the importance of explainability is broadly recognized for AI (Monroe, 2018) and has been extensively studied, for example, for decision support and recommender systems (Nunes and Jannach, 2017; Zhang and Chen, 2020), it has not received due attention for CIS systems.

Challenges described above are not new in the field of NLP or IA and different solutions have been proposed for unanswerability detection in QA (Sulem *et al.*, 2022; Choi *et al.*, 2018; Rajpurkar *et al.*, 2018; Reddy *et al.*, 2019) or search result diversification (Jiang *et al.*, 2018; Xu *et al.*, 2017; Tian *et al.*, 2019). However, effectively integrating these techniques into CIS systems to manage nuanced user information needs remains an open challenge.

## 2.3 Explainable AI

Explainability is a critical and active area of research in AI, particularly for CIS systems aiming to convey vast amounts of information in their responses. While these systems have made significant strides in generating and retrieving information, ensuring that users can understand and trust their outputs remains a major challenge. According to established human-AI interaction design guidelines (Amershi *et al.*, 2019), explainable systems should first clarify their capabilities, helping users grasp what the system can and cannot do. Next, they should provide transparency regarding performance, indicating how likely the system is to make mistakes. Finally, systems should offer explanations for their actions, enabling users to understand the reasoning behind the system's behavior. However, explanations can have both positive and negative effects. They may sometimes lead users to overtrust the system, even when it is wrong, or form inaccurate mental models of how the system operates (Cau *et al.*, 2023). This effect is particularly evident for users with low domain expertise, who may exhibit overconfidence in their decision-making or overreliance on the AI's suggestions.

### 2.3.1 Communicating Confidence and Capabilities

Various studies have explored how different explanation styles and interfaces impact user understanding and interaction with AI systems. Both interactive and "whitebox" explanations can improve user understanding of the system, with the interactive approach showing higher effectiveness at the cost of being time-consuming (Cheng *et al.*, 2019). The effectiveness of explainability also depends on factors like reasoning styles, and the presence of rationales or examples on user trust and satisfaction (Cau *et al.*, 2023; Tsai *et al.*, 2021). Depending on the task at hand and the domain of the system, explanations may come in different forms, informing the user about system confidence, capabilities, limitations of the system output, inner workings of the system, etc.

Confidence modeling has been extensively researched, with applications in tasks such as machine reading comprehension, machine translation, and speech recognition, as well as scenarios involving out-of-distribution inputs (Lakshminarayanan *et al.*, 2017; Niehues and Pham, 2019; Ovadia *et al.*, 2019). Systems can leverage predictive uncertainty to determine whether a question is answerable and abstain from generating responses when confidence is low to avoid the dissemination of unreliable content (Wang *et al.*, 2020). In conversational search, uncertainty estimates can guide decisions between asking clarifying questions and providing potential answers (Aliannejadi *et al.*, 2019; Penha and Hauff, 2021). Human studies have demonstrated that presenting confidence scores can help calibrate user trust in AI systems, though effective decision-making often requires users to bring complementary knowledge to address AI errors (Zhang *et al.*, 2020c). Advances in fine-tuning models to produce calibrated linguistic expressions of uncertainty can further enhance the interpretability of system confidence revealed to the user (Chaudhry *et al.*, 2024).

Communication of the limitations and capabilities of AI systems is another important component of explainability. Various methods, such as using natural language messages, additional user interface elements, or granular confidence scales, can help users understand the system's constraints and reliability of the provided output (Rechkemmer and Yin, 2022; Lu and Yin, 2021; Shani *et al.*, 2013). Confidence reporting, particularly in the context of recommender systems, enhances users' ability to make informed choices (Shani *et al.*, 2013). Research on human-AI interaction emphasizes that both the confidence display and the system's demonstrated accuracy influence users' trust and willingness to rely on AI predictions (Rechkemmer and Yin, 2022). Transparency of the system reliability cues presented to users is an enabler of trust judgments rather than a guarantee of trust itself (Liao and Sundar, 2022). User trust judgments often rely on heuristics rather than purely analytical judgments, leading to quick but potentially flawed decisions and underscoring the need for thoughtful interaction design and clear formulation of explanations (Liao and Sundar, 2022).

### 2.3.2 Evaluating Explanations

The evaluation of explanations in AI systems is multifaceted, focusing on metrics related to trustworthiness, transparency, and reliability. Trustworthiness is often assessed through dimensions such as perceived ability of the system (Toader *et al.*, 2019), desire to use (Tsai *et al.*, 2021), benevolence, and anthropomorphism (Toader *et al.*, 2019), which gauge users' confidence in the system's intentions and competence (Radensky *et al.*, 2023; Liao and Sundar, 2022). Transparency is measured by its impact on user awareness, perceived correctness, interpretability, and accountability. These measures capture how well users understand the system's processes, outputs, and fairness (Rader *et al.*, 2018). Explanations improve awareness of system operations but may not always help users evaluate the correctness or consistency of outputs (Rader *et al.*, 2018). Reliability assessments consider the interplay between human intuition, perceived competence, and reliance on AI systems. User's reliance may be influenced by their perceived competence or misconceptions about their own or the system's capabilities (He *et al.*, 2023b; Chen *et al.*, 2023).

## 2.4 Summary

This chapter has laid the theoretical and conceptual groundwork essential for the research presented in this thesis. By surveying the evolution of information access systems and key NLP tasks, it has provided the necessary context for understanding the challenges and motivations behind conversational search. Starting with classical and neural approaches to document retrieval, the chapter established a foundation for the CIS baseline system discussed in Chapter 3. It then introduced the principles of text representation and natural language modeling, emphasizing their importance in facilitating more natural user-system interactions. The chapter also delved into conversational search as a dynamic and user-centric process. It reviewed key tasks such as conversational response generation and the evaluation of generated responses in the context of RAG. These discussions directly inform the research presented in Chapters 4–6. Finally, we addressed the growing importance of explainability in AI, particularly within information access systems. By exploring existing approaches to developing and evaluating explainable components, we provided the necessary background for the contributions in Chapter 7, which investigates strategies for making CIS responses more transparent.

# Part I

Understanding CIS Limitations

# Chapter 3

# Establishing a Baseline

*One learns from books and example only that certain things can be done. Actual learning requires that you do those things.*

— **Frank Herbert**

The last few years have seen an acceleration of research on multi-turn, natural language, and long-term user modeling capabilities of search systems with an attempt to make them more conversational (Zamani *et al.*, 2023). The Conversational Assistance Track at the Text REtrieval Conference (TREC CAsT) (Dalton *et al.*, 2019, 2020, 2021) has been a key enabler of progress in this area, by providing a reusable test collection for conversational search. The task at TREC CAsT is to identify relevant content from a collection of passages, "for conversational queries that evolve through a trajectory of a discussion on a topic" (Dalton *et al.*, 2021) (see Section 2.2.1). Over the years, query rewriting, passage retrieval, and passage reranking have emerged as the main components, which are combined in a pipeline architecture. Clearly, the ranking components can directly benefit from advances in dense/hybrid passage retrieval (Luan *et al.*, 2021), and are indeed critical to overall system performance. However, what makes the task interesting from a conversational perspective, and different from passage retrieval, is the problem of query rewriting (Kumar and Callan, 2020; Lin *et al.*, 2021; Vakulenko *et al.*, 2021c; Yu *et al.*, 2020; Mele *et al.*, 2020).

In this chapter, we lay a solid foundation for the exploration and development of advanced features and components of conversational information-seeking (CIS) systems, which are covered in later chapters, by establishing a strong CIS baseline for retrieval. Specifically, we address the following research question: **What are strong baselines for passage retrieval in CIS systems? (RQ1a).**[1] It has been shown that the best-performing systems at TREC form a very competitive reference point for effectiveness comparison (Armstrong *et al.*, 2009). This means that even if one's ultimate research interest lies in query rewriting, demonstrating strong absolute performance for conversational search requires a high degree of effectiveness from all system components. Our main objective is to reproduce (1) the best-performing baseline provided by the track organizers (Dalton *et al.*, 2021) and (2) the top-performing (documented) system (Yan *et al.*, 2021) from the 2021 edition of TREC CAsT.[2] These two approaches are seen as representatives of a strong baseline and the state of the art, respectively. It is worth noting that the system description papers accompanying TREC submissions are not peer-reviewed and there is no explicit or implicit reproducibility requirement making reproducibility particularly challenging and a study such as this particularly insightful.

Both selected systems follow a two-stage *retrieve-then-rerank* pipeline architecture with queries rewritten based on conversational context. The baseline system (Dalton *et al.*, 2021) uses a fine-tuned query rewriting model fine-tuned on CANARD (Elgohary *et al.*, 2019), first-pass retrieval based on BM25, and a pointwise reranker. The top participating system (Yan *et al.*, 2021) uses a different dataset for fine-tuning the query rewriting model and employs more advanced ranking components: a combination of sparse-dense retrieval with pseudo relevance feedback for first-pass retrieval, and pointwise/pairwise reranking. Since the two selected systems follow the same basic two-stage retrieval pipeline, we perform additional experiments in order to better understand how each pipeline component contributes to overall effectiveness. To shed light on the generalizability of findings, we report results on both the 2020 and 2021 editions of TREC CAsT. Since the query rewriter influences the effectiveness of both first-pass retrieval and reranking, we also perform experiments using a different retrieval pipeline, which can utilize different query rewriting methods for the two ranking stages. The reproducibility process offers valuable insights and highlights the challenges of replicating systems submitted to TREC CAsT. The experiments with different pipeline architectures and query rewriters help us to better understand the main factors contributing to system performance.

All resources developed within this study (source code, runfiles) are available under: `https://github.com/iai-group/ecir2023-reproducibility`.

[1] Baselines for response generation (RQ1b) are discussed in Section 6.3.2.

[2] We refer to the `clarke-cc` run by the WaterlooClarke group as the top-performing system in this chapter. Note that the `mono-duo-rerank` run, submitted by the h2oloo group, achieves higher performance (Dalton *et al.*, 2021). However, it is not accompanied by a system description, making reproducibility impossible.

This chapter is based on the following paper:

> Łajewska and Balog (2023a): *From Baseline to Top Performer: A Reproducibility Study of Approaches*, ECIR '23

## 3.1 CAsT Systems Overview

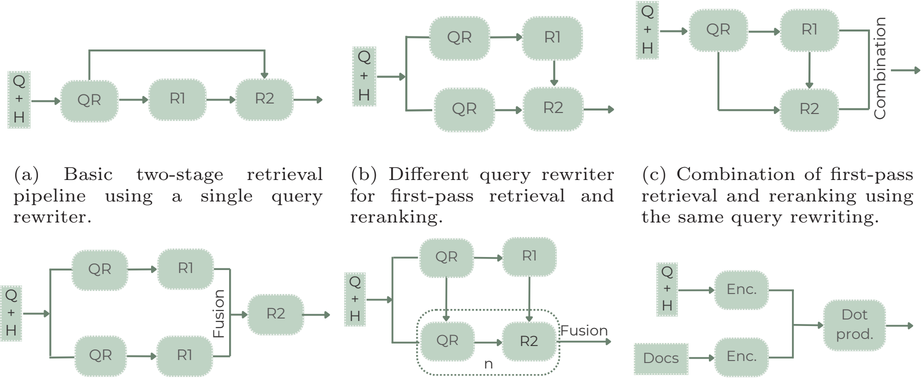
We provide an overview of query rewriting approaches and ranking architectures used at TREC CAsT (see Section 2.2.1 for more details about the track).

### 3.1.1 Query Rewriting

The goal of query rewriting is to handle common conversational phenomena such as omission, coreference (Dalton *et al.*, 2019), zero anaphora, topic change, and topic return (Voskarides *et al.*, 2020). Approaches can be broadly categorized into unsupervised, supervised feature-based, and (weakly-)supervised neural methods. Unsupervised query rewriting methods expand the original query with terms from the conversation history, for example, from previous utterances based on BM25 score (Yilmaz *et al.*, 2019), cosine similarity (Voskarides *et al.*, 2019), or other frequency-based signals (Lin *et al.*, 2021). Supervised feature-based methods use linguistic features based on dependency parsing, coreference resolution, named entity resolution, or part-of-speech tagging (Mele *et al.*, 2020). Supervised neural query rewriting approaches utilize large pre-trained language models and in particular generative models such as GPT-2 (Vakulenko *et al.*, 2021a) or T5 model (Yan *et al.*, 2021; Ju *et al.*, 2021; Chang *et al.*, 2020). These models are fine-tuned on a conversational dataset, such as CANARD (Vakulenko *et al.*, 2021a; Lin *et al.*, 2021; Ju *et al.*, 2021; Chang *et al.*, 2020; Vakulenko *et al.*, 2021c) or QReCC (Yan *et al.*, 2021) (see Section 2.2.1). The generated query reformulations may further be expanded with terms from conversation history (Vakulenko *et al.*, 2021a), with paraphrases (Ju *et al.*, 2021), or related sentences from semantically related documents (Chang *et al.*, 2020). Weakly supervised neural query rewriting methods aim to fine-tune large pre-trained language models (Yu *et al.*, 2020) or term selection classifiers (Kumar and Callan, 2020) with weak supervision data that is created using rule-based or self-supervised approaches. The best results are reported using a combination of term-based query expansion with generative models for query reformulation (Kumar and Callan, 2020; Lin *et al.*, 2021; Vakulenko *et al.*, 2021a).

### 3.1.2 Pipeline Architectures

Systems participating in TREC CAsT exhibit a wide variety of approaches, not only in terms of component-level choices but also in terms of the overall architectures of their ranking pipelines. The most common choice is a two-stage retrieval pipeline with a query rewriting module. Different variants of this cascade architecture include systems with the same rewriting method used for both first-pass retrieval and reranking (Chang *et al.*, 2020; Vakulenko *et al.*, 2021b; Yan *et al.*, 2021; Ju *et al.*, 2021; Vakulenko *et al.*, 2021a; Yu *et al.*, 2020; Vakulenko *et al.*, 2021c; Mele *et al.*, 2020) (Figure 3.1a), different query rewriting modules for both stages (Yang *et al.*, 2019) (Figure 3.1b), or using



(a) Basic two-stage retrieval pipeline using a single query rewriter.

(b) Different query rewriter for first-pass retrieval and reranking.

(c) Combination of first-pass retrieval and reranking using the same query rewriting.

(d) Reranking of fused first-pass results that use different query rewriters.

(e) Fusion of multiple passage rerankings using different rewrites.

(f) Few-shot conversational dense retrieval.

Figure 3.1: Pipeline architectures for conversational search (**Q+H**: raw query and conversational history; **QR**: query rewriter; **R1**: first-pass retriever; **R2**: reranker; **Enc.**: encoder; **Docs**: document collection; **Dot prod.**: dot product).

rewriting only for first-pass retrieval (Gemmell and Dalton, 2020; Yang *et al.*, 2019).

More advanced architectures may use a two-stage retrieval pipeline with the same query rewriter for each stage, but combine the scores obtained from retrieval and reranking to produce a final ranking (Voskarides *et al.*, 2019) (Figure 3.1c) or use two different versions of the rewritten query for first-pass retrieval and a fusion of the ranked lists for the reranking stage (Lin *et al.*, 2021) (Figure 3.1d). Another architecture variant consists of first-pass retrieval using the rewritten query, followed by a fusion of multiple contextualized passage reranking of several different rewrites (Kumar and Callan, 2020) (Figure 3.1e). An alternative to the retrieve-then-rerank approach is a few-shot conversational dense retrieval system that learns contextualized embeddings of utterances and documents in the collection, and scores documents solely using the dot product of the embeddings (Yu *et al.*, 2021) (Figure 3.1f).

## 3.2 Selected Approaches

We present the two approaches from TREC CAsT'21 that we aim to reproduce in this paper: (1) the best-performing official baseline provided by the track organizers' and (2) the top-performing documented system submitted by participants. These approaches may be regarded as representatives of a strong baseline and of the state of the art, respectively. Both may be seen as instantiations of the basic two-stage retrieval pipeline approach (cf. Figure 3.1a), with query rewriting, first-pass retrieval, and reranking components, as shown in Table 3.1. In this section, we focus on a high-level description of these approaches,

Table 3.1: Overview of approaches reproduced in this chapter.

| | **Query rewriting** | **First-pass retrieval** | **Reranking** |
|---|---|---|---|
| BaselineOrganizers | T5 fine-tuned on CANARD | BM25 | monoT5 |
| WaterlooClarke | T5 fine-tuned on QReCC | BM25 with PRF + ANCE | mono/duoT5 |

based on the corresponding TREC papers; specific implementation details are discussed in Section 3.3.

### 3.2.1 Organizers' Baseline

Of the several baselines provided by the track organizers, `org_auto_bm25_t5` was the best-performing run (Dalton *et al.*, 2021); this will be referred to as the **BaselineOrganizers** approach henceforth. The query rewriting component is using T5 fine-tuned on CANARD for generative query rewriting. The rewriter uses all previous queries and the three previous canonical responses as context. For first-pass retrieval, BM25 is used to collect the top 1000 documents from the collection. The documents are reranked with a pointwise (mono) T5 model trained on MS MARCO (Campos *et al.*, 2016).

### 3.2.2 Top Performer: WaterlooClarke

The top-performing documented system was the `clarke-cc` run by Yan *et al.* (2021); this will be referred to as the **WaterlooClarke** approach henceforth. The query rewriting component is based on a T5 model that is fine-tuned on the QReCC dataset (Anantha *et al.*, 2021). For context, the rewriter uses previously rewritten utterances and the last canonical result. First-pass retrieval comprises two sub-components: a sparse and a dense retriever. The sparse retriever utilizes a BM25 with pseudo-relevance feedback (PRF), with the parameters tuned to maximize recall. PRF is run over both the target corpus and the C4 corpus.[3] The dense retriever is based on the ANCE approach (Xiong *et al.*, 2020). Both retrieval systems return the top 1000 documents that are merged into one final ranking. Reranking is performed using a pointwise T5 reranker, followed by another reranking of the top 50 documents, using pairwise duoT5 (Pradeep *et al.*, 2021).

## 3.3 Reproducibility Experiments

In this section, we seek to answer the question: Can the organizers' baseline and the best performing system at the TREC CAsT'21 be reproduced? We describe the implementation details of the two systems and discuss their end-to-end

[3] `https://huggingface.co/datasets/allenai/c4`

performance with respect to the results reported in the track overview (Dalton *et al.*, 2021).

### 3.3.1 Baseline Implementation

We base the implementation solely on the description of the track organizers' `org_auto_bm25_t5` baseline in the overview paper (Dalton *et al.*, 2021), without resorting to additional communication with the authors.

The passage collection is indexed using Elasticsearch, using the built-in analyzer for tokenization, stopwords removal, and KStem stemming. For query rewriting, we use a pre-existing T5 model that has been fine-tuned on the CANARD dataset (`castorini/t5-base-canard`).[4] Our implementation is based on the Hugging Face transformers library.[5] According to (Dalton *et al.*, 2021), the context for the query rewriter is of the form:

$$q_1, q_2, \ldots, q_{i-3}, r_{i-3}, q_{i-2}, r_{i-2}, q_{i-1}, r_{i-1}, q_i,$$

where $q_i$ and $r_i$ are the $i$th raw query and canonical response, respectively. Contexts exceeding the allowed model input length are not handled. This, however, can result in trimming the input in a way that the raw query that is to be rewritten is removed. To increase the quality of the rewriting by ensuring the correct form of the input and benefiting from previous rewrites, we alternatively use:

$$\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_{i-1}, trim(r_{i-1}), q_i,$$

where $\hat{q}_i$ is the $i$th rewritten query and *trim* is a function that cuts the canonical response if the length of the input is longer than the capacity of the model. For first-pass retrieval, the passages are ranked using BM25 on a `catch_all` field (concatenating the `title` and `body` fields) in the 2021 index and on the `body` field for the 2020 index. We initially used the parameters reported by the organizers (k1=4.46, b=0.82), but then achieved better results with the default parameters (k1=1.2, b=0.75). The top 1000 candidates for each turn are reranked using the T5 model introduced by Nogueira *et al.* (2020), which has been published on Hugging Face (`castorini/monot5-base-msmarco`).[6]

### 3.3.2 WaterlooClarke Implementation

We base our implementation on the WaterlooClarke group's TREC paper (Yan *et al.*, 2021). Additional information on specific details was obtained from the authors via email communication and inferred from the implementation made available.[7]

[4] https://huggingface.co/castorini/t5-base-canard

[5] https://github.com/huggingface/transformers

[6] https://huggingface.co/castorini/monot5-base-msmarco

[7] https://github.com/claclark/Cottontail/blob/main/apps/treccast21.cc

The approach requires two indices: an approximate nearest neighbor (ANN) index for ANCE dense retrieval and an inverted index for BM25. The authors use ANCE's own implementation[8] and a publicly released model checkpoint (passage ANCE(FirstP)) for the ANN index.[†] We use Pyterrier's plugin[9] for creating the ANN index, which is based on the original paper, and allows for easier integration with other modules in our pipeline. For building the ANN index we use MS MARCO Passage and TREC CAR collections provided by the ir_datasets package,[10] and implement our own generator for the WaPo 2020, MS MARCO Documents, and KILT collections. No additional preprocessing is performed when building the dense retrieval index. The inverted index used by BM25 is the same as in Section 3.3.1.

The query reformulation step in WaterlooClarke is based on a T5 model trained on the QReCC dataset (Anantha *et al.*, 2021). All the previous rewritten utterances and the canonical response for the last utterance are used as context to reformulate the current question, i.e., the input is given as:

$$\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_{i-1}, trim(r_{i-1}), q_i.$$

If the length of the input sentence exceeds 512, the answer passage is cut off.[†] The authors fine-tune a pretrained `t5-base` model[11] with the training partition of the QReCC dataset for 3 epochs, using the original test partition as a validation set.[†] The train batch size is equal to 2 and the learning rate is $5 \times 10^{-5}$.[†] We use the Simple Transformers library[12] for the fine-tuning procedure (as opposed to PyTorch Lightning[13] and Hugging Face transformers used by the authors[†]).

There are two first-pass rankers involved: (1) sparse retrieval using BM25 with pseudo relevance feedback (PRF) and (2) dense retrieval using ANCE (Xiong *et al.*, 2020). The final sparse retrieval ranking is a fusion of two rankings.[†] PRF is applied on the top 17 documents to expand the query with the top 26 terms; the expanded query is then scored using BM25 to generate the first sparse ranking. Additionally, the authors use the top 16 weighted answer candidates generated by a statistical question-answering method ran against the C4 corpus to create the second ranking (answer candidates are used by BM25).[†] The first and the second ranking produced by the sparse retrieval are fused with Reciprocal Rank Fusion (RRF) (Cormack *et al.*, 2009).[†] There is no further information disclosed about the question-answering system used (neither in the paper nor in the GitHub repository). Therefore, we skip the second ranking in reproducibility and focus on standard BM25 with PRF. The BM25 parameters

[8] https://github.com/microsoft/ANCE

[†] Missing information provided by the authors in personal communication.

[9] https://github.com/terrierteam/pyterrier_ance

[10] https://github.com/allenai/ir_datasets

[11] https://huggingface.co/t5-base

[12] https://simpletransformers.ai/

[13] https://www.pytorchlightning.ai/

Table 3.2: Reproducibility experiments on the TREC CAsT'21 dataset. The results from Dalton *et al.* (2021) are indicated with [*].

| Approach | R@500 | MAP | MRR | NDCG | NDCG@3 |
|---|---|---|---|---|---|
| BaselineOrganizers@TREC'21 [*] | 0.6360 | 0.2910 | 0.6070 | 0.5040 | 0.4360 |
| BaselineOrganizers-QR-BM25 | 0.5632 | 0.2268 | 0.4947 | 0.4317 | 0.3457 |
| BaselineOrganizers-BM25 | 0.5894 | 0.2546 | 0.5405 | 0.4672 | 0.3966 |
| BaselineOrganizers | 0.6472 | 0.2628 | 0.5354 | 0.4885 | 0.3968 |
| WaterlooClarke@TREC'21 [*] | 0.8690 | 0.3620 | 0.6840 | 0.6400 | 0.5140 |
| WaterlooClarke@TREC'21 (runfile) | 0.8534 | 0.3494 | 0.6626 | 0.6240 | 0.4950 |
| WaterlooClarke reproduced by us | 0.6915 | 0.2864 | 0.5712 | 0.5176 | 0.4151 |

are tuned to maximize recall over manually rewritten questions from previous years. The exact details of this remain unclear. We tune BM25 parameters on our 2020 and 2021 indices and take the average of the best parameters found for each year (b=0.45, k1=0.95) since the parameters used in their code (b=0.45, k1=1.18) gave worse results on our indices. For query expansion, since the choice of PRF algorithm could not be resolved, we opted for RM3 (Lavrenko and Croft, 2001), which we implemented from scratch.

The results of sparse and dense retrieval are fused to generate the final set of 1000 candidate passages for reranking. Since the fusion method is not stated in the paper, we assume that this step also employs RRF; we utilize the TrecTools library,[14] which implements a RRF as defined in (Cormack *et al.*, 2009).

The reranking stage in this approach is based on a pointwise monoT5 reranker (on all candidate passages), followed by a pairwise duoT5 reranker (on the top 50 passages reranked by monoT5). The original reranking implementation is based on the Pyaggle library[15] with the default model checkpoints. Our implementation of duoT5 is based on the Hugging Face transformers library and the `castorini/duot5-base-msmarco` model published on Hugging Face.[16]

### 3.3.3 Results

Table 3.2 reports our results on the CAsT'21 collection. Following the official setup, we consider measures with both binary and graded relevance. The main measure is NDCG@3; other measures are computed with a rank cutoff of 500. For binary measures, we apply a relevance threshold of 2.

For the baseline, the results reported in the overview paper (Dalton *et al.*, 2021) are included verbatim and regarded as the reference, since the raw runfile (`org_auto_bm25_t5`) is not available in the TREC archive. We include results using the original query rewriting method and reported BM25 param-

[14] https://github.com/joaopalotti/trectools

[15] https://github.com/castorini/pygaggle

[16] https://huggingface.co/castorini/duo5-base-msmarco

eters (BaselineOrganizers-QR-BM25), using the improved query rewriter while keeping the reported BM25 parameters (BaselineOrganizers-BM25), and finally using the improved query rewriter with default BM25 parameters (BaselineOrganizers). We find that the latest variant performs best; it is still 9% below the reference result in terms of NDCG@3, but 2% better in terms of Recall@500.

Regarding WaterlooClarke, the performance of our reproduced system is 19% lower in terms of NDCG@3 and 20% lower in terms of Recall@500 than the official results reported for this approach. The discrepancy in the results is most likely caused by the lack of the C4-based question-answering step performed in first-pass retrieval. This element of the system is not sufficiently described in the paper nor has been resolved via personal email communication. Surprisingly, we observe discrepancies between the official results reported in the overview paper and a direct evaluation of the `clarke-cc` runfile taken from the TREC archive (cf. rows 5 vs. 6 in Table 3.2). The latter results are lower, with a relative drop of almost 4% in NDCG@3, which is a non-negligible difference. We cannot explain this discrepancy; however, it also puts into question the results reported in the track overview. When comparing our reproduced results against their runfile, the relative differences are under 16% and 19% in terms of NDCG@3 and Recall@500, respectively.

Overall, according to the track overview paper, the relative differences between BaselineOrganizers and WaterlooClarke are 18% and 37% in terms of NDCG@3 and Recall@500, respectively (cf. rows 1 vs. 5 in Table 3.2). However, the respective differences in our reproduced approaches are 5% and 7% (cf. rows 4 vs. 7 in Table 3.2). Moreover, these differences are no longer statistically significant, based on a paired t-test with $p < 0.05$. The same test does indicate significant differences when performed against the WaterlooClarke runfile.

### 3.3.4 Summary

In summary, neither approach could be fully reproduced due to key information missing. In the case of BaselineOrganizers, the specifics of the models used for query rewriting and reranking were lacking, and the formulation of input sequences for query rewriting was underspecified (esp. with regards to exceeding the length limits of the model). As for WaterlooClarke, the complexity of the system and shortages in technical details made it impossible to fully implement the system. Most notably, the involvement of a question-answering system for sparse retrieval is not even mentioned in the paper. We do want to acknowledge the kind, helpful, and open communication by the authors via email, which allowed us to resolve questions around the query rewriting model and its parameters, the BM25 and PRF parameters used, and the rank fusion method employed. Nevertheless, after several rounds of email exchanges, we are still missing details about the PRF algorithm, the question-answering system employed, the exact approach used for tuning the BM25 parameters, the preprocessing employed for the inverted index, and the method used for combining sparse and dense rankings. It is also worth noting that while BM25 parameters were shared for both approaches, those parameters were not the optimal ones

Table 3.3: Variants of a two-stage retrieval pipeline on TREC CAsT'20 and '21. T5_C and T5_Q indicate T5-based query rewriters trained on CANARD and QReCC datasets, respectively. m/dT5 stands for mono/duoT5 reranker.

**TREC CAsT'20**

| Approach | R@1000 | MAP | MRR | NDCG | NDCG@3 |
|---|---|---|---|---|---|
| T5_C + BM25 + monoT5 | 0.5276 | 0.2191 | 0.5457 | 0.4353 | 0.3789 |
| T5_Q + BM25 + monoT5 | 0.5100 | 0.2056 | 0.5106 | 0.4065 | 0.3618 |
| T5_C + ANCE/BM25 + m/dT5 | 0.6781 | 0.2540 | 0.5512 | 0.5027 | 0.4052 |
| T5_Q + ANCE/BM25 + m/dT5 | 0.6449 | 0.2443 | 0.5357 | 0.4804 | 0.4061 |
| T5_C + ANCE/BM25/PRF + m/dT5 | **0.6878** | **0.2555** | **0.5541** | **0.5063** | **0.4086** |
| T5_Q + ANCE/BM25/PRF + m/dT5 | 0.6608 | 0.2451 | 0.5355 | 0.4840 | 0.4052 |

**TREC CAsT'21**

| Approach | R@500 | MAP | MRR | NDCG | NDCG@3 |
|---|---|---|---|---|---|
| T5_C + BM25 + monoT5 | 0.6472 | 0.2628 | 0.5354 | 0.4885 | 0.3968 |
| T5_Q + BM25 + monoT5 | 0.6018 | 0.2530 | 0.5369 | 0.4670 | 0.3933 |
| T5_C + ANCE/BM25 + m/dT5 | 0.7259 | 0.2886 | 0.5575 | 0.5316 | 0.4068 |
| T5_Q + ANCE/BM25 + m/dT5 | 0.6799 | 0.2843 | 0.5702 | 0.5135 | **0.4159** |
| T5_C + ANCE/BM25/PRF + m/dT5 | **0.7306** | **0.2915** | 0.5573 | **0.5330** | 0.4061 |
| T5_Q + ANCE/BM25/PRF + m/dT5 | 0.6915 | 0.2864 | **0.5712** | 0.5176 | 0.4151 |

for us, which is likely due to differences in document preprocessing. It, however, means that BM25 parameters alone, without further details on preprocessing or collection statistics, are only moderately useful. We shall reflect more generally on some of these challenges and possible remedies in Section 3.6.

## 3.4 Additional Experiments

We have reproduced two approaches, BaselineOrganizers and WaterlooClarke, which follow the same basic two-stage retrieval pipeline (cf. Figure 3.1a), but differ in each of the query rewriting, first-pass retrieval, and reranking components. We experiment with different configurations of this basic pipeline to understand which changes contribute most to overall performance (Section 3.4.1). Additionally, we consider a different pipeline architecture (Section 3.4.2). In both sets of experiments, we are interested in the generalizability of findings, therefore we also report results on the TREC CAsT'20 dataset. (Note that the rank cut-off for the 2020 collection is 1000, while for 2021 it is 500.)

### 3.4.1 Variants of a Two-Stage Retrieval Pipeline

In this experiment, we gradually switch out the components of a baseline system (BaselineOrganizers) with components of a state-of-the-art system (Water-

Table 3.4: Performance of query rewriting approaches with different variants of the two-stage pipeline on the TREC CAsT'20 and '21 datasets. The highest scores for each year are in boldface.

| | | **R2** | | | |
|---|---|---|---|---|---|
| | | **T5_CANARD** | | **T5_QReCC** | |
| | | **Recall** | **NDCG@3** | **Recall** | **NDCG@3** |
| **R1** | **T5_CANARD** | 2020: **0.6878**<br>2021: **0.7306** | 2020: **0.4086**<br>2021: 0.4061 | 2020: **0.6878**<br>2021: 0.7267 | 2020: 0.3923<br>2021: 0.4166 |
| | **T5_QReCC** | 2020: 0.6608<br>2021: 0.6879 | 2020: **0.4086**<br>2021: **0.4176** | 2020: 0.6608<br>2021: 0.6915 | 2020: 0.4052<br>2021: 0.4151 |

looClarke). The results are presented in Table 3.3; the first and last rows within each block correspond to BaselineOrganizers and WaterlooClarke, respectively. Our observations are as follows. First, when changing the dataset used for training the T5-based query rewriter from CANARD to QReCC (rows 1 vs. 2, 3 vs. 4, and 5 vs. 6 in Table 3.3) we observe a noticeable drop (3%–7%) in terms of recall, with smaller differences in NDCG@3 (below 2%, with one exception). Second, using more advanced retrieval methods (ANCE/BM25 instead of BM25 for first-pass ranking and mono/duoT5 instead of monoT5 for reranking; rows 1 vs. 3 and 2 vs. 4 in Table 3.3) does yield consistent improvements across metrics and datasets: +12%–29% in recall and +3%–12% in NDCG@3. Finally, using pseudo relevance feedback for first-pass retrieval (rows 3 vs. 5 and 4 vs. 6 in Table 3.3) results in small but consistent improvements in terms of recall (1%–2%) with negligible differences in NDCG@3 (<1%). It should be noted that none of the above differences are statistically significant, thereby the results are merely indicative. However, in terms of overall trends, our results are in line with the tendencies reported by Yan *et al.* (2021). Namely, that adding PRF and combining sparse and dense retrieval methods for first-pass retrieval improves performance.

### 3.4.2 Using a Different Pipeline Architecture

It is clear that query rewriting has a direct impact on both ranking steps: first-pass retrieval (R1) and reranking (R2). Still, it remains to be seen whether the two stages are impacted in the same way. The basic two-stage retrieval pipeline (cf. Figure 3.1a) uses the same query rewriter for both ranking stages and therefore cannot be used to answer this question. We thus switch to a different pipeline architecture—one that uses a different query rewriter component for R1 and R2, but is identical to the basic pipeline in the ranking components (cf. Figure 3.1b).

Table 3.4 presents the results for the possible four-way combinations of query rewriters, T5_CANARD and T5_Q, and ranking stages, R1 and R2. The ranking components follow the WaterlooClarke approach (i.e., using T5_QReCC for both R1 and R2 corresponds to the last row in Table 3.3). The results reveal some interesting tendencies that generalize across both datasets (even though the differences are not statistically significant). Using T5_CANARD for first-pass retrieval results in the highest recall. However, the overall best combination in terms of final ranking (NDCG@3) is when T5_QReCC is employed in first-pass retrieval and T5_CANARD is used in reranking. Overall, we observe meaningful relative improvements for recall (up to 6%) and negligible improvements for NDCG@3 (≤1%) on both datasets over the WaterlooClarke approach.

## 3.5 Reflections on Reproducibility of TREC Systems

TREC papers can range anywhere from vague system descriptions to full-fledged research papers, which can make reproducibility a real challenge; this has certainly been the case for this study. We acknowledge that reproducibility is not a requirement for TREC submissions. Still, since they are often used for reference comparison in terms of absolute system performance on a given test collection, cf. (Armstrong *et al.*, 2009), it is worth considering how easy or difficult it is to reproduce them. Specifically, we have selected two approaches for our study: the best performing baseline by the track organizers and the best performing participant submission (that was accompanied by a paper) from the 2021 edition of TREC CAsT. We have decided against personal communication with the track organizers (thus implicitly subjecting them to a higher virtual bar-of-standard) while making the best effort to resolve any missing details with the participant team over email.

Generally, key missing information includes the names of specific algorithms and models used and detailed-enough descriptions of procedures of constructing inputs to neural models and ways of obtaining models' parameters. We wish to note that sharing model parameters in some cases is not enough; consider, e.g., the simple case of BM25, where the length normalization parameter alone is not meaningful if collection statistics markedly differ due to how the collection is preprocessed. Given that multi-stage ranking architectures are common at TREC CAsT, but also beyond that, sharing intermediate results from the different components would be immensely valuable. These could include rewritten or expanded queries, a set of candidate document IDs, and intermediate document rankings. Sharing them would not only support reproducibility but also facilitate component-level evaluation.

We attempted to clarify the discrepancies between the results in this paper and those reported in the track overview via email communication with the track organizers. There is a difference in tooling: they used Pyserini[17] for building the index, while we used Elasticsearch. Differences in collection preprocessing

[17] https://github.com/castorini/pyserini

(tokenization, stemming, stopword removal, etc.) may contribute to the gap in the results. Regarding the runfile, we were pointed to the track's GitHub repository[18] containing the raw runfile (`org_automatic_results_1000.v1.0.run`). However, evaluating this runfile against the official qrels still yields results different from those reported in the track overview paper (in parentheses): Recall@500 is 0.623 (vs. 0.636), MAP is 0.282 (vs. 0.291), and NDCG@3 is 0.424 (vs. 0.436). This is "in alignment" with the case of the WaterlooClarke (`clarke-cc`) runfile, in the sense that there is a mismatch between the numbers reported in the track overview paper and the evaluation of the actual runfiles (with the latter being lower).

## 3.6 Conclusions

In this chapter, we have attempted to reproduce approaches for the task of conversational passage retrieval, in the context of TREC CAsT. Overall, our reproducibility efforts have met with moderate success. Surprisingly, we have managed to come closer to reproducing the participant's submission (WaterlooClarke) than the organizers' baseline. In the case of the former, there is a missing sparse retrieval component that can well explain the difference. As for the organizers' results, the discrepancies between the reported results in the track overview paper and the actual runfiles found in the TREC archive would be worth a follow-up investigation.

Since both reproduced systems follow the same basic two-stage retrieval pipeline, we conducted additional experiments to explore different configurations of this pipeline. Our results align with previous research (Yan *et al.*, 2021), demonstrating that more advanced retrieval models consistently improve performance across metrics and datasets, while incorporating relevance feedback in the first-pass retrieval yields small but consistent gains in recall and NDCG@3. Additionally, we experimented with various combinations of query rewriting methods within a different retrieval pipeline, showing that applying different methods at different stages can be beneficial. In answer to **RQ1a** *(What are strong baselines for passage retrieval in CIS systems?)*, we have found that a combination of sparse and dense retrieval with pseudo relevance feedback for first-pass retrieval and pointwise/pairwise for reranking preceded by a fine-tuned query rewriting component represents a strong baseline for the conversational passage retrieval system. With a strong retrieval baseline established, the next step is to explore techniques that transition the system's output from ranked passages to a truly conversational, coherent, and informative natural language response.

[18] https://github.com/daltonj/treccastweb/tree/master/2021/baselines

# Chapter 4

# Limitations of CIS Systems

*Good judgment comes from experience, and experience - well, that comes from poor judgment.*

— **Alan Alexander Milne**

Conversational information-seeking (CIS) research currently centers on retrieval components, such as passage retrieval, reranking, and query rewriting (see Chapter 3). However, the core difficulty lies in effectively assembling the retrieved information into a trustworthy and reliable conversational response that the user will ultimately interact with. The task of synthesizing information from the top retrieved passages into a single response is called *conversational response generation* (Ren *et al.*, 2021). Unfortunately, generated responses are susceptible to limitations, including hallucinations when no answer is found (Ji *et al.*, 2023), biased responses only partially answering the question (Gao and Shah, 2020), or factual errors (Tang *et al.*, 2023). These limitations potentially lead to inaccuracies, pitfalls, and biases, which may not always be evident to users, particularly those who lack familiarity with the search topic or the necessary background knowledge. As individuals without specific training can only distinguish between human-generated and auto-generated texts at a level close to random chance (Clark *et al.*, 2021), factually incorrect, unsupported, biased, or incomplete information may be easily overlooked.

This chapter investigates users' ability to recognize pitfalls in CIS systems related to *query answerability* and *response incompleteness* by addressing the following research question: **Which limitations in the responses are de-**

**tectable by users? (RQ2.1)**; see Table 4.1 for illustrative examples. We hypothesize that untrained users cannot identify these problems in CIS interactions. More specifically, we aim to answer the following questions:

- **RQ2.1a:** Can users effectively recognize the problem of *query answerability* and the problem of multiple viewpoints leading to *response incompleteness* in system responses?

- **RQ2.1b:** How do inaccurate, incomplete, and/or biased responses impact the user experience?

We design and conduct two crowdsourcing-based studies to determine whether users can effectively recognize these two problems in responses based on a subset of topics from the TREC Conversational Assistance (CAsT) datasets (Dalton *et al.*, 2020; Owoicho *et al.*, 2022) with inaccuracies or biases manually injected in a controlled manner.

Query answerability can be defined at different levels, which includes determining whether the answer is present within the top relevant passages, the entire corpus, or general world knowledge. Additionally, when "no answer found" is the outcome, the system must transparently reveal this to the user and suggest ways to continue the conversation. In this chapter, we focus on (i) the consequences of generating responses from passages that do not contain the answer, which results in non-factual or hallucinated content, and (ii) the impact of source presentation. The variants of responses in the *answerability study* (i.e., study one) differ in factual correctness (Kryscinski *et al.*, 2020) and the presence/validity of the information source (Bolotova-Baranova *et al.*, 2023; Liu *et al.*, 2023a).

The issue of response incompleteness encompasses a range of challenges, such as presenting biased information that covers only one facet or viewpoint, determining which pieces of information to include given response length limitations, and transparency regarding the relevant information not covered. In this chapter, we focus on the subtask of viewpoint/facet diversification and examine the impact of balanced viewpoint coverage in responses. The variants of the responses in the *viewpoints study* (i.e., study two) vary in diversity (in terms of viewpoints and/or facets) (Helberger *et al.*, 2018) and balance in covering various viewpoints/facets in the response.

The resources developed in this study, including the manually generated CIS responses and scripts for data analysis are made publicly available at `https://github.com/iai-group/sigirap2024-resgen`. Additional results from the user studies are available in Appendix A.

This chapter is based on the following paper:

Łajewska *et al.* (2024a): *Can Users Detect Biases or Factual Errors in Generated Responses in Conversational Information-Seeking?*, SIGIR-AP '24

Table 4.1: Example problems of *query answerability* and *response incompleteness*: the first response contains factual errors and is based on sources that do not provide an answer to the question (Malbec wine is not produced in Penedès, Spain). The second response mentions multiple viewpoints, but only one is covered in detail, resulting in a biased answer.

| ***Query Answerability*** | ***Response Incompleteness*** |
|---|---|
| To combine hiking and Malbec wine, plan a trip to the Penedès region in Catalonia. You can explore the Montserrat mountain range, which offers fantastic hiking opportunities, and then visit renowned wineries in the Penedès, known for its exceptional Malbec wine production...<br>https://www.winetourism.com/wine-tasting-tours-in-penedes/ | The Watergate scandal had a profoundly negative impact on President Nixon's legacy, overshadowing many of his domestic achievements. It tarnished his reputation as a corrupt politician, making him a symbol of political scandal and misconduct in both American politics and popular culture. While he did enact significant legislation like creating the Environmental Protection Agency, his presidency is primarily remembered for the scandal, leading to his resignation and becoming synonymous with political corruption and disgrace. |

## 4.1 Related Work

CIS responses aim to synthesize information from multiple sources while balancing conciseness and completeness (see Section 2.2.2). One of the primary expectations from generated responses is to equip users with the necessary tools for assessing the reliability and accuracy of the provided information (Liu, 2023). While generative language models enhance the fluency of the generated text, issues such as hallucinations, biases, unanswerability, and source subjectivity affect the reliability of response (see Section 2.2.4). Given the potential flaws that may result from these challenges, conversational response generation should involve system revealment and promote a more informed user experience (Azzopardi *et al.*, 2018; Radlinski and Craswell, 2017).

Evaluating response quality in CIS systems presents unique challenges, as traditional offline evaluation measures like ROUGE (Lin, 2004) and NDCG (Järvelin and Kekäläinen, 2002) fail to fully capture the complexities of conversational context, multi-turn dialogue coherence, and the overall user experience in conversational interactions. Evaluating CIS responses from a user perspective involves multiple dimensions (Sakai, 2023), including trust and fairness (Zamani *et al.*, 2023), credibility (Bink *et al.*, 2022), reliability (Lu and Yin, 2021; Rechkemmer and Yin, 2022), verifiability (Liu *et al.*, 2023a), factual correctness, transparency (e.g., information sources, ranking, and consolidation process) (Shah and Bender, 2022), relevance, naturalness, conciseness (Owoicho *et al.*, 2022), informativeness (supporting user in increasing their information literacy) (Shah and Bender, 2022), perceived satisfaction, and usefulness (ter Hoeve *et al.*, 2022; Zheng *et al.*, 2022; Cambazoglu *et al.*, 2021) (see Section 2.2.3). However, directly asking users to assess these dimensions may not be reliable as users may interpret the concepts differently (the problem of indirect observables) (Kelly, 2007).

## 4.2 Methodology

We aim to investigate if users can recognize inaccuracies in CIS system responses and how these inaccuracies impact the user experience—hereafter, we use *response* to refer to *CIS system response.* We conduct two crowdsourcing studies employing a within-subject design that investigates the problems of:

- Query answerability through an *answerability study* with the focus on factual errors and quality of the information sources accompanying the response.
- Response incompleteness through a *viewpoints study* with the focus on the balance of viewpoints and/or facets in the response.

For each study, we select ten queries susceptible to one of the identified problems (i.e., answerability or incompleteness). For each query, we manually create response variants differing in terms of two controlled dimensions (1) factual correctness and (2) source presence/validity in the *answerability study*; and (1) facet/viewpoint diversity and (2) balanced facet/viewpoint presentation in the *viewpoints study*. Workers are presented with a set of queries with responses and asked to indicate their perception of the controlled dimensions listed above, as well as their overall satisfaction. We consider a simplified scenario involving a set of topics that are particularly susceptible to these issues, and we manually introduce isolated, easily detectable errors. We acknowledge that in real-world conversations such errors are likely to be much harder to identify. This chapter presents only a preliminary study, and exploring more realistic and complex scenarios is left for future work.

We aim to investigate users' ability to detect pitfalls in responses in a scenario that closely mirrors real-life system interactions. In actual situations, a user poses a query, receives a single system response, and must then judge whether this response is useful and satisfying. To replicate this setting, we provide each worker with a set of identical queries and a single version of the response for each query. This way, we may include different variants of the response in one task without the differences being too conspicuous when all possible variants of the response for a given query are presented consecutively. These response sets are carefully balanced in terms of accuracy, ensuring that users encounter in their microtasks—hereafter, Human Intelligence Tasks (HITs)—responses of different quality, without those differences being overly apparent.

### 4.2.1 Experimental Design

Crowd workers are presented with ten query-response pairs in each HIT and asked to assess the provided responses. Responses differ in their quality and accuracy along different controlled dimensions. Each response is an instance of one of the experimental conditions. In the *answerability study*, we consider four different experimental conditions $EC^A$ (resulting in four response variants for each query), and in the *viewpoints study*, three $EC^V$ (with three response vari-

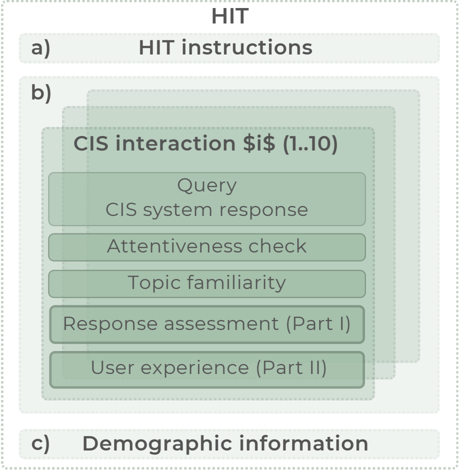

HIT
a) HIT instructions
b)
CIS interaction $i$ (1..10)
Query
CIS system response
Attentiveness check
Topic familiarity
Response assessment (Part I)
User experience (Part II)
c) Demographic information

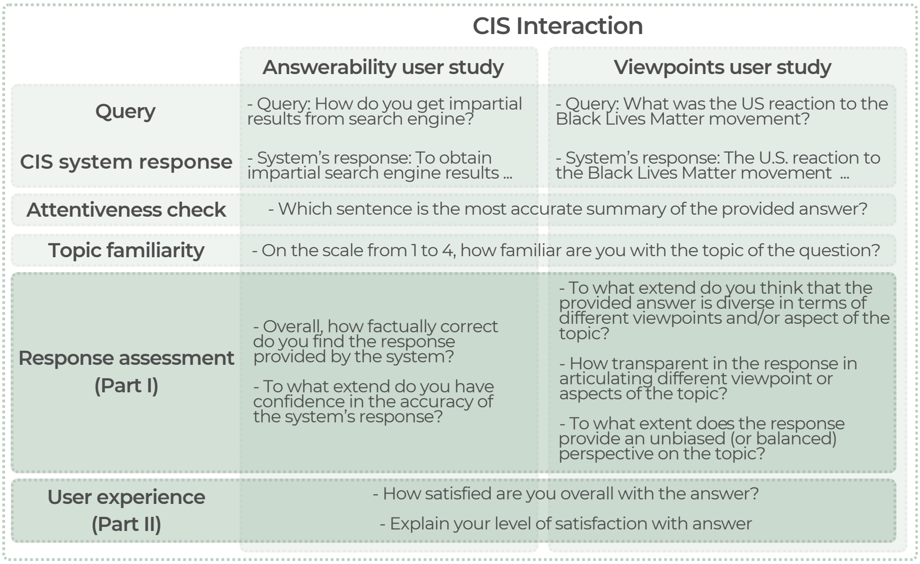
Figure 4.1: High-level design of the user studies.

ants for each query). The experimental conditions of manually crafted responses for both user studies are described in Section 4.3.1.

Both our studies follow the Graeco-Latin square design, which ensures the rotation and randomization of queries and response variants, as well as no overlap in sets of query-response pairs between HITs (Kelly, 2007). Each query-response pair appears in three different HITs, where each HIT contains a different set of ten query-response pairs. Query-response pairs appear in the HITs in a random order. Considering grouping factors that arise whenever one annotator rates multiple responses, we ensure that each crowd worker completed only a single HIT for a given user study (but they were allowed to participate in both user studies). This way, we attempt to balance the need for a large enough annotator pool with a sufficient task size to be worthwhile to the crowd workers (Steen and Markert, 2021).

### 4.2.2 Tasks

The designs of the *answerability study* and the *viewpoints study* follow the same principle: workers are asked to complete one HIT, consisting of ten query-response pairs. The task consists of (a) HIT instructions; (b) ten CIS interactions; and (c) demographics questionnaire as seen in Figure 4.1. Workers are not given specific examples of query-response pairs in the instructions to avoid biasing them. We decompose each user study into multiple subsections using independent CIS interactions to facilitate atomic microtask crowdsourcing (Gadiraju *et al.*, 2015). Each CIS interaction contains one query-response pair, followed by (1) a corresponding attentiveness check, (2) a measurement of the worker's familiarity with the topic, (3) a CIS response assessment (Part I) ,



| | CIS Interaction | |
|---|---|---|
| | Answerability user study | Viewpoints user study |
| Query | - Query: How do you get impartial results from search engine? | - Query: What was the US reaction to the Black Lives Matter movement? |
| CIS system response | - System's response: To obtain impartial search engine results ... | - System's response: The U.S. reaction to the Black Lives Matter movement ... |
| Attentiveness check | - Which sentence is the most accurate summary of the provided answer? | |
| Topic familiarity | - On the scale from 1 to 4, how familiar are you with the topic of the question? | |
| Response assessment (Part I) | - Overall, how factually correct do you find the response provided by the system?<br>- To what extend do you have confidence in the accuracy of the system's response? | - To what extend do you think that the provided answer is diverse in terms of different viewpoints and/or aspect of the topic?<br>- How transparent in the response in articulating different viewpoint or aspects of the topic?<br>- To what extent does the response provide an unbiased (or balanced) perspective on the topic? |
| User experience (Part II) | - How satisfied are you overall with the answer?<br>- Explain your level of satisfaction with answer | |

Figure 4.2: Questions provided to crowd workers in our user studies.

and (4) a measurement of user experience (Part II).[1] The wording of the questions in all parts of the user studies follows questions proposed by Tang *et al.* (2022) for evaluating the factual consistency of summaries (see Figure 4.2). Both studies finish with a short demographics questionnaire asking workers' age, education level, and gender.

**Attentiveness Check** We present workers with an additional question for each CIS interaction for which we have a ground truth answer to serve as an attention check, which enables us to detect poorly performing workers, cheat submissions, or bots (Gadiraju *et al.*, 2015). Each attention check question consists of three sentences related to the query's topic, one of them being a summary of the provided response. Sentences are provided in a random order and workers are asked to select the best summary (Bolotova-Baranova *et al.*, 2023). Submissions that failed on more than 3/10 attentiveness questions were rejected.

**Topic Familiarity** In this part of the CIS interaction task, crowd workers are asked to rate their familiarity with the query topic to help us assess the task difficulty and condition the collected data on users' background knowledge (Krishna *et al.*, 2021).

**Part I: Response Assessment** In Part I, workers are asked to evaluate the dimensions of the response presented for a given query. Since we are investi-

[1] The only difference between the *answerability study* and *viewpoints study* are the response dimensions for which we are collecting crowd workers' ratings in Part I.

Table 4.2: Controlled vs. user-judged response dimensions.

| **User Study** | **Response Dimension** | |
|---|---|---|
| | **Controlled** | **User-judged** |
| Answerability | (1) Factual Correctness<br>(2) Source Presence/Validity | Factual Correctness<br>Confidence in Answer Accuracy |
| Viewpoint | (1) Diversity<br>(2) Balance | Diversity + Transparency<br>Balance/Bias |

gating different response dimensions for the *answerability study* and *viewpoints study*, each study's response assessment part is different. The questions asked per study are related to the dimensions we identified for each problem and are answered by workers on four-point Likert scales. To increase the ecological validity of our experiments (and avoid making the assessment task too artificial), the dimensions used to control the generation of response (*controlled response dimensions*) do not always directly map to the dimensions that workers are asked to assess (*user-judged response dimensions*) (see Table 4.2). In the case of response dimension (2) in the *answerability study* (source presence/validity), simply asking workers whether the source is present or the link is valid would be too apparent and would violate the user study by directly suggesting some specific user behavior (i.e., clicking the link). Therefore, we attempt to capture this dimension by asking about the worker's confidence in the accuracy of the answer. In the case of response dimension (1) in the *viewpoints study* (diversity), it is not enough to ask how diverse the topic is, since recognizing the lack of diversity requires some knowledge about the topic. Therefore, we include an additional user-judged response dimension related to transparency in articulating different viewpoints or facets of the topic. Dimension (2) in the *viewpoints study* (balance) is provided with an additional explanation to ensure a common understanding of the underlying concept. Namely, we ask to assess the unbiased (or balanced) perspective on the topic.

**Part II: User Experience** In the final part of each CIS interaction, we pose a question about the overall satisfaction with the response (a proxy for the user experience). It is followed by a required open text field for workers to elaborate on their decision.

### 4.2.3 Data Analysis Methods

To address RQ2.1a, we assess if workers can detect flaws and inaccuracies in the responses based on their ratings for user-judged response dimensions. We use two-way ANOVA (Fisher, 1992) for analyzing the results, where the different controlled response dimensions, representing different variants of the responses, are factor variables. A separate ANOVA is performed for each user-judged response dimension (dependent variable) with the two controlled dimensions used in a given study as independent variables. Additionally, three-way ANOVA is used to investigate whether the controlled response dimensions and the question or user's familiarity with the topic have an effect on users' evaluation of the responses (measured with user-judged response dimensions). We analyze the crowdsourced data with the Python `statsmodels` library[2] and we use a significance level of $\alpha = 0.05$ to report statistical significance. Whenever applicable, the $\omega^2$ unbiased effect size of a given factor is calculated to quantify the magnitude of the variance observed in the model. It is classified based on the scales used by Culpepper *et al.* (2022) ($\omega^2 \geq 0.14$: large effect size; 0.06–0.14: medium; 0.01–0.06: small; $\leq 0$: no effect).

## 4.3 User Study Execution

We used the Amazon Mechanical Turk (AMT) crowdsourcing platform to collect responses from online workers.[3] The studies were run between 15 September 2023 and 4 October 2023.

### 4.3.1 Data

A critical element of the study is selecting query-response pairs that best represent the particular challenges. We manually craft responses for twenty search queries from TREC CAsT'20 (Dalton *et al.*, 2020) and '22 (Owoicho *et al.*, 2022),[4] simulating everyday system interactions under various experimental conditions. The responses are curated by the authors of this work to ensure accordance with defined response dimensions and high data quality.

#### Queries

For each user study, we select ten queries from the topics released in CAsT'20 and '22 that are susceptible to one of the identified problems (i.e., query answerability and response incompleteness) as detailed below.

**Answerability Study** To identify queries with unanswerability issues (i.e., queries for which answers have not been found), we use the information nugget (i.e., a piece of valuable information) annotations from the CAsT-snippets dataset (to be detailed in Chapter 5) to indicate whether the answer or part of it has been found in the top retrieved passages. We aim to select queries not widely covered in the TREC CAsT passage collections and for which retrieving the answer was challenging. Based on the annotations provided in the CAsT-snippets

[2] `https://www.statsmodels.org/`

[3] Our institution does not require ethics approval for this kind of study.

[4] The TREC CAsT'19 dataset is less complex compared to the 2020 and 2022 editions, while the CAsT'21 dataset assesses relevance at the document level instead of passages.

Table 4.3: Queries from the TREC CAsT'20 and '22 datasets used in the *answerability study*.

| ID | TREC ID | Query |
|---|---|---|
| 1 | 146_1-9 | What's the best bike seat |
| 2 | 135_2-3 | How often should I run to lose weight? |
| 3 | 139_2-15 | What are the other natural wonders of the world besides the Great Barrier Reef? |
| 4 | 142_7-1 | I like hiking and Malbec wine. You mentioned some high peaks. How can I hike some high mountains and visit some wineries famous for Malbec? |
| 5 | 144_2-11 | Tell me about the different types of rocket engines. |
| 6 | 147_2-3 | Interesting. What was the basis of the backlash Marvel Studios faced for the Vice President's suggestion that diversity was causing sales to slide? |
| 7 | 149_3-1 | How do you get impartial results from search engines? |
| 8 | 82_6 | What is the role of Co-Extra in GMO food traceability in the EU? |
| 9 | 85_4 | What licenses and permits are needed for a food truck? |
| 10 | 90_5 | Why did the Airbus A380 stop being produced? |

dataset, we select queries that contain annotated snippets in some but not all of the top-5 passages (based on their ground truth relevance scores in the TREC CAsT datasets). This way, we ensure that the query faces unanswerability problems, but some passages contain information that can be used to generate factually correct responses.[5] After selecting potential candidates, we randomly select only one query per topic to maintain the study's topical diversity. The queries used in the *answerability study* are presented in Table 4.3.

**Viewpoints Study** Open-ended queries about complex or contentious topics with multiple facets and/or viewpoints are specifically prone to incomplete responses (Draws *et al.*, 2021b). To identify such queries in TREC CAsT collections, we: (1) manually select a subset of potential candidates and (2) ask crowd workers to prioritize the selected queries in terms of their controversy and broadness. In step (1), we identify queries related to politics, society, environment, science, education, and technology. Queries strongly dependent on the conversational context or requiring background knowledge are not considered. In step (2), we run a small crowdsourcing task where workers are presented with a question and asked to assess its controversy and broadness on an ordinal scale of 1–5. Based on the collected judgments, we select the top 12 queries for which we generate different variants of the responses. At this stage, we select two additional queries to run an additional validation step. The final ten queries used in the *viewpoints study* are presented in Table 4.4.

[5] Note that answerability can be determined w.r.t. a document (e.g., SQuAD 2.0 (Rajpurkar *et al.*, 2018)), corpus (e.g., TREC CAsT (Dalton *et al.*, 2019)), knowledge base (Pathiyan Cherumanal *et al.*, 2024), or external expert knowledge. In this chapter, we consider answerability w.r.t. a particular set of retrieved passages.

Table 4.4: Queries from the TREC CAsT'20 and '22 datasets used in the *viewpoints study*.

| ID | TREC ID | Query |
|---|---|---|
| 1 | 137_1-5 | What do other philosophers think about Bostrom's 'simulation argument'? |
| 2 | 105_6 | What was the US reaction to the Black Lives Matter movement? |
| 3 | 102_8 | Can social security be fixed? |
| 4 | 149_2-5 | Are algorithms really biased against people of colour |
| 5 | 136_1-13 | What effects did the Watergate scandal have on President Nixon's legacy? |
| 6 | 138_1-9 | Do you think social media might play a role in my son's low self-esteem? |
| 7 | 91_7 | What do users of social networks get in return for by giving up their privacy? |
| 8 | 147_2-1 | What is Marvel Studios' approach to diversity for people of color? |
| 9 | 82_2 | What are the pros and cons of GMO food labeling? |
| 10 | 132_2-1 | That's interesting. Tell me more about how climate change affects developing countries. |

### Responses

The responses were manually created by the authors of this work and are based on the five most relevant passages in the TREC CAsT datasets. The selected passages were first summarised using GPT-3.5, then manually reviewed and embellished to add or remove information, verify the correctness, introduce factual errors, or balance the content depending on the experimental condition. We identify two main dimensions for generating system responses in each user study, acknowledging that these dimensions are not exhaustive. Nevertheless, our hypothesis posits that varying the responses along these dimensions will give us the means to answer our research questions effectively.

**Answerability Study** Failure to find the exact answer to the query in CIS can lead to factual errors and hallucinations (i.e., the introduction of facts that are not true). This is a common problem especially when the response is generated as a summary of partially relevant passages using large language models (Tang *et al.*, 2023). Therefore, we are mostly interested in the following two response dimensions:

1. factual correctness of the included information, and

2. the presence and validity of the source of the information.

The accurate response contains factually correct information along with the source ($EC_1^A$), whereas the flawed response fails to provide a source ($EC_2^A$), contains factually incorrect or unsupported information with an invalid source ($EC_3^A$), or lacks a source altogether ($EC_4^A$); see Table 4.5. The flawed response may contain various factual inconsistencies, such as negation and number, entity, or antonym swaps (Kryscinski *et al.*, 2020), as well as fully hallucinated content

Table 4.5: Schema for experimental conditions ($EC_1^A$–$EC_4^A$) in the *answerability study*. The last two columns contain different variants of CIS system response along with the source for Query 4 (cf. Table 4.3).

| **Exp.** | **Res. Dimensions** | | **CIS System Response** | **Source** |
|---|---|---|---|---|
| **Cond.** | **Fact.Cor.** | **Source** | | |
| $EC_1^A$<br>Factually correct + valid source | ✓ | ✓ | *You can combine your love for hiking and Malbec wine by visiting Mendoza, Argentina. This picturesque city is nestled in the Andes and is renowned for its vineyards...* | https://wanderingtrader.com/argentina/top-5-argentina-tourist-attractions/ |
| $EC_2^A$<br>Factually correct + no source | ✓ | × | Same as above | – |
| $EC_3^A$<br>Factually incorrect + invalid source | × | ✓<br>(invalid) | *To combine hiking and Malbec wine, plan a trip to the Penedès region in Catalonia. You can explore the Montserrat mountain range, which offers fantastic hiking opportunities, and then visit renowned wineries in the Penedès, known for its exceptional Malbec wine production...* | https://www.winetourism.com/wine-tasting-tours-in-penedes/<br>(The link is valid but the article is a website with Wine Tasting & Tours in Penedès, Spain where Malbec wine is not produced.) |
| $EC_4^A$<br>Factually incorrect + no source | × | × | Same as above | – |

not supported by any source information (Ji *et al.*, 2023; Liu *et al.*, 2023a). An invalid source indicates a mismatch between the source's name and content, a topically relevant source that does not support the specific facts in the response, or a source with a broken link. Following the setup proposed for evaluating the usefulness of supporting documents in the WikiHowQA benchmark (Bolotova-Baranova *et al.*, 2023), we allow workers to freely examine the sources linked in the responses to evaluate their correctness and relevance.

**Viewpoints Study** Research on debated topics typically represents viewpoints in a binary fashion (in favor/against). However, viewpoints are additionally characterized by stance, i.e., the degree of strength (e.g., slight support vs. strong favor) and the logic of evaluation (underlying reason or perspective behind the stance) (Draws *et al.*, 2022). Our user study does not address the stance or evaluation logic and focuses on a widely understood diversity of viewpoints and facets. Crowd workers are asked to judge whether the expressed viewpoints or described topic facets are diverse enough or not. While investigating queries that are likely to result in incomplete responses, we are interested

Table 4.6: Schema for experimental conditions ($EC_1^V$–$EC_3^V$) in the *viewpoints study*. The last column contains different variants of CIS system response for Query 5 (cf. Table 4.4).

| **Exp.** | **Res. Dimensions** | | **CIS System Response** |
|---|---|---|---|
| **Cond.** | **Diversity** | **Balance** | |
| $EC_1^V$<br>Multiple viewpoints covered to the same extent | ✓ | ✓ | *The Watergate scandal initially had a severely damaging effect on President Nixon's legacy, tarnishing his image as a corrupt politician and overshadowing his accomplishments. However, in later years, Nixon's reputation underwent a degree of rehabilitation, with recognition of his roles in civil rights and diplomacy. He transitioned from being primarily associated with scandal to being acknowledged as a respected elder statesman, although his legacy remains marked by the shadow of Watergate.* |
| $EC_2^V$<br>Multiple viewpoints mentioned, one covered in detail | ✓ | × | *The Watergate scandal had a profoundly negative impact on President Nixon's legacy, overshadowing many of his domestic achievements. It tarnished his reputation as a corrupt politician, making him a symbol of political scandal and misconduct in both American politics and popular culture. While he did enact significant legislation like creating the Environmental Protection Agency, his presidency is primarily remembered for the scandal, leading to his resignation and becoming synonymous with political corruption and disgrace.* |
| $EC_3^V$<br>Single viewpoint mentioned and covered | × | × | *The Watergate scandal had a profoundly negative impact on President Nixon's legacy. It tarnished his reputation as a corrupt politician, making him a symbol of political scandal and misconduct in both American politics and popular culture. Nixon's resignation and the scandal's fallout reinforced public skepticism and criticism of the presidency, leaving a lasting impression as one of the most Shakespearean and disgraceful episodes in presidential history.* |

in the following two dimensions:

1. response diversity in terms of different viewpoints and/or facets mentioned, and
2. balance in the amount of information provided for each viewpoint and/or facet.

The accurate response equally covers various points of view and/or facets of the topic to the same extent ($EC_1^V$). The flawed response mentions several viewpoints and/or topic facets but elaborates only on one of them ($EC_2^V$) or mentions only one ($EC_3^V$); see Table 4.6.[6]

We introduce an additional step for the *viewpoints study* to validate our proposed response dimensions: diversity, and balance. This step, addressing

[6] Note that a text discussing a single viewpoint or facet cannot be unbalanced; therefore, an experimental condition with a lack of diversity and balance is not applicable.

the subjectivity of controversy and topic broadness, aids in filtering out non-representative query-response pairs. We create small surveys where expert annotators are presented with three topics and lists of recommended resources used to generate the responses. Expert annotators are asked to explore the provided resources to become familiar with the given topic. Then, they are presented with different response variants and asked to judge the diversity and balance of each of the provided query-response pairs. For each of the twelve queries, we collect ratings between 1–5 for diversity and balance from three different expert annotators. We employ Ph.D. students for their academic skills in exploring new domains, assuming their ratings reflect users highly familiar with the topics (i.e., experts). We exclude the query for which the response variant corresponding to $EC_1^V$ (multiple viewpoints covered to the same extent) is judged as not diverse enough and the query for which the response variant corresponding to $EC_3^V$ (single viewpoint mentioned and covered) is judged as too balanced.

### 4.3.2 Workers

Crowd workers with an approval rate greater than 97%, more than 5,000 approved HITs, and located in the US were qualified to participate in the studies. Workers were paid $3 USD for successful HIT completion. The reward was estimated based on the time needed by an expert to complete the task (the time was increased by 30%) and the federal minimum wage in the US ($7.25 USD per hour). Three different workers assessed each query-response pair to avoid repeated judgments that would reduce the reliability of the study (Steen and Markert, 2021). This user study setup gave us 12 (3 workers × 4 answer variants per query) different HITs for the *answerability study* and 9 (3 workers × 3 answer variants per query) for the *viewpoints study*. This resulted in 36 annotators for *answerability study* and 27 annotators for *viewpoints study*. The power analysis,[7] employing results of one-way ANOVA with the experimental condition as an independent variable and the user-reported values for the main response dimension (factual correctness for the *answerability study* and diversity for the *viewpoints study*) as a dependent variable, was conducted using data collected in the first run. The results of the power analysis indicated that the *viewpoints study* had a strong "true" effect when it existed. In contrast, the low power of *answerability study* suggested a low statistical sensitivity—aligning with our intuition that users are unlikely to detect hallucinations. To increase the power of *answerability study*, we collected more data from five additional workers per HIT in the second run with the same worker requirements and rewards (see Table 4.7 for descriptive statistics). Ten submissions out of 133 released HITs were discarded due to failed attentiveness checks.

The *answerability study* involved 96 workers: 44 male and 52 female (no workers reported "other" or "prefer not to say"). Thirty-four workers self-reported to be in the 18–30 age group, 35 in the 31–45 group, 19 in the 46–60, and seven in the 60+ group. One participant did not report on age. Regarding education,

[7] Calculated using the scripts at https://waseda.app.box.com/v/SIGIR2016PACK

Table 4.7: User studies setup in numbers. Numbers in the parentheses refer to the second data collection run.

| | **Answerability** | **Viewpoints** |
|---|---|---|
| #queries per user study | 10 | 10 |
| #experimental cond. (#resp. per query) | 4 | 3 |
| #crowd workers per HIT | 3 (+5) | 3 |
| #different HITs | 12 | 9 |
| #crowd workers per query-response | 9 (+15) | 9 |
| #query-response pairs annotations | 360 (+600) | 270 |

two workers held a Ph.D. or higher, 15 had a master's degree, 59 had a bachelor's degree, and 19 had high school. One participant did not report on education. The *viewpoints study* involved 27 workers: 15 male and 12 female (with none selecting "Other" or "Prefer not to say"). Three workers self-reported to be in the 18–30 age group, 12 in the 31–45 group, 10 in the 46–60, and two in the 60+ group. Two workers had a master's degree, 16 had a bachelor's degree, and 8 had a high school education. One participant did not report on education.

## 4.4 Results and Discussion

The analysis of data obtained from the crowdsourcing experiments is performed using the methods described in Section 4.2.3.

### 4.4.1 Users' Ability to Recognize Problems

Table 4.8 shows the results of the two-way ANOVA performed to answer RQ2.1a (*Can users effectively recognize problems related to query answerability and response incompleteness in system responses?*). Controlled response dimensions are treated as independent variables, and a given response dimension (i.e., self-reported worker ratings) as a dependent variable. Statistically significant results indicate an effect of the experimental condition on a given response dimension.

**Effect of controlled response dimension manipulation on response user ratings** We do not observe any statistically significant effect of manipulating the controlled response dimensions on user ratings in the *answerability study* (upper part of Table 4.8), suggesting that users cannot recognize pitfalls in the responses or do not associate them with any of the response dimensions. On the other hand, results for the *viewpoints study* (lower part of Table 4.8) show small or medium effect on self-reported worker ratings meaning that users can correctly identify the problems related to viewpoint diversity and balance.

Table 4.8: Results of two-way ANOVA. Statistically significant effects are in bold. Effect size: L=Large, M=Medium, S=Small (see Section 4.2.3).

| **Dependent Variable (User-Judged)** | **Independent Variable(s) (Controlled)** | ***p*-value** | **Effect Size** |
|---|---|---|---|
| *Answerability Study* | | | |
| Factual Correctness | **Contr. Fact. Corr.** | **0.014** | **-** |
| | Contr. Source | 0.664 | - |
| | Contr. Fact. Corr. × Contr. Source | 0.267 | - |
| Conf. in Answer Acc. | Contr. Fact. Corr. | 0.244 | - |
| | Contr. Source | 0.763 | - |
| | Contr. Fact. Corr. × Contr. Source | 0.575 | - |
| Overall Satisfaction | Contr. Fact. Corr. | 0.306 | - |
| | Contr. Source | 0.394 | - |
| | Contr. Fact. Corr. × Contr. Source | 0.267 | - |
| *Viewpoints Study* | | | |
| Diversity | **Contr. Diversity** | **0.0** | **M** |
| | Contr. Balance | 1.0 | - |
| | **Contr. Diversity × Contr. Balance** | **0.0** | **M** |
| Transparency | **Contr. Diversity** | **0.0** | **M** |
| | Contr. Balance | 1.0 | - |
| | **Contr. Diversity × Contr. Balance** | **0.0** | **M** |
| Balance | **Contr. Diversity** | **0.0** | **S** |
| | Contr. Balance | 1.0 | - |
| | **Contr. Diversity × Contr. Balance** | **0.0** | **S** |
| Overall Satisfaction | **Contr. Diversity** | **0.0** | **S** |
| | Contr. Balance | 1.0 | - |
| | **Contr. Diversity × Contr. Balance** | **0.0** | **M** |

**Effect of the interaction between query and controlled response dimensions on user ratings** The three-way ANOVA results in Table 4.9 show that the query and interaction between the query and the controlled response dimensions (especially factual correctness) significantly affect all response dimensions in the *answerability study*, which aligns with findings from other information retrieval experiments, highlighting the topic-dependent nature of user judgments (Culpepper *et al.*, 2022; Alaofi *et al.*, 2022). It indicates that the perceived factual correctness may vary based on the query, despite the consistent experimental condition. In the *viewpoints study*, only the diversity and overall satisfaction with the response are affected by the interaction between the query and controlled response dimensions, suggesting that the *viewpoints study* is more robust w.r.t. topic/query variability.

**Effect of the interaction between user background knowledge and experimental condition** The topic familiarity reported by workers is a proxy for user background knowledge. Even though we anticipated that the topic familiarity would influence the ratings reported by the workers for different response dimensions, we did not observe a statistically significant association of the interaction between the familiarity and experimental condition on any of the response dimensions. This holds for both user studies (see Table 4.10).

### 4.4.2 User Experience

This section discusses the results to answer RQ2.1b (*How do factually incorrect, inaccurate, incomplete, and/or biased responses impact the user experience?).*

**Correlation between user-reported response dimensions and the overall satisfaction** Table 4.11 shows the Pearson correlation coefficient $r$ calculated for overall satisfaction—a proxy for user experience—, and user-reported response dimensions. For both user studies, we observe a moderately strong correlation ($0.6 < r < 0.8$) between user satisfaction and other user-judged dimensions. This suggests that satisfaction is a fairly good indicator of the overall user experience. Correlations for the *answerability study* are lower than for the *viewpoints study*. As we discussed in Section 4.4.1, we do not observe a statistically significant effect of the controlled response dimension on user ratings for the *answerability study*. This implies that users find these response dimensions important and associate them with satisfaction, but they are not able to identify them correctly in system responses. On the other hand, results for the *viewpoints study* suggest that users can correctly identify these dimensions and use them as indicators for their satisfaction.

**Effect of query and response quality on overall satisfaction** In both studies, the query significantly affects overall satisfaction (see Table 4.12). We do not observe a statistically significant association between controlled response dimensions and overall satisfaction in the *answerability study*, which suggests that response quality does not influence worker's perception of satisfaction (see Table 4.8). The opposite observation is made in the *viewpoints study*, implying that workers can spot response inaccuracies. The three-way ANOVA (Table 4.9) shows that a small- or medium-size effect of the query leads to a statistically significant effect of the interaction between query and response variant on the overall satisfaction for both studies. This indicates that, in terms of user satisfaction, both studies are sensitive to topic variability that may impact the results. For future work, using a larger number of queries, especially for the *answerability study*, may increase the sensitivity of the experiment.

### 4.4.3 Further Analysis

**Rating distributions for response dimensions** In the *answerability study*, the ratings for user-judged response dimensions, topic familiarity, and overall

Table 4.9: Results of three-way ANOVA. Stat. significant effects are in bold. Effect size: L=Large, M=Medium, S=Small (see Section 4.2.3).

| **Dependent Variable** | **Independent Variable(s)** | $p$**-value** | **Effect** |
|---|---|---|---|
| **(User-Judged)** | **(Controlled)** | | **Size** |
| *Answerability Study* | | | |
| Factual Correctness | **Query** | **0.0** | **S** |
| | **Contr. Fact. Corr. × Query** | **0.002** | **S** |
| | **Contr. Source × Query** | **0.048** | **-** |
| | Contr. Fact. Corr. × Contr. Source × Query | 0.439 | - |
| Conf. in Answer Acc. | **Query** | **0.015** | **S** |
| | **Contr. Fact. Corr. × Query** | **0.0** | **S** |
| | Contr. Source × Query | 0.118 | - |
| | Contr. Fact. Corr. × Contr. Source × Query | 0.341 | - |
| Overall Satisfaction | **Query** | **0.0** | **S** |
| | **Contr. Fact. Corr. × Query** | **0.0** | **S** |
| | Contr. Source × Query | 0.339 | - |
| | Contr. Fact. Corr. × Contr. Source × Query | 0.598 | - |
| *Viewpoints Study* | | | |
| Diversity | Query | 0.147 | S |
| | Contr. Diversity × Query | 0.101 | S |
| | Contr. Balance × Query | 1.0 | - |
| | **Contr. Diversity × Contr. Balance × Query** | **0.016** | **S** |
| Transparency | Query | 0.35 | - |
| | Contr. Diversity × Query | 0.582 | - |
| | Contr. Balance × Query | 1.0 | - |
| | Contr. Diversity × Contr. Balance × Query | 0.689 | - |
| Balance | **Query** | **0.012** | **S** |
| | Contr. Diversity × Query | 0.559 | - |
| | Contr. Balance × Query | 1.0 | - |
| | Contr. Diversity × Contr. Balance × Query | 0.316 | - |
| Overall Satisfaction | **Query** | **0.001** | **M** |
| | Contr. Diversity × Query | 0.599 | - |
| | Contr. Balance × Query | 1.0 | - |
| | **Contr. Diversity × Contr. Balance × Query** | **0.034** | **S** |

satisfaction per query are concentrated around higher values (3 and 4) for all response dimensions apart from familiarity (see Figure 4.3). It means that workers are not very critical in evaluating these dimensions or cannot identify the pitfalls related to them. Workers report that they are rather unfamiliar with most of the query topics. In the *viewpoints study*, the ratings for familiarity are

Table 4.10: Results of three-way ANOVA. Stat. significant effects are in bold. Effect size: L=Large, M=Medium, S=Small (see Section 4.2.3).

| **Dependent Variable**<br>**(User-Judged)** | **Independent Variable(s)**<br>**(Controlled)** | $p$**-value** | **Effect**<br>**Size** |
|---|---|---|---|
| *Answerability Study* | | | |
| Fact. Corr. | **Familiarity** | **0.006** | **S** |
| | Contr. Fact. Corr. × Familiarity | 0.962 | – |
| | Contr. Source × Familiarity | 0.275 | – |
| | Contr. Fact. Corr. × Contr. Source × Familiarity | 0.56 | – |
| Conf. in Answer Acc. | **Familiarity** | **0.0** | **S** |
| | Contr. Fact. Corr. × Familiarity | 0.894 | – |
| | Contr. Source × Familiarity | 0.556 | – |
| | Contr. Fact. Corr. × Contr. Source × Familiarity | 0.348 | – |
| Overall Satisfaction | **Familiarity** | **0.0** | **M** |
| | Contr. Fact. Corr. × Familiarity | 0.544 | – |
| | Contr. Source × Familiarity | 0.381 | – |
| | Contr. Fact. Corr. × Contr. Source × Familiarity | 0.777 | – |
| *Viewpoints Study* | | | |
| Diversity | Familiarity | 0.816 | – |
| | Contr. Diversity × Familiarity | 0.056 | S |
| | Contr. Balance × Familiarity | 1.0 | – |
| | Contr. Diversity × Contr. Balance × Familiarity | 0.628 | – |
| Transparency | Familiarity | 0.788 | – |
| | Contr. Diversity × Familiarity | 0.257 | – |
| | Contr. Balance × Familiarity | 1.0 | – |
| | Contr. Diversity × Contr. Balance × Familiarity | 0.316 | – |
| Balance | Familiarity | 0.89 | – |
| | Contr. Diversity × Familiarity | 0.325 | – |
| | Contr. Balance × Familiarity | 1.0 | – |
| | Contr. Diversity × Contr. Balance × Familiarity | 0.242 | – |
| Overall Satisfaction | Familiarity | 0.358 | – |
| | Contr. Diversity × Familiarity | 0.187 | – |
| | Contr. Balance × Familiarity | 1.0 | – |
| | Contr. Diversity × Contr. Balance × Familiarity | 0.38 | – |

more spread. A wide range of diversity ratings is observed per query, unlike for other response dimensions. Even though the ratings are more spread than for the *answerability study*, most of the ratings concentrate around a higher value (i.e., 3).

Table 4.11: Pearson correlation between user-reported response dimensions and their overall satisfaction with the system's response.

| **Response Dimension** | **Correlation Coefficient** |
|---|---|
| *Answerability Study* | |
| Factual Correctness | 0.634 |
| Conf. in Answer Acc. | 0.660 |
| *Viewpoints Study* | |
| Diversity | 0.720 |
| Transparency | 0.727 |
| Balance | 0.785 |

**Effect of background knowledge on the response dimensions** According to the results of one-way ANOVA with familiarity used as an independent variable (see Table 4.12), we obtain different results for the two studies. In the *answerability study*, the worker's background knowledge impacts how accurate or satisfying they find the response. Whereas, in the *viewpoints study*, none of the response dimensions is significantly affected by users' topic familiarity.

**Effect of the query on the response dimensions** In both user studies the topic familiarity and overall user satisfaction are significantly affected by the query (see Table 4.12). It means that user background knowledge and response satisfaction depend on the query, not necessarily on the response. It confirms that, to get meaningful results, one must include many different study topics, which is indeed what we tried to ensure with our query selection processes. Statistically significant differences in response dimensions between queries are observed for all dimensions in the *answerability study*, while only for balance in the *viewpoints study*. This suggests that the former studies' setup is more query-dependent than the latter. The results are more generalizable in the *viewpoints study*, even after collecting additional data according to the power analysis results for the *answerability study*. The high effect of the query on all the response dimensions in the *answerability study* also justifies the significant effects of the interactions between the query and the controlled response dimensions observed in the three-way ANOVA.

### 4.4.4 Qualitative Analysis

To validate our findings, we characterize workers' user experience by analyzing their natural language comments. We manually inspect all the 960 worker comments in the *answerability study* and 270 in the *viewpoints study*. We followed an inductive approach (Williams, 2008) to identify themes in the comments. After consensus among the authors, one of the authors labeled all comments.

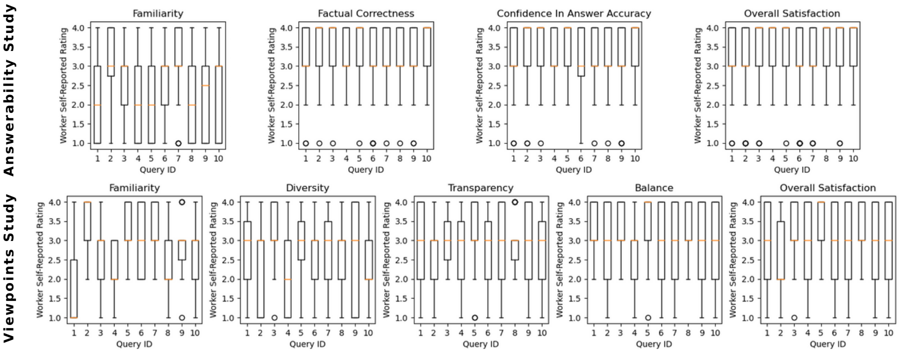


Figure 4.3: Distribution of user-judged response dimensions per query in the two user studies. Yellow lines indicate the median values.

Next, we counted how many workers mentioned a particular aspect.[8] In the reporting below, numbers in parentheses indicate the proportion of high (when the aspect mentioned in the comments is positive) or low (when the aspect is negative) satisfaction ratings corresponding to specific aspects mentioned in the comments. For instance, if three workers mention a positive aspect like “factual correctness” in the open text field, but only two assign high satisfaction scores on the four-point Likert scale, this is reported as 2/3. Conversely, if a negative aspect like “bias” is identified by five workers and four of them express low satisfaction, it is recorded as 4/5.

Coherence, fluency, naturalness, details, and logic of the response mentioned in the comments are almost always accompanied by high satisfaction ratings (178/185 in the *answerability study* and 23/23 in the *viewpoints study*). In the *answerability study*, comments mentioning positive aspects such as factual correctness (126/133), information completeness (99/100), agreement with the response (59/60), presence (53/64), and credibility (18/23) of the source are accompanied by high satisfaction ratings (3 or 4 on the Likert scale). However, high satisfaction ratings are not always paired with positive comments. Some comments associated with high satisfaction ratings indicate negative aspects, such as lack of source (4/21) or invalid source (2/11). Additionally, highlighting missing or incomplete information (60/156) does not always cause a decrease in the satisfaction rating. In the *viewpoints study*, positive comments indicating high diversity (55/58), balance (6/6), lack of bias (12/13), completeness of the provided response (22/22), or agreement with the answer (8/9) are accompanied by high satisfaction ratings. However, some responses describing negative aspects such as bias (14/25) or lack of diversity (43/64) are still given high satisfaction ratings. Most of the responses described as not diverse (43/64) or

[8] While adhering to an established qualitative analysis approach, the authors acknowledge that personal interpretation may introduce some degree of subjectivity in the interpretation and categorization of the data.

Table 4.12: Results of one-way ANOVA. Statistically significant effects are in bold. Effect size: L=Large, M=Medium, S=Small (see Section 4.2.3).

| Dependent Variable | Independent Variable(s) | $p$-value | Effect Size |
|---|---|---|---|
| *Answerability Study* | | | |
| **Familiarity** | Query | **0.0** | **M** |
| **Factual Corr.** | | **0.0** | **S** |
| **Conf. in Answer Acc.** | | **0.019** | **S** |
| **Overall Satisfaction** | | **0.0** | **S** |
| **Factual Correctness** | Familiarity | **0.005** | **S** |
| **Conf. in Answer Acc.** | | **0.0** | **S** |
| **Overall Satisfaction** | | **0.0** | **M** |
| *Viewpoints Study* | | | |
| **Familiarity** | Query | **0.0** | **L** |
| Diversity | | 0.338 | - |
| Transparency | | 0.458 | - |
| **Balance** | | **0.027** | **S** |
| **Overall Satisfaction** | | **0.005** | **S** |
| Diversity | Familiarity | 0.375 | - |
| transparency | | 0.478 | - |
| Balance | | 0.639 | - |
| Overall Satisfaction | | 0.378 | - |

imbalanced (12/22) are accompanied by low satisfaction ratings (1 or 2 on the Likert scale). Additional aspects mentioned in the comments include the usefulness (33/35 comments accompanied by high satisfaction rating) and subjectivity (10/24 comments accompanied by low satisfaction rating) of the response in the *answerability study*, and lack of source (4/12 comments accompanied by low satisfaction rating) in the *viewpoints study*. It is worth pointing out that while usefulness is a common indicator of successful completion of the search task (Cambazoglu *et al.*, 2021; Liu, 2023), it is only mentioned in 2.8% of the comments.

Although satisfaction ratings are skewed (see Figure 4.3) and other scales such as magnitude estimation (Turpin *et al.*, 2015) may give us more informative ratings, they are roughly aligned with the dimensions we aim to capture. It implies that the dimensions we use to differentiate between response variants impact user satisfaction. Comments from the *answerability study* suggest that satisfaction is associated with both factual correctness and source validity. The frequent user references to factual correctness in comments imply a significant focus on this aspect when evaluating responses. Even though we observe a high correlation between user-reported response dimensions and overall satisfaction, we do not observe a statistically significant effect of the controlled factual correctness on user ratings for this response dimension. This implies that users find these response dimensions important and associate them with their satisfaction, but they are not able to identify factuality correctly in the responses. Additionally, one-way ANOVA for the *answerability study* revealed that the overall satisfaction is affected by the query and topic familiarity, not the controlled response dimensions (see Table 4.12). This also explains why the response dimensions mentioned in the comments do not completely align with the actual flaws in the responses. The results' sensitivity to the topics may suggest that including more queries in further studies might reveal the effect of factual correctness and source validity on overall satisfaction. In the *viewpoints study*, our qualitative analysis shows that user satisfaction is more linked to viewpoint diversity and response completeness than information balance, differing from quantitative findings. It can follow from the fact that the concept of response diversity is better understood by users and is easier to identify. Nevertheless, the qualitative analysis shows that selected response dimensions are indeed common indicators of user satisfaction.

## 4.5 Discussion

**Summary of findings** Users generally find it easier to perceive viewpoints than to assess factual correctness. In the *answerability study*, crowd workers demonstrate a limited ability to detect pitfalls in responses compared to the *viewpoints study*, highlighting the challenge of identifying factual errors without topic-specific knowledge. In terms of user satisfaction, in the *answerability study* it strongly correlates with confidence in answer accuracy, highlighting the importance of valid sources. In the *viewpoints study*, satisfaction is tied to perceived balance, with users preferring unbiased responses that equally cover all viewpoints. Satisfaction scores reported by users do not always align with their comments—additional aspects revealed in free-text user comments refer to source credibility, as well as the completeness, usefulness, and subjectivity of the provided information—, indicating a potential discrepancy between reported and actual satisfaction levels. Users may also associate their satisfaction with response fluency, which can be easily ensured by existing generative search engines. However, it does not guarantee the accuracy or proper citation of all statements (Liu *et al.*, 2023a).

**Implications of our findings** The conclusions drawn from these studies inform the design of future response generation methods and highlight important challenges that still need to be addressed. Simply relying on the relevance of the top retrieved passages does not guarantee the generation of a satisfying response. Future response generation approaches must ensure the completeness, diversity, balance, objectivity, and factual correctness of responses, along with proper attribution to credible sources. Additionally, the response should inform users of potential inaccuracies and help them assess the presented information objectively, by providing sources or system capability details. Including these explanations ensures transparent and effective interactions with the system (see Chapter 7). Another open challenge is the evaluation of the generated responses. To the best of our knowledge, there are no CIS datasets with ground truth judgments for the identified response dimensions. Our study designs and experimental protocol can serve as a blueprint for human evaluation of responses across multiple dimensions, supporting data collection for a broader range of experimental conditions, more complex multi-turn settings, and additional queries/topics.

**Lessons learned from executing user studies** Reflecting on our experiences from conducting these user studies, several key lessons that may be found useful by the community have become apparent:

1. The effectiveness of these studies depends on the careful selection of representative queries and responses to the problems being investigated.

2. Incorporating validation steps, especially in experiments that involve subjectivity, helps mitigate biases introduced by study designers.

3. Implementing attentiveness checks is crucial for ensuring the quality of collected data and maintaining the credibility of the gathered information.

4. While qualitatively analyzing responses from crowd workers in natural language may incur higher costs, it can reveal unforeseen dimensions and challenges. Furthermore, natural language responses to open-ended questions serve as reliable indicators of data quality—fluent, relevant, and informative responses from crowd workers typically accompany meaningful data.

5. Developing operational definitions of explored dimensions and refining them iteratively during the initial stages of experimentation and design fosters a deeper understanding of the dimensions under examination and facilitates necessary adjustments before data collection begins.

**Limitations** Due to the complexity of the user studies and the costs involved, some simplifications were made, such as focusing on single-turn interactions and a limited number of queries. As a result, these experiments do not fully reflect the dynamic nature of real-world CIS dialogues, where user needs and context change over multiple turns. Future work will explore more topics, particularly for the *answerability study*, to enhance result sensitivity, and use other scales to capture overall satisfaction (e.g., magnitude estimation (Turpin *et al.*, 2015)). Another limitation is relying on Amazon MTurk crowd workers, who may not fully represent the diversity of CIS system users. These studies do not fully control participants' own biases, which is left for future investigation. Lastly, the findings of this work are limited to the properties of the test collection used in our experiments. Future experiments should also explore answerability on broader levels—such as ranking, corpus, and expert knowledge—while considering the system's transparency when no answer is found, as well as a wider spectrum of topics, viewpoints, and responses. Despite these limitations, the experiments serve as a first step toward understanding challenges in CIS response generation and highlight key open questions for further research.

## 4.6 Conclusions

Response generation poses various challenges in CIS systems. To study this, we proposed two crowdsourcing-based study designs to investigate unanswerable questions and incomplete responses from a user perspective in the scenario inspired by the TREC CAsT benchmark. We explored users' ability to recognize factual inaccuracies, pitfalls, and biases in terms of viewpoint diversity by controlling experimental conditions in manually crafted responses simulating CIS system interactions. In answer to **RQ2.1** *(Which limitations in the responses are detectable by users?)*, our findings indicate that users are more adept at detecting viewpoint diversity issues and response biases than factual errors or problems related to source validity. These results provide evidence that CIS system responses cannot be limited to a simple synthesis of the retrieved information and source attribution alone is insufficient to ensure effective interaction with the system. We believe CIS responses should explicitly inform users about potential inaccuracies and provide aid to assess the presented information objectively (e.g., by including credible sources or information about system capabilities).

The lessons learned from these experiments serve as a roadmap for constructing transparent and reliable responses. Insights revealing users' difficulty in detecting factual errors prompt further experiments on unanswerability detection (see Chapter 5) and directly communicating response limitations to users (see Chapter 7). Additionally, considerations around response diversity and completeness inspire our future work on clustering information from retrieved passages by topic facets and ensuring information density when generating responses (see Chapter 6).

# Part II

Addressing CIS Limitations

# Chapter 5

# Snippet-level Annotations for Predicting Query Answerability

*To learn which questions are unanswerable, and not to answer them: this skill is most needful in times of stress and darkness.*

— **Ursula K. Le Guin**

A large fraction of research on conversational information seeking (CIS) to date has focused on the problem of retrieving relevant passages. The task of conversational passage retrieval requires advances in query rewriting (Lin *et al.*, 2021; Vakulenko *et al.*, 2021a,c) and can also directly benefit from research on multi-stage passage retrieval (Luan *et al.*, 2021) (see Chapter 3). However, identifying relevant passages is only an intermediate step. Ultimately, the information contained in these passages would need to be synthesized into a single answer. *Conversational response generation* is the task of encapsulating the most relevant pieces of information in an easily consumable unit (Culpepper *et al.*, 2018) (see Section 2.2.2). Including it in the CIS pipeline would increase the naturalness of the conversation (Trippas *et al.*, 2020, 2018).

There are at least two main challenges involved in the task of response generation: identifying key pieces of information from relevant results (e.g., paragraphs) and summarizing them in a concise answer. Correspondingly, Ren *et al.* (2021) proposes to split the task into two stages: (1) identification of supporting snippets and (2) summarization of selected snippets. In this chapter, we focus on the problem of (1), and more specifically on building a snippet dataset with high-quality annotations using crowdsourcing. We develop critical CIS resources to support the grounded and transparent response generation methods proposed later in this thesis.

The significance of being able to identify relevant snippets is twofold. First, it enables the training of models that can ground the generated answers in actual statements. Natural language generation models are susceptible to hallucinations, especially if the query is insufficiently covered in the corpus, or the retrieved documents contain redundant, complementary, or contradictory information (Ji *et al.*, 2023). Therefore, employing abstractive summarization methods on top of relevant snippets identified can help to mitigate this problem and provide more control over the generation process, much in the spirit of the two-step process proposed in (Ren *et al.*, 2021) (see Chapter 6). Second, it would enable automatic evaluation of the generated responses quantitatively, in terms of relevant information nuggets included (Pavlu *et al.*, 2012). Response summarization in CIS systems has been piloted in TREC CAsT'22 (Owoicho *et al.*, 2022), where the quality of answer summaries is evaluated by human judges along three dimensions: relevance, naturalness, and conciseness (Owoicho *et al.*, 2022). Having annotations of relevant snippets would enable automatic evaluation of answers in terms of completeness (see Chapter 7).

Even though crowdsourcing has become an established means of collecting human annotations at scale, ensuring data quality can be challenging (Daniel *et al.*, 2019). Indeed, we demonstrate that the seemingly straightforward task of highlighting relevant snippets may not be so simple and deserves closer attention. In the first part of this chapter, we investigate what effective task designs and the trade-offs between worker qualifications and costs to perform the task of snippet annotations are. We address the following research question: **How to identify core information units in the relevant passages that need to be included in the response? (RQ3.1)**. Specifically, we consider paragraph- and sentence-level snippet annotation interfaces, multiple crowdsourcing platforms, and crowd workers with different qualifications as well as expert annotators. Measuring the quality of annotations is challenging because relevant snippet selection is subjective and often there are multiple correct sets of snippets in a given passage. We evaluate the resulting annotations in terms of inter-annotator agreement and similarity to expert annotations using text similarity measures adapted to this task. Based on these results, we set out to create a large-scale dataset, CAsT-snippets, which enriches the TREC CAsT'20 and '22 datasets with snippet-level answer annotations. We follow a setup in which we closely work with a selected pool of highly engaged crowd workers in order to ensure high data quality.

In the second part of this chapter, we explore the application of the CAsT-snippets dataset to query answerability detection addressing the following research question: **How to detect factors contributing to incorrect, incomplete, or biased responses? (RQ2.2)**. In an ideal scenario, when the passages from the top of the ranking answer the question, the task of response generation boils down to summarization (Owoicho *et al.*, 2022). However, it is often the case that the answer to the user's question is not contained in the top retrieved passages. In such cases, summaries generated from those passages would result in hallucinations (Tang *et al.*, 2023; Cao *et al.*, 2016). Therefore, we propose a mechanism for detecting unanswerable questions for which the correct answer is not present in the corpus or cannot be retrieved. More specifically, given a set of top-ranked passages that have been identified as most relevant to the given question, we predict if the question can be answered (at least partially) based on information contained in those passages. This enables us to move beyond the notion of passage relevance and focus more on the actual presence of the information that answers the question. Introducing this additional step of answerability prediction in the CIS pipeline, to be performed after the passage retrieval and before the response generation steps, could help mitigate hallucinations and factual errors. It would enable the system to transparently communicate to the user if the answer to the query could not be found, instead of generating a response from only marginally relevant passages.

To fill this gap, we extend the CAsT-snippets dataset with answerability labels on three levels: (1) sentences, (2) passages, and (3) rankings (i.e., top-ranked passages retrieved by a CIS system), introducing a *CAsT-answerability* dataset, to train and evaluate methods for question answerability prediction. Notably, we generate input passage rankings with various degrees of difficulty in answerability prediction, mixing passages that contain answers with those with no answers, in a controlled way. As a result, passage rankings range from all passages containing an answer to "no answer found in the corpus." This extended dataset is then used to develop a baseline approach for predicting answerability based on an input ranking. Our proposed approach predicts which sentences from the top-ranked passages contribute to the answer and aggregates the obtained answerability scores on the passage and ranking levels.

The resources presented in this chapter are made publicly available. The CAsT-snippets dataset and code for computing evaluation measures are available at `https://github.com/iai-group/CAsT-snippets`. The CAsT-answerability dataset and the implementation of our proposed answerability prediction method can be found at `https://github.com/iai-group/answerability-prediction`. Additional information about datasets is provided in Appendix B.

This chapter is based on the following papers:

Łajewska and Balog (2023b): *Towards Filling the Gap in Conversational Search: From Passage Retrieval to Conversational Response Generation*, CIKM '23 🏆

Łajewska and Balog (2024a): *Towards Reliable and Factual Response Generation: Detecting Unanswerable Questions in Information-seeking Conversations*, ECIR '24 🏆

## 5.1 Related Work

Some dimensions of responses are not reliably covered by currently available automatic metrics and require manual evaluation (e.g., coherence and relevance) (Fabbri *et al.*, 2021), while others (e.g., completeness) can be evaluated automatically, provided that more fine-grained annotations are available. Information nuggets have been proposed as an alternative to automatically assign relevance judgments to documents and/or evaluate retrieval systems (Pavlu *et al.*, 2012) (see Section 2.2.3). Our work aims to contribute to this type of evaluation by studying ways to collect snippet-level annotations. A task similar to snippet annotation (or information nuggets identification) has been broadly researched in QA systems. In most available datasets for reading comprehension focused mainly on factoid questions, the generated response is a single entity or a short segment of text from the passage (Rajpurkar *et al.*, 2016; Campos *et al.*, 2015; Tan *et al.*, 2018; Choi *et al.*, 2018).

Crowdsourcing provides a scalable means to the completion of large amounts of labeling or annotation tasks that require human intelligence (Gadiraju *et al.*, 2015). The actual quality of the results is influenced by the workers, software platform (Vakharia and Lease, 2013), task design (Eickhoff, 2018), and quality measures employed (Daniel *et al.*, 2019). In this chapter, we attempt to understand what setup is needed to effectively perform the task of snippet annotation.

Relevant annotation efforts include QuaC (Choi *et al.*, 2018), which is a dataset of QA dialogues. However, it is limited to sections of Wikipedia articles and contains only dialogues about a biased sample of entities of type person. Queries in CAsT datasets are much more diverse, both in terms of the expected type of answer and in the topics discussed. Most relevant to our work is the paper by Ren *et al.* (2021), where crowd workers are asked to respond to queries from the TREC CAsT'19 dataset while being presented with SERPs. The response generation task is divided into three stages: (optional) query rewriting, finding supporting sentences in results displayed on a SERP, and summarizing them into a short conversational response. We focus only on the supporting evidence-finding step, which is performed on a finer (snippet-level) granularity, and explore various task designs to ensure high data quality.

In the second part of this chapter, we leverage the collected data to train an unanswerability detector for response generation. The problem of unanswerability has been addressed in the context of machine reading comprehension (MRC) (Huang *et al.*, 2019a; Hu *et al.*, 2019) and extractive question-answering (QA) (Asai and Choi, 2021; Liao *et al.*, 2022; Godin *et al.*, 2019). Solutions proposed include answerability prediction using prompt-tuning (Liao *et al.*, 2022), modeling high-level semantic relationships between objects from question and context (Huang *et al.*, 2019a), and combining the output of reading and verification modules in MRC systems (Hu *et al.*, 2019; Zhang *et al.*, 2020a). Ren *et al.* (2021) acknowledge the challenge of unanswerability in conversational search; however, their approach does not explicitly address it. This work aims to bridge that gap by integrating unanswerability detection into the response

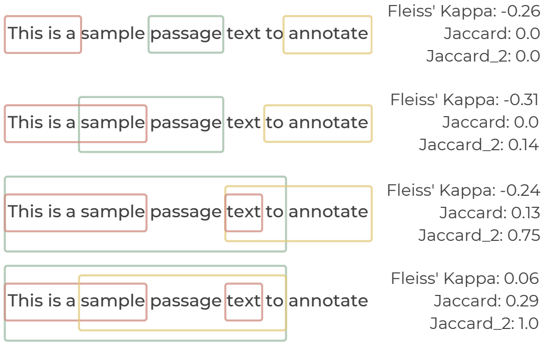


Figure 5.1: Visualization of annotations made by 3 workers on a sample text. The values of Fleiss' Kappa indicate poor agreement. On the other hand, Jaccard similarity offers more granular results which are easier to interpret in this specific scenario. Krippendorf's Alpha with nominal weight function gives analogous results to Fleiss' Kappa.

generation process. Our proposed solution for answerability prediction is based on a sentence-level classifier that is learned on CIS-specific training data, and can further be augmented with QA answerability data.

## 5.2 Evaluating Snippet Annotations

Traditional metrics for inter-annotator agreement such as Fleiss' Kappa or Krippendorff's Alpha are designed to assess categorical annotations and rely on a binary notion of agreement. In our case, we are more interested in measuring the degree to which snippets selected by different workers overlap (see Figure 5.1). We define evaluation measures to compare the agreement between annotators and across crowd workers and expert annotators.

### 5.2.1 Inter-annotator Agreement

We define inter-annotator agreement in terms of Jaccard similarity. Given an input text $t$ annotated by $n$ workers $(w_1, \dots, w_n)$, we count the length of the snippets chosen by all annotators and divide it by the length of snippets chosen by any annotator. Formally:

$$J(t) = \frac{|\bigcap_{i=1}^{n} snippets(t, w_i)|}{|\bigcup_{i=1}^{n} snippets(t, w_i)|} \;, \tag{5.1}$$

where $snippets(t, w_i)$ denotes the set of intervals selected by worker $w_i$ in text $t$. The intersection and union of snippet intervals are calculated on the character level.

We also consider a less strict variant of the measure, termed Jaccard$_k$, which takes only those intervals into account that are chosen by at least $k$ annotators.

Formally, in the numerator in Eq. (5.1) we count the length of intervals that appear in at least $k$ annotations made by different workers, while the denominator remains unchanged.

### 5.2.2 Similarity to Reference Annotations

To measure the similarity of snippet annotations by crowd workers against reference annotations by experts, we follow a logic similar to ROUGE-1, which considers the overlap of unigrams between the system and reference summaries (Lin, 2004). Specifically, we employ the ROUGE-like measures proposed in (Iskender *et al.*, 2021). For every input text $t$, we have annotations made by $n$ different crowd workers $(w_i)$ and reference annotations by $m$ different experts $(e_j)$. First, we define precision and recall of the snippets in text $t$ between a pair of annotators $w_i$ and $e_j$:

$$\begin{aligned} p_t^{i,j} &= \frac{|snippets(t, w_i) \cap snippets(t, e_j)|}{|snippets(t, w_i)|}, \\ r_t^{i,j} &= \frac{|snippets(t, w_i) \cap snippets(t, e_j)|}{|snippets(t, e_j)|}. \end{aligned}$$

We compute the F1 score as the harmonic mean of precision and recall: $f1_t^{i,j} = 2 \times p_t^{i,j} \times r_t^{i,j} / (p_t^{i,j} + r_t^{i,j})$.

Next, we aggregate these measures for a given crowd worker $i$ against all $(m)$ expert annotations: precision as $P_t^i = \frac{1}{m} \sum_{j=0}^{m} p_t^{i,j}$, recall as $R_t^i = \frac{1}{m} \sum_{j=0}^{m} r_t^{i,j}$, and F1 score as $F1_t^i = \frac{1}{m} \sum_{j=0}^{m} f1_t^{i,j}$.

Finally, we aggregate the annotations across all $(n)$ crowd workers in three different ways:

- *Mean* $(\overline{P}_t, \overline{R}_t, \overline{F1}_t)$, by simply averaging $P_t^i$, $R_t^i$, and $F1_t^i$ over all crowd workers.
- *Majority* $(P_t^{\gg}, R_t^{\gg}, F1_t^{\gg})$, where we consider a single crowd worker snippet annotation, which is taken as the union of intervals that are selected by the majority of workers.
- *Similarity* $(P_t^{\simeq}, R_t^{\simeq}, F1_t^{\simeq})$, where we only consider the snippet annotation by crowd worker $w_i$ that is most similar to the annotations of other crowd workers in terms of $f1_t^{i,j}$.

## 5.3 A Preliminary Study of Snippet Annotation

To ensure that we get high-quality snippet-level annotations, we first perform a preliminary study where we compare different task designs, platforms, and worker pools, by annotating two topics selected from the TREC CAsT'22 dataset, with markedly different characteristics, comprising 22 queries in total. The first

**Query:** I remember Glasgow hosting COP26 last year, but unfortunately I was out of the loop. What was the conference about?

**Passage:** HOME - UN Climate Change Conference (COP26) at the SEC – Glasgow 2021 Uniting the world to tackle climate change. The UK will host the 26th UN Climate Change Conference of the Parties (COP26) in Glasgow on 1 – 12 November 2021. The COP26 summit will bring parties together to accelerate action towards the goals of the Paris Agreement and the UN Framework Convention on Climate Change. The UK is committed to working with all countries and joining forces with civil society, companies and people on the frontline of climate change to inspire climate action ahead of COP26. COP26 @COP26 · May 25, 2021 1397069926800654339 We need to accelerate the #RaceToZero Join wef, MPPindustry, topnigel & gmunozabogabir for a series of events demonstrating the need for systemic change to accelerate the global transition to net zero. Starting May 27th Learn more #ClimateBreakthroughs | #COP26 Twitter 1397069926800654339 COP26 COP26 · May 24, 2021 1396737733649846273 #TechForOurPlanet is a new challenge programme for #CleanTech startups to pilot and showcase their solutions at #COP26! Innovators can apply to six challenges focusing around core climate issues and government priorities.

Figure 5.2: Sample from the CAsT-snippets dataset with highlighted expert annotations.

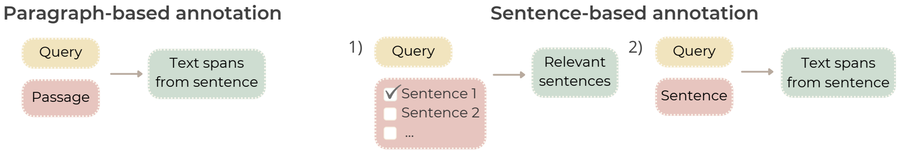


Figure 5.3: Illustration of different designs for the snippet annotation task.

topic (ID 132) has 12 turns (i.e., queries) and focuses on listing several independent pieces of information, which requires workers to choose multiple keywords or phrases within each passage. The second topic (ID 133) consists of 10 turns with questions about recipes, where several consecutive sentences contain relevant bits of information to be included in the answer.

### 5.3.1 Task Designs

We task crowd workers with the identification of snippets in a provided text that contains key pieces of the answer to a given query. Text snippets are required to be short, concise, informative, self-contained, and cannot overlap. Each snippet is supposed to contain one piece of information, so it can be treated as an information nugget. An example expert-annotated snippet is shown in Figure 5.2. Specifically, we identify snippets in paragraphs that have been labeled as relevant answers to the question. These passages can be long, which makes the annotation task cognitively demanding. Therefore, we consider two designs of the task: paragraph-based and sentence-based; see Figure 5.3.

In the *paragraph-based* annotation task, workers are asked to identify all text snippets in a given passage that are relevant to the input query. Since paragraphs can be lengthy, we also consider a simplified, *sentence-based* variant of this task, which lets workers operate on the significantly shorter text and

Table 5.1: Task configurations used for data collection; values are averaged for the annotation of a single paragraph.

| Task Variant | Annotator | Time | # workers | Acceptance rate | Cost |
|---|---|---|---|---|---|
| Paragraph | MTurk regular | 182s | 5 | 50% | $0.36 |
| | MTurk master | 63s | 5 | 90% | $0.38 |
| | Prolific | 154s | 5 | 79% | $0.51 |
| | Expert | 96s | 3 | - | - |
| Sentence | MTurk regular | 977s | 3 | 72% | $0.43 |
| | MTurk master | 305s | 3 | 87% | $0.56 |

enforces shorter text snippet selection. Specifically, the task is divided into: (1) relevant sentence selection, and (2) snippet annotation in relevant sentences. In sub-task (1), crowd workers are presented with a question and a passage that is split into sentences. They are asked to choose sentences that contain information relevant to the query. This is a straightforward task that resembles extractive summarization (Zhou *et al.*, 2018). Sub-task (2) is very similar to the paragraph-based annotation task, the only difference is that workers are presented with a relevant sentence instead of an entire passage.

### 5.3.2 Platforms and Workers

We set up the annotation task on two crowdsourcing platforms: Amazon MTurk[1] and Prolific.[2] MTurk offers an easily customizable web-based annotation interface and it is possible to filter workers based on qualifications. Prolific has more limited options in terms of the annotation interface, but the qualification of workers is claimed to be higher than on MTurk.[3] Additionally, we employ a group of expert annotators (Ph.D. students) who have been trained to perform this annotation task; they also use the MTurk platform, but in sandbox mode, i.e., without receiving payment.

### 5.3.3 Task Configurations

Table 5.1 shows the different task configurations we experiment with. The paragraph-based annotation task, which is regarded as the cognitively more demanding variant, is performed with workers from both crowdsourcing platforms as well as with expert annotators. The sentence-based variant of the task is executed only on MTurk. All tasks on MTurk are performed with both regular

[1] https://www.mturk.com/

[2] https://app.prolific.co/

[3] https://www.prolific.co/prolific-vs-mturk

Table 5.2: Inter-annotator agreements (J and $\mathrm{J}_{\boldsymbol{k}}$). The number of annotators for every input text is shown in parentheses.

| **Task Variant** | **Annotator** | **J** | $\mathbf{J}_{\boldsymbol{k}}$ | | |
|---|---|---|---|---|---|
| | | | $\boldsymbol{k}=\mathbf{4}$ | $\boldsymbol{k}=\mathbf{3}$ | $\boldsymbol{k}=\mathbf{2}$ |
| Paragraph | MTurk regular ($n=5$) | 0.02 | 0.08 | 0.21 | 0.48 |
| | MTurk master ($n=5$) | 0.18 | 0.35 | 0.53 | 0.73 |
| | Prolific ($n=5$) | 0.14 | 0.27 | 0.44 | 0.65 |
| | Expert ($m=3$) | 0.25 | - | - | 0.54 |
| Sentence | MTurk regular ($n=3$) | 0.35 | - | - | 0.71 |
| | MTurk master ($n=3$) | 0.47 | - | - | 0.76 |

and master workers.[4] The remuneration varies depending on the platform used, with payments on MTurk ensuring that the payment is above the US federal minimum wage and on Prolific being determined by the recommended payment per minute suggested on the platform. The reported numbers correspond to the average cost of annotating one paragraph (based on the average number of sentences per paragraph in the case of the sentence-based variant) and also include the platform fee. The total cost of this preliminary study was $1.2k. Table 5.1 also reports the number of workers assigned for each sample in the given task variant, the median time taken to annotate a paragraph, and the acceptance rate after a manual quality check. The acceptance rate is given only to tentatively present the difficulty of different variants of the task and it is not indicative of the quality of the final accepted annotations.

### 5.3.4 Quality Control

To ensure that the collected data is of the highest quality, we define several automatic quality control criteria before the final manual verification of results. In the paragraph-based task variant, annotations longer than 50% of the passage and annotations not contained in the intersection of the intervals chosen by at least two other crowd workers are flagged. In the first sub-task of the sentence-based variant, we flag submissions where fewer than one sentence or more than 75% of sentences are chosen. Additionally, only submissions that have at least one common sentence chosen with other crowd workers are accepted. The second sub-task applies the same quality control criteria as the paragraph-based variant, with the maximum length of the snippet increased to 75% of the sentence. Importantly, the automatic quality control criteria are only used to flag submissions that require additional attention. All results are manually verified by the author of this thesis, and responses that do not meet the task requirements are rejected.

[4] MTurk Master is a qualification earned through a proven track record of quality work.

Table 5.3: Similarity against reference (expert) annotations.

| **Task variant** | **Annotator** | $\overline{F1}$ | $F1^{\gg}$ | $F1^{\simeq}$ |
|---|---|---|---|---|
| Paragraph-based | MTurk regular | 0.36 | 0.32 | 0.45 |
| | MTurk master | **0.54** | **0.60** | **0.61** |
| | Prolific | 0.50 | 0.54 | 0.57 |
| Sentence-based | MTurk regular | 0.31 | 0.33 | 0.34 |
| | MTurk master | 0.41 | 0.43 | 0.44 |

### 5.3.5 Results

We report on the inter-annotator agreement and similarity against reference annotations on the two topics selected for this preliminary study in Tables 5.2 and 5.3 respectively.

In the paragraph-based variant, we observe better agreement ($J$) between MTurk masters than between Prolific workers, yet there is a big gap between crowd workers and experts. The relative ordering between workers is: MTurk masters > Prolific > MTurk regular, which also holds for the more relaxed version of the agreement measure ($J_k$). We notice that for $J_2$, the agreement between expert annotators is lower than for MTurk masters and Prolific workers; however, there are only 3 experts (vs. 5 crowd workers), hence it is not fair to directly compare these numbers. The generally low agreement scores highlight the difficulty of the task in the paragraph-based form.

On the simplified sentence-based variant, we indeed observe a much higher agreement between MTurk workers.[5] Also, the differences between regular workers and masters are not as large as in the paragraph-based variant. We note that the two task variants (sentence-based and paragraph-based) cannot be compared directly in terms of inter-annotator agreement because the probability of choosing the same snippets by different workers is much higher in a single sentence than in an entire paragraph.

Table 5.3 reports on the quality of worker annotations, with respect to their similarity to the reference (expert) annotations. These results are consistent for all measures and are also in line with the observations made in terms of inter-annotator agreement. Namely, MTurk masters achieve the best results, followed by Prolific workers, and then MTurk regular workers. The same holds for MTurk workers on the sentence-based variant of the task. We notice that the absolute scores are much closer for paragraph- and sentence-based annotations than for inter-annotator agreement (with sentence-based performing even slightly better on $F1^{\gg}$ for regular workers). Overall, we find that the paragraph-based variant yields higher-quality data than the sentence-based one.

[5] Given that MTurk masters outperformed Prolific workers in the paragraph-based variant, sentence-based annotations are only performed on MTurk.

### 5.3.6 Discussion

Our preliminary exploration of different task designs, platforms, and workers has led us to the conclusion that the highest-quality annotations for this specific task can be collected on the MTurk platform using a paragraph-based task design. The main challenge in collecting snippet annotations turned out to be the process of quality control that cannot be automated due to the nature of this task. Even for expert annotators, who performed the task attentively, the inter-annotator agreement is low. Therefore, a low similarity between snippets selected by a worker and reference annotations does not imply that the worker did an inferior job. Moving forward to collecting annotations at scale, we opt for recruiting a smaller group of crowd workers, using a qualification task, and working closely with them by providing continuous feedback on their work.

## 5.4 The CAsT-snippets Dataset

This section describes our large-scale data collection effort. We perform annotations on the TREC CAsT'20 and '22 datasets.[6] Each dataset comprises of a set of information-seeking dialogues (i.e., topics) with a sequence of questions (i.e., queries) within each. The input to the snippet annotation task consists of queries and corresponding passages. We consider the top 5 passages for each query with respect to their relevance labels in the ground truth (ranging from 0 to 4). If there are fewer than 5 passages available for the query at the highest relevance level, then we fill up the remaining slots with passages one relevance level below. If there are more passages available, then we cluster them using $k$-means clustering and pick a random passage per cluster. For example, if we have 3 highly relevant passages for a given query and 10 relevant passages, we choose all the passages with relevance level 4 and populate the remaining two places by splitting the passages with a relevance level 3 into two clusters and then choosing a random passage from each cluster. Selecting the passages for annotation this way ensures that they are both relevant and diverse. Even though we mostly consider highly relevant and relevant passages, some of them do not contain a direct answer to the question, which makes the snippet annotation task even more challenging. For each of the 371 queries in the TREC CAsT'20 and '22 datasets, the top 5 passages are annotated by 3 crowd workers, resulting in a total of 1,855 query-passage pairs.

### 5.4.1 Setup

The annotation task was released only to a small group of trained crowd workers, who were selected through a qualification task. The qualification task contained a detailed description of the problem at hand, examples of correct annotations, a quiz, and 10 query-passage pairs to be annotated; it was made available to

[6] The 2019 dataset has relatively low complexity compared to these two, while the 2021 dataset provides relevance assessments on the level of documents instead of passages.

both master and regular MTurk workers to reach a bigger audience. From the 20 workers who completed the qualification task, we chose 15 that had the highest quality results (independently of their MTurk Master qualification). Each worker received feedback on the provided responses and was given an opportunity to ask their own questions about the task. Several rounds of discussion that emerged from the qualification task resulted in an extended set of guidelines addressing the challenging aspects of the annotation task. They contain detailed instructions for crowd workers, a list of tricky cases along with recommendations on how to proceed, a brief description of the problems that we plan to address using the collected data, and toy examples illustrating how much context should be included in the span. The extended guidelines are available in Appendix B.

The process of data collection was divided into daily batches and conducted over approximately two weeks. The reason was to both avoid worker fatigue and also to allow for continuous feedback along the way. Each batch contained questions about one specific topic, which amounts to 46 query-passage pairs on average, and was annotated by 3 different workers. Workers received \$0.3 for each query-passage pair. A bonus of \$2 was paid for every batch completed within 24 hours upon release. The total cost of large-scale data collection was \$2.1k.

The training of the annotators did not end at the qualification task but continued throughout the whole data collection process. Crowd workers were provided with feedback after each submitted batch. From each batch, random data samples with low agreement were selected and verified manually by an expert (the author of this thesis). Incorrect data annotations were flagged and discussed individually with crowd workers. After each batch, general comments and suggestions were shared with all workers. We used Slack[7] as the main communication platform; there, workers could also share challenging cases and benefit collectively from discussions and from expert guidance. The Slack channel was widely used by crowd workers from their own initiative through the whole data annotations process to brainstorm about tricky query-passage pairs and align on their understanding of the task.

### 5.4.2 Statistics

In comparison to the results of the preliminary study (cf. Table 5.2) on the same set of queries, we find that the inter-annotator agreement ($J$=0.38 and $J_2$ =0.62) exceeds even that of expert annotations and the similarity with expert annotations ($\overline{\boldsymbol{F1}}$ =0.54) matches those of the best-performing MTurk master workers. These results indicate that the collected data is of high quality and attest to the success of our annotation setup with continuous feedback.

Table 5.4 provides a comparison against other related datasets. We note that there are not only more snippets annotated for each input text in our dataset,

[7] https://slack.com/

Table 5.4: Comparison against other datasets.

| **Dataset** | **Input text** | **Avg. snippet length (tokens)** | **# snippets per annotation** |
|---|---|---|---|
| CAsT-snippets | Paragraph | 39.6 | 2.3 |
| SaaC (Ren *et al.*, 2021) | Top 10 passages | 23.8 | 1.5 |
| QuaC (Choi *et al.*, 2018) | Wikipedia article | 14.6 | 1 |

but they are also longer on average, which follows from the information-seeking nature of queries.

We note that there is a number of query-passage pairs where annotators did not find any snippet relevant to the query, despite the passage being labeled as relevant by TREC assessors (77 such passages selected by all three annotators and 111 selected by two of the annotators).

### 5.4.3 Feedback from Crowd Workers

The close collaboration with crowd workers at every stage of data annotation has revealed several interesting aspects concerning the problem of snippet annotations. One of the most significant challenges was determining the appropriate amount of context to include in each span, striking a balance between conciseness and being self-contained. This issue is closely related to "conditional responses," where the span answers the question only under some specific condition or within a related situation (e.g., a medical condition is mentioned as a symptom, whereas the user was searching for treatment of that condition). Context also needs to be considered for justification of selected answers, particularly for yes/no responses. Moreover, temporal considerations, such as time mismatches between queries and passages, and the subjectivity of statements in the passage further compounded the challenge.

The second challenge identified by the crowd workers pertains to questions for which only a partial answer can be found in the passage. Deciding whether a span partially answers a question or is only somewhat relevant and should not be selected proved to be highly subjective. Additionally, the crowdsourcing process revealed that even passages with high relevance scores in the ground truth sometimes do not contain the exact answer to the question, resulting in cases of unanswerability.

The third noteworthy observation highlighted by several crowd workers concerns the background knowledge required to select a correct span or determine that the passage does not answer the question. This raises questions about the necessary general/expert knowledge required to annotate responses. Crowd workers worked on batches containing questions from specific areas, and while the TREC CAsT dataset assumes that the information needed to understand the question is included in the conversational context, some contextual knowledge may be missing if the system cannot find a highly relevant passage containing

Table 5.5: Statistics for the CAsT-answerability dataset.

| | **Answerable** | **Unanswerable** |
|---|---|---|
| #question-sentence pairs (train+test) | 6,395 | 19,043 |
| #question-passage pairs (train+test) | 1,778 | 1,932 |
| #question-ranking pairs (test) | 4,035 | 504 |

the answer. Moreover, even with access to previous questions and responses within a given topic, crowd workers still encounter challenges when annotating data from topics outside their areas of interest and expertise.

Our task design also included a confidence field for each annotation task, allowing workers to express their level of confidence in the selected spans. We analyzed the collected data to determine whether high confidence levels among workers corresponded to high inter-annotator agreement. Surprisingly, we did not observe any significant relationship between the two measures. We suspect that the confidence scores reported by crowd workers are more closely related to their familiarity with the topic in a given batch.

## 5.5 The CAsT-answerability Dataset

To prevent responses from being generated based on passages that lack relevant information—potentially leading to hallucinations—we aim to detect such cases in advance. To predict answerability in CIS dialogues, we build upon the CAsT-snippets dataset, which contains snippet-level annotations for the top 5 retrieved results. To balance the collection, we also include 5 randomly selected non-relevant passages to each question. The resulting dataset, named *CAsT-answerability*, contains around 1.8k answerable and 1.9k unanswerable question-passage pairs. We further consider answerability on the level of sentences and on the level of rankings, as follows. For sentence-level answerability, we leverage annotations of information nuggets from the CAsT-snippets dataset as follows: each sentence that overlaps with an information nugget, as per annotations in the originating CAsT-snippets dataset, is labeled as 1 (answer in the sentence), otherwise as 0 (no answer in the sentence).

For ranking-level answerability, which is the ultimate task we are addressing, we consider different input rankings, i.e., sets of $n = 3$ passages, for the same input question. Specifically, for each unique input test question (38), we generate all possible $n$-element subsets of passages available for this question (both containing and not containing an answer), thereby simulating passage rankings of varying quality. These rankings represent inputs with various degrees of difficulty for the same question, ranging from all passages containing an answer to a single passage with an answer to "no answer found in the corpus." This yields a total of 4.5k question-ranking pairs, of which 0.5k are unanswerable.[8]

[8] Examples of answerability scores on various levels are provided in Appendix B.

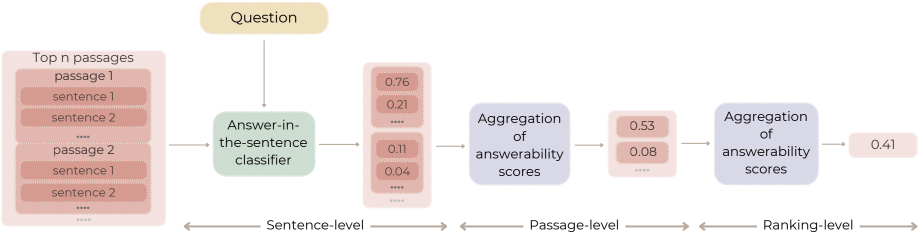


Figure 5.4: Overview of our answerability detection approach.

Overall, our CAsT-answerability dataset contains binary answerability labels on three levels: sentence, passage, and ranking. Sentence- and passage-level answerability is divided into train (90%), and test (10%) portions; the splitting is done on the question level to avoid information leakage. Ranking-level answerability has only a test set. See Table 5.5 for a summary.

## 5.6 Answerability Detection

The challenge of answerability in CIS arises from the fact that the answer is typically not confined to a single entity or text snippet, but rather spans across multiple sentences or even multiple passages. Note that answerability extends beyond the general notion of relevance and asks for the presence of a specific answer. At the core of our approach is a sentence-level classifier that can distinguish sentences that contribute to the answer from ones that do not. These sentence-level estimates are then aggregated on the passage level and then further on the ranking level (i.e., set of top-n passages) to determine whether the question is answerable; see Figure 5.4. Operating on the sentence level is a design decision that has the added benefit that a future response generation step may take a filtered set of sentences that contribute to the final answer as input.

### 5.6.1 Answer-in-the-Sentence Classifier

The answer-in-the-sentence classifier is trained on sentence-level data from the train portion of the CAsT-answerability dataset. In some of the experiments, this data is augmented by data from the SQuAD 2.0 dataset (Rajpurkar *et al.*, 2018) to provide the classifier with additional training material and thus guidance in terms of questions that can be answered with a short snippet contained in a single sentence. Data from SQuAD 2.0 is downsampled to be balanced in terms of the number of answerable and unanswerable question-sentence pairs. The classifier is built using a BERT transformer model with a sequence classification head on top.[9] Each data sample contains `question [SEP] sentence`

[9] https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertForSequenceClassification

as input and a binary answerability label. The output of the classifier is the probability that the sentence contains (part of) the answer to the question.

### 5.6.2 Aggregation of Sentence-level Answerability Scores

In reality, answers are not confined to a single sentence but can be spread across several passages. We thus need a method to aggregate results obtained from the sentence-level classifier to decide whether the question can be answered given (1) a particular passage or (2) a set of top-ranked passages, referred to as a ranking.

We consider two simple aggregation functions, *max* and *mean*, noting that more advanced score- and/or content-based fusion techniques could also be applied in the future (Kurland and Culpepper, 2018). Intuitively, *max* is expected to work particularly well for factoid questions where the answer is relatively short and usually contained in a single sentence, while *mean* should capture the cases where pieces of the answer are spread over several sentences within the passage or across passages. The aggregated answerability score for a given passage is compared against a fixed threshold; passages with an aggregated score exceeding this threshold are identified as containing the answer. We set the threshold values on a validation partition (10% of the dataset, sampled from the training partition); 0.5 for max and 0.25 for mean.

An analogous procedure is repeated for the top $n = 3$ passages in the ranking to decide on ranking-level answerability. Here, the aggregation methods take the passage-level answerability scores as input (obtained using max or mean aggregation of sentence-level probabilities). The resulting values are compared against a fixed threshold (using the same values as for passage-level aggregation) to yield a final ranking-level answerability prediction.

## 5.7 Answerability Prediction Results

Table 5.6 presents the answerability results on the sentence, passage, and ranking levels on the test partition of CAsT-answerability in terms of accuracy.

**Does data augmentation help answerability detection?** On the sentence level, we find that augmenting the CAsT-answerability dataset with additional training examples from SQuAD 2.0 improves performance. These improvements also carry over to the first aggregation step on the passage level. However, the best ranking-level results are obtained by aggregating results obtained from the classifier trained only on CAsT-answerability. It may result from the fact that SQuAD 2.0 training data focuses on questions with short-span answers (like entities or numbers) confined to a single sentence. This could mislead the classifier to overlook answers spanning multiple sentences or passages. Thus, while sentence-level answerability prediction benefits from augmented data, this does not translate to effective passage or ranking-level answerability prediction.

Table 5.6: Answerability detection results in terms of classification accuracy. The best scores for each level are in boldface. For the augmented classifier (rows 5–8), significant differences against the respective method (rows 1–4) are indicated by $^*$. ChatGPT results are tested against the best classifier in rows 1–8. We use McNemar's test with $p < 0.05$.

| **Classifier** | **Sentence** | **Passage** | | **Ranking** | |
|---|---|---|---|---|---|
| | **Acc.** | **Aggr.** | **Acc.** | **Aggr.** | **Acc.** |
| CAsT-answerability | 0.752 | Max | 0.634 | Max | 0.790 |
| | | | | Mean | **0.891** |
| | | Mean | 0.589 | Max | 0.332 |
| | | | | Mean | 0.829 |
| CAsT-answerability augmented with SQuAD 2.0 | **0.779**$^*$ | Max | 0.676$^*$ | Max | 0.810$^*$ |
| | | | | Mean | 0.848$^*$ |
| | | Mean | 0.639$^*$ | Max | 0.468$^*$ |
| | | | | Mean | 0.672$^*$ |
| ChatGPT passage-level (zero-shot) | | | **0.787**$^*$ | T=0.33 | 0.839$^*$ |
| | | | | T=0.66 | 0.623$^*$ |
| ChatGPT ranking-level (zero-shot) | | | | | 0.669$^*$ |
| ChatGPT ranking-level (two-shot) | | | | | 0.601$^*$ |

**Which of the two aggregation methods performs better?** In all cases, max aggregation on the passage level followed by mean aggregation on the ranking level gives the best results. Intuitively, this configuration captures single sentences with high answerability scores in individual passages (max aggregation on passage level) that give a high average score for the whole ranking (mean aggregation on ranking level) for answerable questions.

**How competitive are these baselines in absolute terms?** Ours is a novel task, with no established baselines to compare against. However, using a large language model (LLM) for generating the final response from the top retrieved passages is a natural choice. Therefore, for reference, we compare against a state-of-the-art LLM, using the most recent snapshot of GPT-3.5 (`gpt-3.5-turbo-0301`) via the ChatGPT API. We consider two settings: given a passage (analogous to the passage-level setup) and given a set of passages as input (analogous to the ranking-level setup). We prompt the model to verify whether the question is answerable in the provided passage(s) and return 0 or 1 accordingly.[10] In the passage-level setup, the passage-level predictions returned

[10] The prompts are available in Appendix B.

by ChatGPT are aggregated using fixed thresholds to obtain a ranking-level prediction. The max aggregation boils down to checking whether any of the passages is predicted to contain the answer. In the case of mean aggregation, a threshold of 0.33 or 0.66 (based on the fact that binary values are returned for passage-level answerability predictions) would mean that 1 or 2 out of 3 passages, respectively, contain the answer. In the ranking-level setup, we experiment with both a zero-shot setting, where neither examples nor context is given to the model, and a two-shot setting containing a question followed by two sentences (one positive and one negative example) extracted from the passage. We observe that the passage-level answerability scores of ChatGPT are higher than ours, but after ranking-level aggregation, it is no longer the case. Further, performing the ranking-level task directly results in significantly lower performance. These results indicate that LLMs have a limited ability to detect answerability without additional guidance. Our baseline methods trained on small datasets and based on simple classifiers with multi-step results aggregation turn out to be more effective for answerability prediction and thus represent a strong baseline.

## 5.8 Conclusions

We have introduced two datasets for conversational information seeking based on TREC CAsT'20 and '22 datasets: (1) CAsT-snippets, containing snippet-level annotations of top passages, and (2) CAsT-answerability, with answerability labels on sentence, passage and ranking levels. To answer **RQ3.1** *(How to identify core information units in the relevant passages that need to be included in the response?)*, we conducted a preliminary study to explore different task designs, platforms, and worker pools for snippet annotation. The insights from this study informed our decision to collaborate closely with a pool of highly engaged crowd workers, releasing tasks in daily batches and providing continuous feedback. The answerability labels in the CAsT-answerability dataset are derived from the information nugget annotations.

Our direct communication with crowd workers throughout the data annotation process revealed multiple challenges that need to be addressed in conversational response generation: (1) Selecting spans for questions when only a partial answer is present is challenging and appears to be highly subjective. (2) Temporal considerations may exclude some spans as they are not valid answers given the time specified in the query. However, assessing the temporal validity of text may be challenging based solely on short text passages without a larger context. (3) Passages originating from blogs or comments very often contain subjective opinions. Should such subjective opinions be marked up as answers? (4) What kind of background knowledge should be assumed when the passage does not contain a direct answer but the answer may be inferred from the text? (5) How much content is needed for open-ended questions? (6) When is evidence or additional information needed for a factoid question and when is an entity alone sufficient as an answer?

Unanswerable questions pose a challenge in conversational information seeking. The utility of the CAsT-snippets dataset has been demonstrated on the task of unanswerability detection. To answer **RQ2.2** *(How to detect factors contributing to incorrect, incomplete, or biased responses?)*, we have presented a baseline approach based on the idea of sentence-level answerability classification and multi-step results aggregation and evaluated multiple instantiations of this approach with different configurations on the CAsT-answerability dataset. Despite their simplicity, our baselines have been shown to outperform a state-of-the-art LLM on the task of answerability prediction.

In this chapter, we explored the use of snippet-level answer annotations for detecting unanswerable questions. However, this idea of operating on snippets (or information nuggets) has many additional applications related to enhancing the reliability and factual accuracy of responses in CIS. It supports the development of answer generation methods that are grounded in relevant snippets in paragraphs, it allows for the automatic evaluation of the generated answers in terms of completeness (see Chapter 6), and it enables more granular source attribution (see Chapter 7).

# Chapter 6

# Grounded Response Generation

*If you do not know where you come from, then you don't know where you are, and if you don't know where you are, then you don't know where you're going. And if you don't know where you're going, you're probably going wrong.*

— **Terry Pratchett**

Current commercial generative search engines often appear informative but frequently contain unsupported statements and inaccurate citations, highlighting the difficulty of achieving grounded responses (Liu *et al.*, 2023a). Injecting evidence into LLM prompts for retrieval-augmented generation (RAG), similar to a retrieve-then-generate pipeline, significantly influences answers to fully mitigate hallucinations or ungrounded responses in systems like ChatGPT (Koopman and Zuccon, 2023; Lewis *et al.*, 2020). However, redundant information and overly long contexts can lead to the "lost in the middle" problem, where models experience significantly degraded performance when they need to access relevant information in the middle of long contexts (Liu *et al.*, 2024). Consequently, post-retrieval efforts focus on selecting essential information, emphasizing critical sections, and shortening the context to avoid information overload diluting key details with irrelevant content (Gao *et al.*, 2023b).

In conversational information-seeking (CIS) systems that limit responses to a few sentences, information about the scope of the answer and the extent to which it is covered is often hidden from the user. In the space of complex, often exploratory queries, there exists a natural trade-off between the completeness and succinctness of the responses. The extent to which a given topic is covered in the response, both in terms of breadth of diverse information and in-depth details, needs to be determined by the system based on the retrieved sources, user preferences, and previous interactions (Gienapp *et al.*, 2024). The information about the fraction of answers covered is essential for users to decide on the following interactions with the system (Azzopardi *et al.*, 2018). A system that generates the response and is aware of the part of relevant search space the information covers can suggest possible follow-up questions, for addressing additional details or aspects that did not fit in the provided response due to length constraints but may be of interest to the users. System transparency in this regard can help the user navigate the search space and facilitate interaction with the system when addressing complex information needs.

In this chapter, we aim to address the following research question: **How to ensure the grounding of responses in the retrieved sources? (RQ3.2)**. We introduce a modular pipeline for **G**rounded **I**nformation **N**ugget-based **GE**neration of Conversational Information-Seeking **R**esponses (GINGER) that operates on information nuggets—minimal, atomic units of relevant information (Pavlu *et al.*, 2012)—to ensure that responses are rooted in factual evidence and easily verifiable. The multistage pipeline encompasses nugget detection, clustering, ranking, summarization of top clusters, fluency enhancement, and follow-up question generation based on uncovered aspects of the topic. Our approach uniquely addresses three key challenges in CIS:

- *Grounding*: By operating on information nuggets throughout the pipeline, we ensure the grounding of the final response in the source passages and enable easy verifiability of source attribution.
- *Response completeness*: Our method offers control over response completeness, by ensuring the coverage of a required number of query facets in the response within a predefined length limit. This allows adaptability to user preferences, desired diversity, or conversational context.
- *Follow-up questions*: By explicitly modeling information nuggets related to different facets of the query, GINGER can suggest relevant (and answerable) follow-up questions based on facets that could not be covered in the response due to length constraints.

These features can significantly enhance the user experience in conversational search scenarios, mitigating the potential information loss when transitioning from traditional SERP-based interactions.

We introduce a competitive response generation baseline inspired by retrieval-augmented open-domain question answering to compare with GINGER, answering the following research question: **What are strong baselines for response generation in CIS systems? (RQ1b)**. We evaluate our response generation method using the TREC CAsT'20 and '22 datasets through both automatic and human evaluation. For automatic evaluation, we assess response grounding, faithfulness, answer relevance, and completeness using natural language inference and LLM-based RAG evaluation techniques. For human evaluation, we conduct side-by-side comparisons via crowdsourcing, measuring coherence, correctness, sufficiency, conciseness, engagement, and completeness of responses. Additionally, we extend our evaluation to the TREC RAG'24 dataset, analyzing the core response generation capabilities of GINGER using the AutoNuggetizer evaluation framework introduced in the track (Pradeep *et al.*, 2024). We perform an ablation study to examine the impact of different system components and compare GINGER's results against the top-performing systems submitted to the RAG'24 track.

All the resources developed in this chapter are available online: https://github.com/iai-group/ginger-response-generation/. Additional results and analysis can be found in Appendix C.

This chapter is based on the following papers:

Łajewska and Balog (2025): *GINGER: Grounded Information Nugget-Based Generation of Responses*, SIGIR '25

Łajewska and Balog (2024b): *The University of Stavanger (IAI) at the TREC 2024 Retrieval-Augmented Generation Track*, TREC '24

Łajewska and Balog: *X-GINGER: Explainable and Grounded Conversational Response Generation* [submitted]

## 6.1 Related Work

**Response Grounding** Existing search engine responses often exhibit high fluency and perceived utility but frequently contain unsupported statements or inaccurate citations (Liu *et al.*, 2023a) (see Section 2.2.2). Despite advancements in LLMs, abstractive summaries still suffer from hallucinations and factual errors (Ladhak *et al.*, 2022; Tang *et al.*, 2023; Falke *et al.*, 2019; Tang *et al.*, 2022; Ji *et al.*, 2023; Koopman and Zuccon, 2023). Source attribution, which measures the accuracy and support of generated statements through citations (Rashkin *et al.*, 2021), and verifiability, which requires that each statement is fully supported by in-line citations, are key concepts to address these challenges (Liu *et al.*, 2023a; Schuster *et al.*, 2023). Systems with high citation precision might lack fluency, while those with lower precision risk misleading users by appearing more fluent and relevant (Liu *et al.*, 2023a). To ensure high citation precision while maintaining fluency, we propose a method that extracts and groups atomic statements from sources, summarizing them with LLMs. Statements are referred to as "atomic/semantic content units" (Nenkova *et al.*, 2007; Liu *et al.*, 2023b) or "information nuggets" in traditional IR (Pavlu *et al.*, 2012; Sakai, 2023). The Knowledge Selection of Large Language Models (KS-LLM) method, which identifies valuable information from evidence documents using triples and evidence sentences, is similar to using information nuggets for response generation but leaves knowledge synthesizing to LLMs, which are prone to factual errors (Zheng *et al.*, 2024).

**Follow-up Questions** Clarifying questions refine the understanding of the initial query, while follow-up questions build upon the given information to explore related topics further. Nevertheless, they serve the same goal of supporting the user in navigating search space and creating a more effective and user-focused conversational experience. A common approach for generating clarifying questions is based on query facet detection (Wang, Zhenduo *et al.*, 2023; Samarinas *et al.*, 2022). In this work, we explicitly model facets of the response and generate follow-up questions based on those that cannot be included due to length constraints.

**TREC RAG'24** The RAG track at TREC has been launched in 2024 with a focus on combining retrieval methods for finding relevant information within large corpora with LLMs to enhance the ability of systems to produce relevant, accurate, and contextually appropriate content (Pradeep *et al.*, 2024). The track is divided into three tasks: Retrieval (R), which involves ranking and retrieving the most relevant segments from the corpus; Augmented Generation (AG), which requires generating RAG answers using top-k relevant segments from a baseline retrieval system provided by organizers; and Retrieval-Augmented Generation (RAG), where participants generate RAG answers with attributions using their retrieval system and chunking technique. Our work focuses on the augmented generation task, similar to in-context retrieval augmented

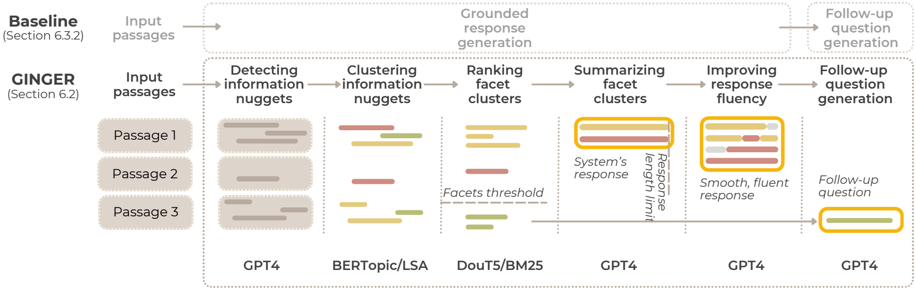


Figure 6.1: High-level overview of our nugget-based response generation pipeline, called GINGER.

language models without modifying model weights (Ram *et al.*, 2023; Muhlgay *et al.*, 2023). A trend toward retrieval context curation is evident in TREC RAG'24 submissions. A common approach to AG involves generating responses based on the top 20 retrieved documents, often in a single step using proprietary models, with an optional post-processing phase. Several submissions adopt a multi-step approach, such as segment clustering followed by extracting, combining, and condensing relevant information (Fröbe *et al.*, 2024). Similarly, some approaches emphasize verifying key facts across multiple documents, eliminating redundant content, prioritizing facts by relevance, and enhancing clarity and coherence (Farzi and Dietz, 2024b). In this work, we take this a step further by not only curating the LLM context but also decomposing the response generation process to mitigate the negative effects of irrelevant and redundant information by operating on atomic pieces of information.

## 6.2 GINGER: Nugget-Based Multi-Step Response Generation

We present GINGER, a novel method for generating grounded conversational responses by operating on information nuggets. GINGER explicitly models various facets of the query based on retrieved information and generates a concise response that adheres to length constraints. Additionally, it proactively suggests follow-up questions to help users navigate complex information needs, based on aspects that are not covered in the response due to the length limit, inherently supporting mixed-initiative conversations.

Generating grounded, completeness-aware responses is a multistage process, illustrated in Figure 6.1, that includes: (1) detecting information nuggets in top relevant passages, (2) clustering detected nuggets, corresponding to different facets of the query, (3) ranking the clusters with respect to their relevance to the query, (4) summarizing the top-ranked clusters to be included in a final response, (5) refining the response to improve its fluency and coherence, and (6) generating a follow-up question based on the top facet cluster not included in the response. Steps 1-3 aim at curating the context for response generation to mitigate the "lost in the middle" problem related to LLMs, focusing mostly on the beginning and the end of long texts (Liu *et al.*, 2024; Gao *et al.*, 2023b). By operating on information nuggets in all intermediate components of the pipeline we ensure the grounding of the final response in the source passages, ensuring that all information in the final response is entailed by the source (Falke *et al.*, 2019).

We assume that a de-contextualized query and the corresponding ranking of passages with relevance scores are provided as input, as query rewriting and passage retrieval are not the focus of this work. The prompts used by different components of the system are made available in Appendix C.

### 6.2.1 Detecting Information Nuggets

We aim to automatically detect information nuggets using an LLM. The LLM is prompted to annotate input passages with information nuggets containing the key information that answers the query. Specifically, it is instructed to copy the text of the passage and place the annotated information nuggets between specific tags, without modifying the passage content or adding any extra symbols.

### 6.2.2 Clustering Information Nuggets

Next, we proceed to cluster the detected information nuggets with respect to different facets of the query topic. Clustering information nuggets has two main purposes. First, it addresses the problem of information redundancy, which originates from the fact that information nuggets and their variants can appear in multiple documents in different forms but still convey the same information (Pavlu *et al.*, 2012). Second, by clustering redundant information nuggets, we attempt to increase the information density of the generated information (Adams *et al.*, 2023). Nugget clustering is challenging due to the semantic closeness of nuggets within the same topic. We address this by employing a neural topic modeling technique, BERTopic (Grootendorst, 2022), and adjusting its sensitivity on the validation partition of the CIS dataset in order to distinguish nuanced differences between nuggets. Ideally, information nuggets in each cluster represent specific facets of the answer to the query.

### 6.2.3 Ranking Facet Clusters

This step in the pipeline is responsible for the ranking of facet clusters with respect to the input query to determine which clusters are most important and should be prioritized for inclusion in the response, and which may be skipped (Gao *et al.*, 2023b; Liu *et al.*, 2024). Given the relatively low number of facet clusters we observe in practice, we can employ more expensive reranking techniques relying on pairwise comparisons to maximize effectiveness. Specifically, we employ pairwise reranking using duoT5 (Pradeep *et al.*, 2021) by joining the nuggets in clusters and treating them as individual passages.

### 6.2.4 Summarizing Facet Clusters

The response is made up of the summaries of the top $n$ clusters, where $n$ is the facet threshold. This threshold controls the desired response length and may be adjusted based on the information need, task context, or user preferences. Each cluster of information nuggets is summarized independently as a single sentence with the maximum number of words specified in the prompt, to stop the LLM from generating very long sentences (Goyal *et al.*, 2023). We follow the prompt design used for short story summarization (Subbiah *et al.*, 2024) to generate summaries that are short, concise, and only contain the information provided. Previous steps in the pipeline ensure that the most relevant information from retrieved passages is synthesized and the generated summaries are attributed to the sources, allowing summarization to operate in a shorter but more relevant context.

### 6.2.5 Improving Response Fluency

Our modular approach results in a response that is a concatenation of independent summaries of facet clusters, that may lack fluency and consistency. To mitigate this shortcoming, we include an additional step to rephrase the generated response with the help of an LLM. The LLM is prompted not to modify the provided information, nor include any additional content.

### 6.2.6 Follow-up Question Generation

Follow-up questions are generated with an LLM by prompting it to generate a follow-up question based on the original query and the provided aspect of the topic. The follow-up question should start with *"Would you like to learn more about ..."* or an equivalent with the same meaning (Sekulić *et al.*, 2021). The follow-up question targets the aspect of the $n + 1$th facet cluster, that is, the most relevant facet cluster that is not included in the response. This setup ensures that the follow-up question is both relevant to the query topic and is answerable.

## 6.3 Experimental Setup

This section presents the dataset and baselines used in our evaluation and provides technical details on the implementation of `GINGER`.

### 6.3.1 Dataset

We base our evaluation on the test partition of the CAsT-snippets dataset (see Section 5.4). It comprises 44 queries from the TREC CAsT'20 and '22 benchmarks (Dalton *et al.*, 2020; Owoicho *et al.*, 2022) along with information nugget annotations for the top 5 most relevant passages. Queries with no or only one nugget are excluded from the evaluation, as they are either incomparable to the baseline or fall under the problem of query unanswerability, which is covered in Chapter 5.

This work focuses on the problem of response generation and assumes a ranked list of passages to be provided, along with the query, as input. However, the relevance of these passages impacts how well the method grounds responses in specific facts; the absence of direct evidence increases the likelihood of unsupported statements (Liu *et al.*, 2023a). Therefore, we consider three types of input rankings to control for their quality. *Relevant* represents an idealized scenario where the input consists of the top 5 most relevant passages; this corresponds to the setting where the most relevant passages have been correctly retrieved from the corpus. We use this as our default setting, ensuring that we evaluate methods based on their core capabilities, independent of potential noise introduced by passage ranking. *Irrelevant* corresponds to a situation where the answer is not present in the corpus or the retriever is unable to find it. Here, the input contains 5 passages with low relevance scores. In both cases, the passages are selected based on the relevance judgments in the TREC CAsT benchmark. Finally, *retrieved* represents the real-world case where the top 5 passages from a competitive passage ranker are considered. Specifically, we use a multi-stage retrieval system that employs a T5-based query rewriter (fine-tuned on the CANARD dataset and expanded with terms from pseudo-relevance feedback), BM25 for first-pass retrieval, and reranking performed with MonoT5 and DuoT5.

Additional analysis is performed on the Augmented Generation task of TREC RAG'24 (Pradeep *et al.*, 2024) to evaluate `GINGER` against the most recent response generation systems. In all experiments, we utilize the top relevant passages (from MS MARCO V2.1 segment collection) from a fixed list of 100 retrieved results provided by the organizers for all 301 queries. This setup represents the real-world case with the top passages retrieved by a competitive passage ranker. To account for the amount of input information, we consider three sizes of input rankings containing 5, 10, or 20 passages.

### 6.3.2 Baseline

The focus of our method is twofold: (1) the generation of grounded responses, and (2) the generation of useful follow-up questions. To the best of our knowledge, there is no established approach for the combination of these tasks in a CIS setting. Therefore, we compare our system against methods for grounded open-domain QA (Ren *et al.*, 2025) and clarifying question generation (Samarinas *et al.*, 2022) as effective solutions for addressing response and follow-up question generation, respectively.

#### Grounded Response Generation

We aim to compare our proposed method against models that use external knowledge in the generation process. We exclude standard RAG models trained end-to-end with the retrieval component (Lewis *et al.*, 2020; Guu *et al.*, 2020;

Izacard *et al.*, 2023) as our focus is on grounded response generation with a fixed retriever to enable the evaluation of entailment in retrieved facts. We also refrain from using LLM-based generation approaches that rely solely on internal model knowledge (Sun *et al.*, 2022) due to our emphasis on grounding.

We explored the performance of grounded text generation, text summarization, and open-domain QA models in terms of faithfulness on the validation partition of the CAsT-snippets dataset to select the most competitive baseline for our scenario. We found that a pre-trained model proposed for grounded text generation in dialogues that rely on external knowledge (Peng *et al.*, 2022) tended to copy words or phrases directly from the source text, exhibiting a more extractive behavior, which is not desired in a CIS setting where the information from multiple sources needs to be aggregated. Considerably better results in terms of coherence and naturalness on the validation sample were observed when using the approach proposed for open-domain QA in a retrieval-augmented setting with an off-the-shelf LLM without further training (Ren *et al.*, 2025; Ram *et al.*, 2023; Muhlgay *et al.*, 2023). The most recent at the time of writing snapshot of OpenAI's GPT-4 model (`gpt-4-turbo-2024-04-09`) that achieves the highest scores in terms of faithfulness on the task of summary generation (Subbiah *et al.*, 2024) and is the most commonly used LLM architecture in RAG (Ren *et al.*, 2025; Ram *et al.*, 2023; Shi *et al.*, 2024; Huang and Huang, 2024; Muhlgay *et al.*, 2023) is therefore used as a baseline for grounded response generation. Our prompt is inspired by retrieval-augmented QA LLM instruction proposed in Ren *et al.* (2025). The length of the generated summary is limited to around 100 words and 3 sentences, which is controlled in task model prompt (Goyal *et al.*, 2023).[1] The second baseline uses GPT-4 with Chain-of-Thought prompting (Wei *et al.*, 2022) and one ICL demonstration created manually based on a sample from TREC CAsT'22 dataset (`baseline_CoT-top5`).

### Follow-up Question Generation

As a baseline for follow-up question generation, we employ a well-established method for generating clarifying questions. It comprises two steps: (1) facet extraction and (2) template-based question generation. For the first step, we aggregate facets extracted using sequence labeling and extreme multi-label classification, along with facets generated using auto-regressive text generation, based on relevance and diversity (Samarinas *et al.*, 2022). Combining facets generated by diverse methods, which yield complementary results, leads to significant improvements (Samarinas *et al.*, 2022), resulting in a competitive baseline. For step (2), we follow a query- and facet-conditioned clarifying question generation method (Sekulić *et al.*, 2021). However, instead of the fine-tuned GPT-2 model used in the original paper, we utilize the latest version of GPT-4 for improved fluency and naturalness. Additionally, we resort to a widely used template-based approach for clarifying question construction, providing the LLM with a template to ensure that the specific facet is the focus of the generated ques-

[1] Prompt is available in Appendix C.

tion (Zamani *et al.*, 2020; Sekulić *et al.*, 2022). We use the same follow-up question template in both the baseline and our method (see Section 6.3.2), with the only difference being the facet included in the prompt. This allows us to isolate and evaluate the quality of the facet itself.

### 6.3.3 `GINGER` Implementation

The modular design of `GINGER` allows for independent implementation of individual components of the pipeline. The detection of information nuggets is performed with GPT-4, with the query and passage as input. Information nuggets clustering is based on BERTopic, with parameters set experimentally on samples from the validation partition.[2] The ranking of information nugget clusters is done using duoT5, implemented based on the HuggingFace transformers library and the `castorini/duot5-base-msmarco` model.[3] The top 3 ranked information nuggets clusters are passed to query-based summarization (Tombros and Sanderson, 1998) performed with GPT-4. The length of each cluster summary is limited to one sentence and around 35 words and specified in the prompt (Goyal *et al.*, 2023). To generate a follow-up question based on the most relevant information not included in the response, we utilize GPT-4 with the same prompt as in the baseline method (see Section 6.3.2). The aspect of the topic provided in the prompt is the summary of the most relevant information cluster that is not included in the response. If there are less than 4 clusters for a given query, we use the last cluster summary from the ranking. In all methods, GPT-4 corresponds to the most recent snapshot of GPT-4-turbo (`gpt-4-turbo-2024-04-09`) accessed via the OpenAI API.

## 6.4 Evaluation Methodology

We conduct automatic evaluations of response grounding, faithfulness, answer relevance, and completeness, alongside human evaluations of coherence, correctness, sufficiency, conciseness, engagement, and completeness. To the best of our knowledge, there is no established framework for joint evaluation of response and follow-up question generation. Therefore, we propose a set of automatic metrics inspired by the evaluation of grounded summaries, LLM-based RAG evaluation, and nugget-oriented properties of our method. For human evaluation, we follow the SWAN framework proposed for conversational systems (Sakai, 2023).

### 6.4.1 Automatic Evaluation

Automatic evaluation of responses is limited to reference-free metrics, without comparing to a ground truth (Gienapp *et al.*, 2024), in the absence of CIS

[2] If fewer than four information nuggets are identified in the top $n$ passages, we skip clustering; instead, we treat each nugget as an independent cluster and proceed with standard ranking and summarization.

[3] `https://huggingface.co/castorini/duot5-base-msmarco`

datasets with ground-truth responses. Inspired by human evaluation protocols for summary salience that use fine-grained content units (Liu *et al.*, 2023b) corresponding to Atomic Content Units in the Pyramid protocol (Nenkova and Passonneau, 2004), our automatic evaluation of grounding and completeness operates on information nuggets annotated in the CAsT-snippets dataset.

### Grounding

In the TREC'06 QA task and the Pyramid method, answers and summaries are evaluated based on their coverage of relevant information nuggets or Semantic Content Units (SCUs), with a focus on recall to derive overall system scores (Nenkova and Passonneau, 2004; Shapira *et al.*, 2019; Bhandari *et al.*, 2020). We follow a similar approach and evaluate response grounding by assessing the entailment of each generated response against the automatically detected information nuggets in input passages (Pavlu *et al.*, 2012; Liu *et al.*, 2023b; Falke *et al.*, 2019). To evaluate the entailment of generated responses, we predict the probability that information nugget is entailed or contradicted by a generated response with Natural Language Inference (NLI) (MacCartney and Manning, 2008). Our implementation of NLI is based on a RoBERTa model[4] trained on the SNLI (Bowman *et al.*, 2015) and MultiNLI[5] datasets.

### Faithfulness and Answer Relevance

To evaluate the factual correctness of the generated text we utilize the Retrieval Augmented Generation Assessment (RAGAs) framework for reference-free evaluation of RAG pipelines (Es *et al.*, 2024). The framework measures faithfulness and answer relevance by prompting an LLM. Faithfulness is defined as the accuracy with which the generated content reflects the information in the retrieved documents, ensuring the generation process avoids misinformation, while answer relevance evaluates whether the response directly addresses the question, not taking into account factuality, but penalizing incompleteness and redundant information.

### Completeness

Completeness measures the extent to which the information need is addressed in the response provided by the system. It can be computed with respect to the top retrieved passages, the whole corpus of documents, or external knowledge. Response completeness can be achieved at the cost of reduced succinctness, especially for complex or exploratory queries. In our proposed method, completeness can be controlled by manipulating the number of passages, the granularity of facet clustering, and the number of facet clusters included in the response. We

[4] https://huggingface.co/cross-encoder/nli-roberta-base

[5] https://cims.nyu.edu/~sbowman/multinli/

measure response completeness in terms of the number of ground-truth information nuggets (provided as part of the CAsT-snippets dataset) entailed by the response. Entailment is calculated using NLI based on the RoBERTa model (same as for grounding evaluation).

### AutoNuggetizer

We use the AutoNuggetizer framework proposed for RAG evaluation and validated during TREC RAG'24 (Pradeep *et al.*, 2024). AutoNuggetizer comprises two steps: nugget creation and nugget assignment. In nugget creation, nuggets are formulated based on relevant documents and classified as either "vital" or "okay" (Voorhees, 2004). The second step, nugget assignment, involves assessing whether a system's response contains specific nuggets from the answer key. The score $V_{strict}$ for system's response is defined as follows:

$$V_{strict} = \frac{\sum_i ss_i^v}{|n^v|}$$

where $n^v$ represents the subset of the vital nuggets; $ss_i^v$ is 1 if the response supports the $i$-th nugget and is 0 otherwise. The score of a system is the mean of the scores across all queries.

We reimplemented the AutoNuggetizer evaluation framework to compute the $V_{strict}$ measure, adhering to the original prompts from Pradeep *et al.* (2024). To make the evaluation more robust and mitigate any potential bias of using the same LLM for response generation and judging, we use the average of the scores generated by three different LLMs (`gpt-4o-2024-08-06`, `claude-3-5-haiku-20241022`, `gemini-1.5-flash`) as the final score (as opposed to the original TREC RAG scores, which are based solely on GPT-4o). We validate our implementation of the evaluation framework by comparing the results for our submitted runs with the official numbers reported by track organizers (Pradeep *et al.*, 2024) (see column "TREC" vs column "avg LLM" in Table 6.7). Even though our V_strict scores are higher than the scores reported in the TREC RAG track, the relative ordering of the systems remains the same.

## 6.4.2 Human Evaluation

The generated responses are assessed by human evaluators along various dimensions to evaluate the performance of the proposed method. We use reference-free evaluation which instructs annotators to assess the response directly (Fabbri *et al.*, 2021) as, to the best of our knowledge, there is no dataset with ground-truth responses in the CIS domain. To determine which response is better, we use pairwise comparison instead of individual ordinal evaluations, focusing on relative rather than absolute measures (Kelly, 2007). In the side-by-side evaluation, users are asked to choose their preferred response and follow-up question with respect to a given dimension, which is a common setup for collecting preference annotations in summarization (Goyal *et al.*, 2023) and clarifying question evaluation (Sekulić *et al.*, 2021). To ensure independence of the collected scores we allow each crowd worker to complete only one task.

### Response Dimensions

The main motivations for GINGER are (1) to ensure grounding of the response in specific facts, including in it as many unique informative bits as possible given the length limit, and (2) to generate useful follow-up questions. To measure the effectiveness of our method in achieving that, we utilize the SWAN evaluation framework (Sakai, 2023). We select five response dimensions to capture the main functionalities of our proposed method for human assessment: coherence, correctness, sufficiency, conciseness, and engagement.[6] The focus of the human evaluation is the general quality of the response and follow-up question with the goal of capturing users' preferences towards baseline or our proposed method.

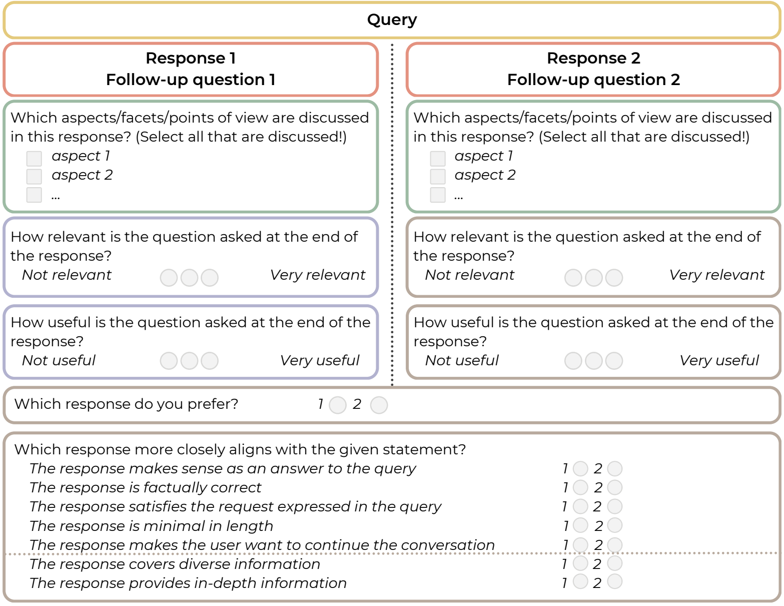
To verify the effectiveness of our method in controlling the completeness of the responses, we run an additional human evaluation task where we compare a *'broad'* response briefly covering different facets of the topic but missing details about different aspects with a *'deep'* response that focuses on one aspect of the answer and discusses it in details at a cost of diversity. The broad response is generated by the standard GINGER setup and contains a one-sentence summary for the top three clusters. The deep response is represented by a three-sentence-long summary of the top cluster. The human evaluation task setup is the same with two additional response dimensions added in the last questionnaire that correspond to information breadth (*The response covers diverse information*) and information depth (*The response provides in-depth information*). We use the collected human scores to investigate the relation between the automatically computed completeness score of the response and user-reported breadth and depth of the provided information (Gienapp *et al.*, 2024). Statements used for information breadth and depth are validated in the pilot study where we explored three different formulations for each dimension.[7]

### Study Design

Human assessments are collected for responses generated with different configurations of our method paired with a baseline response. Each response evaluation task consists of a query, two variants of the response, corresponding attentiveness checks, and questions about response dimensions (see Figure 6.2). One response is generated by the baseline method, while the second response is produced using GINGER. Each response is accompanied by the corresponding follow-

[6] We acknowledge that there are other evaluation criteria that are not taken into account in our experiments (Sakai, 2023; Gienapp *et al.*, 2024). We skip criteria that include verification of the sources and we evaluate response grounding with automatic metrics. Similarly, we do not ask crowd workers about dimensions related to the conversation and leave it for future work. The operational definitions of these response dimensions are presented in the bottom questionnaire in Figure 6.2.

[7] More details can be found in Appendix C.



Query

| Response 1<br>Follow-up question 1 | Response 2<br>Follow-up question 2 |
|---|---|
| Which aspects/facets/points of view are discussed in this response? (Select all that are discussed!)<br>□ *aspect 1*<br>□ *aspect 2*<br>□ ... | Which aspects/facets/points of view are discussed in this response? (Select all that are discussed!)<br>□ *aspect 1*<br>□ *aspect 2*<br>□ ... |
| How relevant is the question asked at the end of the response?<br>*Not relevant* ○○○ *Very relevant* | How relevant is the question asked at the end of the response?<br>*Not relevant* ○○○ *Very relevant* |
| How useful is the question asked at the end of the response?<br>*Not useful* ○○○ *Very useful* | How useful is the question asked at the end of the response?<br>*Not useful* ○○○ *Very useful* |

Which response do you prefer? *1* ○ *2* ○

Which response more closely aligns with the given statement?

| | |
|---|---|
| *The response makes sense as an answer to the query* | *1* ○ *2* ○ |
| *The response is factually correct* | *1* ○ *2* ○ |
| *The response satisfies the request expressed in the query* | *1* ○ *2* ○ |
| *The response is minimal in length* | *1* ○ *2* ○ |
| *The response makes the user want to continue the conversation* | *1* ○ *2* ○ |
| *The response covers diverse information* | *1* ○ *2* ○ |
| *The response provides in-depth information* | *1* ○ *2* ○ |

Figure 6.2: Design of the response evaluation user study. The last two questions in the questionnaire separated with dotted line are used only in the completeness evaluation.

up question. This setup allows for a relative comparison of different variants of GINGER with the baseline method. Following each response, an attentiveness check is presented to the crowd worker. This check involves a list of four aspects, facets, or points of view generated by GPT-4, from which the worker must select all that are discussed in the response. The model is prompted to create two lists of aspects related to the topic of the provided passage. The first list contains 2-5 items covered in the passage, while the second list contains 2-5 items not covered in the passage. The attentiveness check question includes four aspects randomly sampled so that at least one aspect is discussed in the response. The main part of the task contains questions about the user's preferences towards responses and the quality of the follow-up question. Each follow-up question is evaluated in terms of relevance (Samarinas *et al.*, 2022) and usefulness (Sekulić *et al.*, 2021; Wang, Zhenduo *et al.*, 2023) on a 3-point Likert scale. Then, crowd workers are asked to reveal their general preference towards responses not to introduce any biases about the dimensions they should focus on (Goyal *et al.*, 2023). The last part of the task focuses on pair-wise evaluation of the responses along the dimensions selected from the SWAN framework (Sakai, 2023).

### Study Execution

Altogether, we create 44 tasks for human evaluation, each corresponding to a query from the test partition of the CAsT-snippets dataset. Each task is com-

Table 6.1: Automatic evaluation of responses that measures target grounding (nugget entailment (Ent.) and contradiction (Cont.)) and completeness (Comp.)). Statistically significant differences ($p < 0.05$) with respect to the baseline are marked with * (t-test). The best scores for each measure are boldfaced.

| **Method** | **Automatic evaluation** | | |
|---|---|---|---|
| | **Ent. ↑** | **Contr. ↓** | **Comp. ↑** |
| Baseline | 0.34 | 0.10 | 0.25 |
| `GINGER` | **0.61***$^{*}$ | **0.06** | **0.29** |

pleted by 5 crowd workers. Crowd workers with a greater than 97% approval rate, over 10,000 approved tasks, and located in the US were qualified to participate in the study. Workers were paid US $0.25 for successful task completion.[8] Workers who failed to correctly classify 5 out of 8 aspects or more were rejected. The acceptance rate was 70% (67 out of 220 HITs have been rejected due to failed attentiveness checks).

## 6.5 Results on the TREC CAsT Datasets

We verify whether our modular response generation pipeline operating on information nuggets (1) ensures grounding of the response in specific facts from the retrieved sources, (2) generates relevant, answerable follow-up questions, and (3) provides control over response completeness. This analysis is based on the TREC CAsT collections, with additional results on TREC RAG discussed in Section 6.6. Additional results are presented in Appendix C.

### 6.5.1 Grounding and Source Attribution

The results of the automatic and human evaluation are presented in Tables 6.1 and 6.2. Grounding is assessed by computing entailment and contradiction of information nuggets detected in the source passages in the generated response (cf. Section 6.5.1), with high nugget entailment and low nugget contradiction scores being ideal. Completeness is measured as the fraction of ground-truth information nuggets entailed by the final response. Given that automatic metrics may not provide a reliable evaluation of LLM summaries (Goyal *et al.*, 2023), we also conduct a human evaluation to collect user preferences along the response dimensions discussed in Section 6.4.1.

**Does operating on information nuggets instead of passages in generating the response improve grounding and source attribution?** We observe significantly higher nugget entailment for responses generated with our

[8] The average time taken to complete a task was 1.7 minutes (8.82 USD hourly rate).

Table 6.2: Human evaluation of responses that reports the fraction of votes received when compared with the other method for (Coh)erence, (Con)ciseness, (Eng)agingness, (Fac)tuality, (Suf)ficiency, response (Pref)erence, and average scores for follow-up questions (on 3-point Likert scale) in terms of relevance (FQ_rel) and usefulness (FQ_use). Statistically significant differences ($p < 0.05$) with respect to the baseline are marked with * (Chi-square). The best scores for each measure are boldfaced.

| **Method** | **Human evaluation** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Coh** | **Con** | **Eng** | **Fac** | **Suf** | **Pref** | **FQ_rel** | **FQ_use** |
| Baseline | 0.45 | 0.40 | 0.50 | 0.48 | **0.52** | 0.49 | **2.58** | **2.63** |
| GINGER | **0.55** | **0.60*** | **0.50** | **0.52** | 0.48 | **0.51** | **2.58** | 2.60 |

method compared to the baseline that generates the response from entire passages in a single step. This may be due to baseline responses focusing on different information from the passages, not necessarily on the detected snippets. Nevertheless, given the high performance of automatic nugget detection,[9] we can conclude that our method ensures a higher entailment of relevant information in the generated response. In general, by operating on information nuggets, GINGER grounds responses in specific facts and significantly improves source attribution over the baseline.

**Does nugget-based response generation method improve user-perceived response quality?** In human evaluation, we observe significant preference towards GINGER only in terms of response coherence and conciseness. Human scores are comparable for the remaining response dimensions considered in the study. Crowd workers do not observe differences in the factuality of the responses, even though our method significantly outperforms the baseline in grounding and source attribution. This discrepancy may arise from the fact that crowd workers evaluate only the response and follow-up question without access to the sources (documents where the input passages originate from), leaving them unable to verify source attribution. Higher scores for coherence and conciseness may follow from highly structured and dense responses generated by GINGER, where each sentence discusses a specific aspect, compared to baseline-generated responses that may be less organized and contain redundant information. In general, human evaluation shows that the responses generated by baseline and GINGER are comparable with a clear preference towards our method in terms of coherence and conciseness.

[9] Evaluation of automatic nugget detection can be found in Appendix C.

Table 6.3: Response completeness scores (measured by entailed ground-truth information nuggets) and human evaluation of response breadth and depth (fraction of votes).

| Method | Completeness | Resp. breadth | Resp. depth |
|---|---|---|---|
| GINGER broad | **0.31** | **0.58** | **0.57** |
| GINGER deep | $0.17^*$ | $0.42^*$ | $0.43^*$ |

### 6.5.2 Follow-up Question Generation

Focusing on the human evaluation of follow-up questions, Table 6.2 presents the results in its last two columns. We find that follow-up questions generated by our method are on par with the facet-based approach used as our baseline (cf. Section 6.3.2). This demonstrates that GINGER is capable of generating useful follow-up questions based on facet clusters. Notably, those questions are guaranteed to be answerable, which plays a crucial role in dialogue continuity and reliability of the conversational system.

### 6.5.3 Controlling Completeness

To measure GINGER's ability to control the completeness of generated responses, we contrast two settings: *broad* and *deep*. The former uses the standard GINGER setup with top 3 cluster summaries (facet threshold = 3), while the latter is a 3-sentence summary of the top cluster (facet threshold = 1) (see Section 6.4.1). The results presented in Table 6.3 show that broad responses indeed achieve a higher completeness score, indicating that more nuggets are entailed in the response, while deep responses cover a lower number of ground-truth information nuggets, demonstrating that the completeness of the responses can effectively be controlled in GINGER.

We also perform a human evaluation to measure the perceived response breadth and depth. Surprisingly, crowd workers rated broad responses as having both greater diversity and more in-depth information. This unexpected result might be attributed to facet clusters containing a limited number of nuggets, which may hinder in-depth coverage of certain aspects of the topic. Furthermore, we observed no correlation between completeness and perceived breadth and depth;[10] this may be caused by the fact that crowd workers were not provided with the source material, potentially limiting their ability to fully evaluate information diversity and detail.

Table 6.4: Automatic evaluation of responses with varying quality of input. The columns are the same as in Table 6.1. Additionally, faithfulness and response relevance scores generated with the RAGAs framework are reported.

| **Input** | **Method** | **Automatic evaluation** | | | | |
|---|---|---|---|---|---|---|
| | | **Entail.** | **Contrad.** | **Compl.** | **Faithfulness** | **Answer rel.** |
| Relevant | Baseline | 0.34 | 0.10 | 0.25 | **0.79** ± 0.24 | **0.94** ± 0.04 |
| | GINGER | **0.61**$^*$ | **0.06** | **0.29** | 0.69 ± 0.30 | 0.87$^*$ ± 0.14 |
| Retrieved | Baseline | 0.38 | 0.12 | **0.17** | 0.71 ± 0.36 | **0.92** ± 0.15 |
| | GINGER | **0.61**$^*$ | **0.07** | 0.13 | **0.71** ± 0.28 | 0.88 ± 0.06 |
| Irrelevant | Baseline | 0.19 | 0.17 | **0.07** | **0.48** ± 0.36 | 0.73 ± 0.38 |
| | GINGER | **0.55**$^*$ | **0.04**$^*$ | 0.03 | 0.47 ± 0.29 | **0.75** ± 0.27 |

Table 6.5: Human evaluation of responses with varying quality of input. The columns are the same as in Table 6.2.

| **Input** | **Method** | **Human evaluation** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Coh** | **Con** | **Eng** | **Fac** | **Suf** | **Pref** | **FQ_rel** | **FQ_use** |
| Relevant | Baseline | 0.45 | 0.40 | 0.50 | 0.48 | **0.52** | 0.49 | 2.58 | **2.63** |
| | GINGER | **0.55** | **0.60**$^*$ | 0.50 | **0.52** | 0.48 | **0.51** | 2.58 | 2.60 |
| Retrieved | Baseline | **0.69** | **0.53** | **0.64** | **0.64** | **0.62** | **0.65** | **2.55** | **2.61** |
| | GINGER | 0.31$^*$ | 0.47 | 0.36$^*$ | 0.36$^*$ | 0.38$^*$ | 0.35$^*$ | 2.47 | 2.48 |
| Irrelevant | Baseline | **0.75** | **0.52** | **0.74** | **0.76** | **0.73** | **0.71** | **2.53** | **2.53** |
| | GINGER | 0.25* | 0.48 | 0.26* | 0.24* | 0.27* | 0.29* | 2.03* | 2.01* |

### 6.5.4 Robustness

So far we have considered an idealized setting where GINGER was provided with a set of *relevant* passages as input. To evaluate the robustness of our approach, we now consider two additional settings: *retrieved* represents a real-world scenario where a competitive passage ranker is employed, while *irrelevant* corresponds to the situation when the answer is not found in the corpus (see Section 6.3.1). In addition to nugget entailment and completeness scores, we report on faithfulness and answer relevance computed using the RAGAs framework. Since RAGAs scores are non-deterministic, we run the framework 5 times for each response variant and report on the average along with the standard deviation of the runs. Tables 6.4 and 6.5 present the results.

Intuitively, the quality of input passages and the amount of diverse information included in them should have a direct impact on the quality of the

[10] Additional analysis of the correlation between human and corresponding automatic measures can be found in Appendix C.

Table 6.6: Automatic evaluation responses for different variants of GINGER. Statistically significant differences (t-test, $p < 0.05$) with respect to the standard setup (top row) are marked with $^*$.

| Method | Entail. | Contrad. | Compl. | Faithfulness | Answer rel. |
|---|---|---|---|---|---|
| GINGER | 0.61 | 0.06 | 0.29 | 0.69 ± 0.30 | 0.87 ± 0.14 |
| -fluency | 0.70 | 0.07 | 0.28 | 0.72 ± 0.29 | 0.86 ± 0.14 |
| -fluency w/ GTnuggets | 0.51 | 0.06 | **0.48**$^*$ | **0.81**$^*$ ± 0.24 | 0.86 ± 0.14 |
| -fluency w/ BM25 | 0.70 | 0.06 | 0.30 | 0.72 ± 0.29 | 0.86 ± 0.11 |
| -fluency w/ LSA | 0.71 | 0.06 | 0.29 | 0.75 ± 0.26 | **0.88** ± 0.05 |
| -fluency w/ BM25+LSA | **0.74**$^*$ | **0.04** | 0.29 | 0.72 ± 0.27 | 0.87 ± 0.06 |

generated response. To some extent, we observe this tendency in RAGAs scores for different inputs for both methods. Responses generated from irrelevant passages have significantly lower scores for both faithfulness and answer relevance. This also aligns with previous research showing that using more relevant information monotonically improves open-domain QA results for generation (Lewis *et al.*, 2020). However, grounding evaluation performed with RAGAs does not confirm scores reported by entailment, with results showing no significant differences between baseline and GINGER, and no clear preference towards any of the two. This indicates that the evaluation using the RAGAs framework is not sensitive enough and may not be applicable in our case, especially after the recent critique of the automatic evaluation of generative models by means of other models (Sakai, 2023; Faggioli *et al.*, 2023; Gienapp *et al.*, 2024; Rackauckas *et al.*, 2024).

Human evaluation shows a clear preference for the baseline for the *retrieved* and *irrelevant* inputs, despite the low entailment and faithfulness scores observed in automatic evaluation. This suggests that crowd workers prioritize the fluency of the single-step LLM baseline approach over the factually more reliable but possibly less fluent responses by GINGER. Similarly, users' preference for baseline-generated follow-up questions is observed for responses based on irrelevant and retrieved passages. This is expected, as irrelevant passages likely contain a limited number of relevant information nuggets, resulting in fewer facet clusters and insufficient data for our method to generate useful follow-up questions. In contrast, the baseline extracts facets based on the topic and uses autoregressive text generation, which may produce facets not mentioned in the source passages. It may also result in follow-up questions that are not answerable. We notice that human evaluation disagrees with automatic measures, which is unsurprising, as humans struggle with identifying unreliable content in fluent responses (Clark *et al.*, 2021).[10]

### 6.5.5 Ablation Study

The pipeline architecture of GINGER allows for the evaluation of individual components and their impact on the final responses. Specifically, we consider vari-

Table 6.7: Response evaluation with AutoNuggetizer. TREC scores are provided for TREC RAG'24 AG submissions. The remaining scores are based on our reimplementation of the framework.

| **Method** | **V_strict** | | | | |
|---|---|---|---|---|---|
| | **TREC** | **GPT4o** | **Gemini** | **Claude** | **avg LLM** |
| baseline-top5 | 0.247 | 0.332 | 0.468 | 0.525 | 0.442 |
| baseline_CoT-top5 | — | 0.332 | 0.452 | 0.500 | 0.428 |
| Webis | 0.357 | — | — | — | — |
| TREMA | 0.261 | — | — | — | — |
| GINGER-top20 wo/ rewriting | **0.427** | **0.500** | **0.543** | **0.659** | **0.568** |
| GINGER-top10 wo/ rewriting | 0.369 | 0.423 | 0.502 | 0.582 | 0.502 |
| GINGER-top5 wo/ rewriting | 0.213 | 0.263 | 0.392 | 0.431 | 0.362 |
| GINGER-top5 | 0.211 | 0.279 | 0.400 | 0.451 | 0.377 |

ants without the final response fluency improvement step (-fluency), experiment with more traditional approaches for nugget clustering, based on latent semantic analysis (LSA), and for cluster ranking, based on BM25. We also analyze the impact of using ground-truth nuggets instead of automatic nugget detection (GTnuggets). Following our standard setup from before, we provide the method with 5 relevant passages as input. The results are presented in Table 6.6.

We find that entailment, completeness, faithfulness, and answer relevance are not strongly impacted by the modification of specific components. The most significant improvements are observed when ground-truth nuggets are used, suggesting that the quality of nuggets plays a crucial role. We thus conclude that the main contributing factor to GINGER's performance is related to operating on information nuggets, as opposed to the effectiveness of individual components. This highlights the importance of ensuring higher granularity and reducing redundancy of information used for response generation.

## 6.6 Results on the TREC RAG'24 Dataset

Given that GINGER essentially performs retrieval-augmented generation, we seek to measure its capability to generate high-quality responses on the TREC RAG'24 benchmark. Table 6.7 presents results for the two baseline systems (RQ1b), as well as for our nugget-based response generation pipeline (with or without response fluency improvement) using different numbers of retrieved passages. Responses from the top 5 passages are limited to 3 sentences, while those from the top 10 or 20 passages have a 400-word limit. TREC scores, corresponding to AutoNuggetizer scores provided by TREC organizers, are reported for system configurations where we have official evaluation results.

Table 6.8: Response evaluation with AutoNuggetizer on the TREC RAG'24 dataset of responses generated with different variants of GINGER without the fluency enhancement step and with the top 20 passages provided as input.

| Clustering | Ranking | V_strict (avg LLM) |
|---|---|---|
| BERTopic | DuoT5 | **0.568** |
| BERTopic | BM25 | 0.554 |
| LSA | DuoT5 | 0.521 |
| LSA | BM25 | 0.551 |

**Does GINGER improve response generation performance over the baselines?** The best-performing variant of our system outperforms both baseline approaches. Even prompting the model to divide the task into several steps using Chain-of-Thought and providing a ground-truth response as an example does not help in the response generation process. This implies that given the complexity of the queries and the amount of input context to be taken into account, the LLM needs further guidance to generate an accurate answer.

**How does GINGER perform in comparison to other systems submitted to TREC RAG'24?** Based on the initial results provided by TREC RAG organizers (Pradeep *et al.*, 2024), our best system (`GINGER-top20 wo/ rewriting`) is among the top performing AG submissions. Even though several other systems decompose response generation into a multi-step process, GINGER shows higher performance. For reference, we include two other competitive AG submissions in the results table: Webis (`Webis.webis-ag-run0-taskrag`) (Fröbe *et al.*, 2024) and TREMA (`TREMAUNH.Enhanced_Iterative_Fact_Refinement_and_Prioritization`) (Farzi and Dietz, 2024b).

**How does the amount of input information affect GINGER's performance?** By operating on information nuggets throughout the pipeline, GINGER grounds responses in specific facts and effectively synthesizes information from the provided passages. Responses generated from more passages are of higher quality, indicating that the additional context is indeed utilized in the system output (`GINGER-top10 wo/ rewriting` vs. `GINGER-top20 wo/ rewriting`). This suggests that splitting response generation into several independent steps and increasing the granularity of information mitigates the information loss to which LLMs are prone. Even with potential information redundancy, responses of the same length limit score higher with more input context, implying that they include more vital facts.

**How much do individual pipeline components contribute to overall system performance?** GINGER's modular architecture allows us to evaluate individual components and their impact on the final responses. We experiment with traditional Latent Semantic Analysis (LSA) for nugget clustering and BM25 for cluster ranking, as alternatives. Following our best-performing setup, we provide 20 relevant passages as input and limit responses to 400 words. Table 6.8 presents the results of this ablation study. We find that modifying specific components does not strongly impact response quality. Therefore, we conclude that operating on information nuggets, rather than the effectiveness of individual components, is the primary factor contributing to GINGER's performance. This highlights the importance of higher information granularity and reduced redundancy in response generation.

**Does LLM-based fluency enhancement reduce grounding?** Since the responses returned by our method are concatenations of independent facet cluster summaries, the final response may lack fluency and coherence. However, the difference between the responses generated with the fluency improvement and without this step (GINGER-top5 vs GINGER-top5 wo/ rewriting) is not significant. This indicates that LLMs can be used to refine response fluency without sacrificing quality or grounding.

## 6.7 Conclusions

To answer **RQ3.2** *(How to ensure the grounding of responses in the retrieved sources?)*, we have introduced GINGER, an approach for the generation of grounded, completeness-aware conversational responses. By utilizing information nuggets from top retrieved passages, it employs a multi-stage process (clustering, reranking, summarization, fluency enhancement) to generate concise, information-rich text, free of redundancy. Our approach offers several key advantages: maximizing information within response length limits, providing source attribution for verifiability, guiding users with relevant follow-up questions, and allowing control over response completeness. We have compared GINGER against baseline response generation approaches using both automatic and human evaluation. In answer to **RQ1b** *(What are strong baselines for response generation in CIS systems?)*, we have adopted a retrieval-augmented prompting strategy using an off-the-shelf LLM without additional training, along with a Chain-of-Thought baseline that includes a manually curated in-context example from TREC CAsT'22. Automatic evaluation results show that GINGER outperforms the baselines in terms of grounding and source entailment. Evaluation with AutoNuggetizer framework shows that GINGER achieves top performance on the Augmented Generation task at the TREC 2024 RAG track. The human evaluation shows a clear preference towards GINGER in terms of conciseness and confirms that our method generates useful follow-up questions.

The grounded and completeness-aware responses generated by GINGER provide a promising foundation for advancing transparent and explainable responses in CIS. Communicating the completeness of responses to users can aid in navigating the search space more effectively. Additionally, highlighting the information nuggets used in response generation within the sources can enhance transparency by improving source attribution (see Chapter 7).

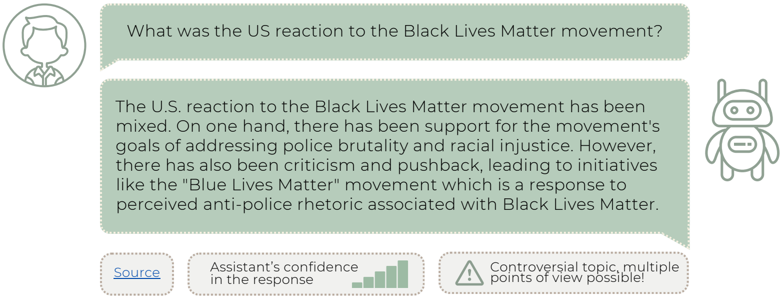
# Chapter 7

# Generating Transparent Responses

*Knowing is not enough; we must apply.*
*Willing is not enough; we must do.*

— **Johann Wolfgang von Goethe**

The increasing reliance on digital information has raised the demand for search systems that are not only factual and reliable but also transparent. Having established a method for synthesizing the requested information into a conversational response grounded in specific facts identified in the passages (see Chapter 6), we turn our attention to the remaining open challenge of ensuring response transparency. In transitioning from traditional search engine result pages to conversational information-seeking systems that limit responses to a few sentences, there is a significant concealment of underlying details such as the ranking of results and specifics about the sources. These details are essential for users to assess the scope, novelty, reliability, and topical relevance of the provided information (Xu and Chen, 2006). Recently proposed retrieval-augmented generation (RAG) systems (Lewis *et al.*, 2020) are claimed to produce more factually correct and diverse content. RAG, however, does not solve issues around transparency, as it is not able to indicate low-confidence responses or identify potential flaws related to limitations of the retrieved results or of the response generation process itself. Since the user is provided only with a short textual response as the final outcome of the generation process, it becomes the responsibility of the conversational system to identify and communicate any potential limitations to its users, ensuring transparency and empowering users to evaluate response quality. While the importance of explainability is broadly recognized



What was the US reaction to the Black Lives Matter movement?

The U.S. reaction to the Black Lives Matter movement has been mixed. On one hand, there has been support for the movement's goals of addressing police brutality and racial injustice. However, there has also been criticism and pushback, leading to initiatives like the "Blue Lives Matter" movement which is a response to perceived anti-police rhetoric associated with Black Lives Matter.

Source
Assistant's confidence in the response
Controversial topic, multiple points of view possible!

Figure 7.1: Information-seeking dialogue with a CIS system with explanations (sources, confidence, and limitations).

for AI (Monroe, 2018) and has been extensively studied, for example, for decision support and recommender systems (Nunes and Jannach, 2017; Zhang and Chen, 2020), it has not received due attention for CIS systems.

In this chapter, we aim to fill this gap by investigating approaches to explaining conversational responses, as a means to increase the transparency of the system addressing the following research question: **How to generate responses transparent about the system's confidence and limitations? (RQ3.3)**. Our focus is on *informational transparency*, disclosing information about the limitations or potential pitfalls in the response generation needed to enable appropriate understanding and assessment, in contrast to *functional* understanding of what the system can do, by exposing its capabilities and limitations or *mechanistic* understanding focused on how the system works (Liao and Vaughan, 2024). In particular, the focus of this study lies on the sources used for generating the response, the system's confidence in the provided information, and potential limitations or pitfalls of the response. In contrast to prior research on reporting system confidence (Cau *et al.*, 2023; Rechkemmer and Yin, 2022) or identifying particular limitations (Zhong *et al.*, 2020; Huang *et al.*, 2019b), our emphasis is on effectively communicating this information to the user in a conversational setting; we thus assume the existence of components that estimate system confidence and perform the detection of limitations.

Specifically, based on the previous research in related domains, we choose to increase the transparency of a CIS system by explaining (1) the origin of presented information, i.e., source (Tsai *et al.*, 2021; Bohnet *et al.*, 2023; Liu *et al.*, 2023a), (2) the system's confidence (Cau *et al.*, 2023; Radensky *et al.*, 2023), and (3) potential limitations of the generated response (Sakaeda and Kawahara, 2022); see Figure 7.1 for an illustration of an enhanced conversational response. Being transparent about these aspects of the response can enable users to make informed judgments about the presented information and increase their perceived usefulness of the response, bridging the gap between system-generated responses and responses verifiable by the user. We investigate users' perception of the system response quality together with the type and quality of explanations. Specifically, we ask the following two questions.

**How does the quality of responses and explanations affect user-perceived response usefulness? (RQ3.3a)** Individuals without specific training can only distinguish between human-generated and auto-generated texts at a level close to random chance (Clark *et al.*, 2021). Indeed, users easily overlook factually incorrect, unsupported, biased, or incomplete information. Therefore, we investigate the impact of the quality of the response and explanations provided by the system on users' assessment of the response. Specifically, imperfect responses in our study include factual errors or lack of viewpoint diversification, while noisy explanations introduce problems related to sources (information subjectivity or lack of support for the response), confidence (incorrect scores), or limitations (irrelevant information about pitfalls).

**What are effective ways to provide explanations to users? (RQ3.3b)** There are multiple approaches to providing users with explanations. One option is to incorporate them as part of the natural language system utterance, ensuring that users are explicitly informed about the confidence and potential pitfalls of the response (Rechkemmer and Yin, 2022). As an alternative, we explore utilizing various user interface elements to effectively convey the response's limitations (Lu and Yin, 2021; Shani *et al.*, 2013) or providing a granular scale of the system's confidence in generated response (Shani *et al.*, 2013). Building on findings from studies in recommender systems and automated decision making (Nunes and Jannach, 2017; Zhang and Chen, 2020), we seek to adapt and explore these concepts within the context of CIS systems.

To answer the above questions, we conduct a crowdsourcing-based user study with 160 participants asking about their perception of responses and explanations that vary in quality and presentation mode. Overall, our study seeks to establish a more trustworthy interaction in CIS dialogues by bridging the gap between system-generated responses and their usefulness to the users, by providing explanations.

This chapter is accompanied by an online repository, containing the manually generated CIS responses and explanations, as well as scripts for data analysis at `https://github.com/iai-group/sigir2024-transparentCIS`.

This chapter is based on the following paper:

> Łajewska *et al.* (2024b): *Explainability for Transparent Conversational Information-Seeking*, SIGIR '24

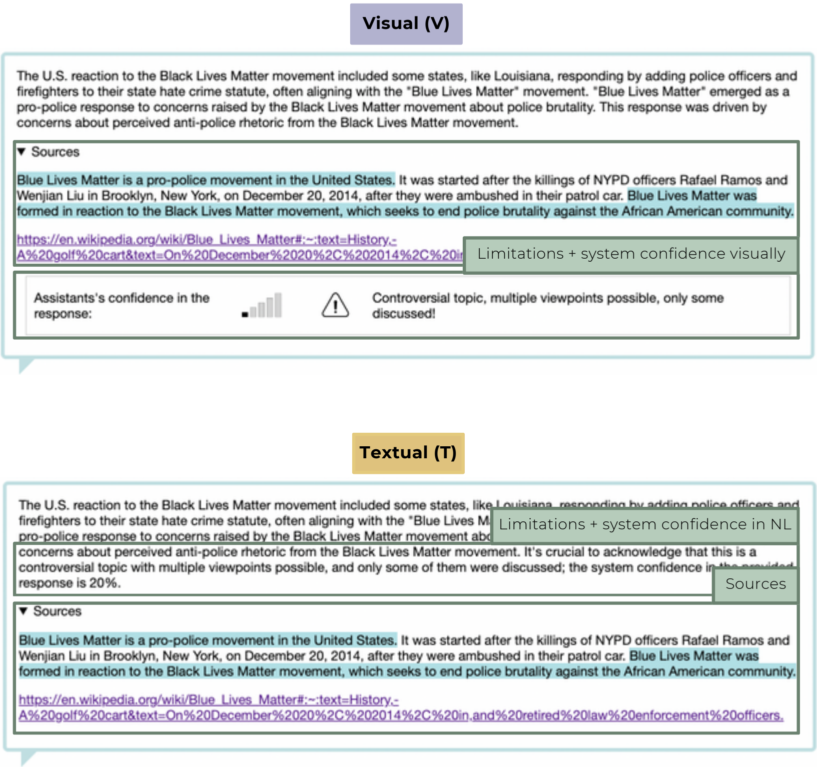
## 7.1 Related Work

Effective user-system interactions, aligning user expectations, and building trust in the systems that we attempt to achieve by communicating explanations about the response to the user are the main axes of explainable AI (XAI). According to human-AI interaction design guidelines, the system should communicate its capabilities, reliability, and the rationale behind its decisions (Amershi *et al.*, 2019). XAI research in the context of decision-making emphasizes the role of explanations in improving user comprehension (Cheng *et al.*, 2019) and increasing human trust in the system (Zhang *et al.*, 2020c). Explanations can vary in terms of presentation format (textual vs. visual) (Zhang *et al.*, 2020c), the level of interactivity (Cheng *et al.*, 2019), complexity (Tsai *et al.*, 2021), or reasoning styles (Cau *et al.*, 2023). Explanations can also be used to reveal the system's confidence (Cau *et al.*, 2023), system accuracy indicators (Kocielnik *et al.*, 2019), data sources (Tsai *et al.*, 2021), answer attribution (Bohnet *et al.*, 2023; Liu *et al.*, 2023a) and the correctness of system suggestions (Cau *et al.*, 2023) (see Section 2.3).

Despite advancements, CIS systems still face limitations such as unanswerability (Sulem *et al.*, 2022; Choi *et al.*, 2018; Rajpurkar *et al.*, 2018; Reddy *et al.*, 2019), biases or lack of viewpoint diversification (Gao and Shah, 2020; Draws *et al.*, 2021b; Sakaeda and Kawahara, 2022; Azzopardi, 2021) (see Section 2.2.4). Even though research has been done in related fields, such as text classification (Zhong *et al.*, 2020; Kim and Allan, 2019), question answering (Liao *et al.*, 2022; Rajpurkar *et al.*, 2018), or reading comprehension (Zhang *et al.*, 2021; Huang *et al.*, 2019b) towards detecting these issues, communicating detected problems to users is still a largely unexplored area in CIS. To ensure the transparency of the system, the response should disclose system capabilities and potential limitations, thereby managing user expectations (Radlinski and Craswell, 2017; Azzopardi *et al.*, 2018). Unlike previous studies that concentrated on detecting limitations, our work emphasizes the effective communication of potential flaws in the system's output to the user. Such limitations can be revealed using natural language utterances (Rechkemmer and Yin, 2022), using analogy (He *et al.*, 2023a), incorporating user interface elements (Lu and Yin, 2021; Koch *et al.*, 2023), or by providing a granular scale of the system's confidence (Shani *et al.*, 2013) (see Section 2.3.1).

## 7.2 Methodology

We aim to investigate the user's perception of the (1) system response quality, (2) type and quality of explanations, and (3) presentation of explanations. We assume a *CIS system* that, given a query, performs the following steps: (1) it retrieves passages and identifies the information nuggets in the top retrieved results containing key pieces of information answering a user query; (2) it synthesizes the identified snippets (i.e., information nuggets) into a concise and natural language response; (3) it returns the system's confidence in the


Visual (V)
Limitations + system confidence visually
Textual (T)
Limitations + system confidence in NL
Sources

The U.S. reaction to the Black Lives Matter movement included some states, like Louisiana, responding by adding police officers and firefighters to their state hate crime statute, often aligning with the "Blue Lives Matter" movement. "Blue Lives Matter" emerged as a pro-police response to concerns raised by the Black Lives Matter movement about police brutality. This response was driven by concerns about perceived anti-police rhetoric from the Black Lives Matter movement.

▼ Sources

Blue Lives Matter is a pro-police movement in the United States. It was started after the killings of NYPD officers Rafael Ramos and Wenjian Liu in Brooklyn, New York, on December 20, 2014, after they were ambushed in their patrol car. Blue Lives Matter was formed in reaction to the Black Lives Matter movement, which seeks to end police brutality against the African American community.

https://en.wikipedia.org/wiki/Blue_Lives_Matter#:~:text=History,-A%20golf%20cart&text=On%20December%2020%2C%202014%2C%20i

Assistants's confidence in the response: ▂▃▅▆▇ ⚠ Controversial topic, multiple viewpoints possible, only some discussed!

The U.S. reaction to the Black Lives Matter movement included some states, like Louisiana, responding by adding police officers and firefighters to their state hate crime statute, often aligning with the "Blue Lives M pro-police response to concerns raised by the Black Lives Matter movement abo concerns about perceived anti-police rhetoric from the Black Lives Matter movement. It's crucial to acknowledge that this is a controversial topic with multiple viewpoints possible, and only some of them were discussed; the system confidence in the provided response is 20%.

▼ Sources

Blue Lives Matter is a pro-police movement in the United States. It was started after the killings of NYPD officers Rafael Ramos and Wenjian Liu in Brooklyn, New York, on December 20, 2014, after they were ambushed in their patrol car. Blue Lives Matter was formed in reaction to the Black Lives Matter movement, which seeks to end police brutality against the African American community.

https://en.wikipedia.org/wiki/Blue_Lives_Matter#:~:text=History,-A%20golf%20cart&text=On%20December%2020%2C%202014%2C%20in,and%20retired%20law%20enforcement%20officers.

Figure 7.2: Examples of responses with explanations for the query: *What was the US reaction to the Black Lives Matter movement?* The response at the top contains limitations and system confidence presented using Visual (V) elements. The variant presented at the bottom contains this information appended at the end of the response in a Textual (T) form. The source is always presented in the same way.

provided response; and (4) based on the provided query, retrieved information nuggets, and returned confidence, it identifies the potential pitfalls and limitations that could have contributed to flaws in the response. We consider three types of explanations the system may provide: (1) the underlying *source*, to help users verify the response's factual correctness and broader context; (2) the system's *confidence* in the provided response, to give users insights about how certain the outcome of response generation is; and (3) potential *limitations or pitfalls* to warn the user about flaws in the response or the source.

The study's main goal is to investigate whether explanations provided by the system can make the user's response assessments easier or increase the information's usefulness. We provide crowd workers with different configurations of responses and explanations, varying in quality and presentation mode, and ask them to indicate their perception of different system response dimensions.

Inspired by work in the area of explainable decision-making systems (Cau *et al.*, 2023), we explore two different ways of presenting the explanations about the response limitations and system confidence: textual and visual presentation; see Figure 7.2.

### 7.2.1 Experimental Design

We have defined ten experimental conditions using different variants of the response and explanations.[1] Covering all combinations of factors (explanation components × quality × presentation mode) exhaustively would be unfeasible. Therefore, we select a subset of experimental conditions that best represent what we are trying to measure in our study. The selected conditions vary along three main dimensions: (1) response quality, (2) quality of the explanations (i.e., source, system confidence, limitations), and (3) presentation style (see Table 7.3). More details about experimental conditions and the different explanation variants can be found in Section 7.3.1.

The ten experimental conditions resulted in ten different human intelligence tasks (HITs). In each HIT, crowd workers are asked to assess responses for ten queries. This is to ensure that the obtained results are to a large extent topic-independent. To avoid repeated judgments that would reduce the reliability of the study, we allow each crowd worker to complete only one HIT (Steen and Markert, 2021), i.e., we employ a between-subject design (Kelly, 2007). In each HIT, the order of query-response pairs is intentionally randomized. This is done to prevent any adverse effects on the given query-response pairs that might occur if they were consistently presented towards the end of the task, where worker fatigue could potentially influence the results.

### 7.2.2 Crowdsourcing Task Design

Figure 7.3 summarizes the design of the crowdsourcing tasks. Each HIT contains ten query-response pairs and is comprised of: I) HIT instructions providing task background; II) a questionnaire about the worker's familiarity with conversational assistants (see Table 7.1); III) a description of the system; IV) ten CIS interactions; V) a post-task questionnaire; and VI) a demographics questionnaire. Workers are not given specific examples of query-response pairs in the instructions to avoid bias. Part III contains a pre-use explanation of the system (Chiang and Yin, 2022). It aims at improving the following competencies of the users: (1) understanding the capabilities of the system, and (2) understanding that the response is limited to 3 sentences only. We decompose part IV of the user study into ten subsections using independent CIS interactions to facilitate atomic microtask crowdsourcing (Gadiraju *et al.*, 2015). Each

[1] We acknowledge that the variants for each transparency dimension are not exhaustive. Various UI elements can be used to present information, and different ways to introduce noise can be explored. However, since response-related explanations have not been explored in conversational search, we limit the first study in this area to solutions previously proposed for similar systems.

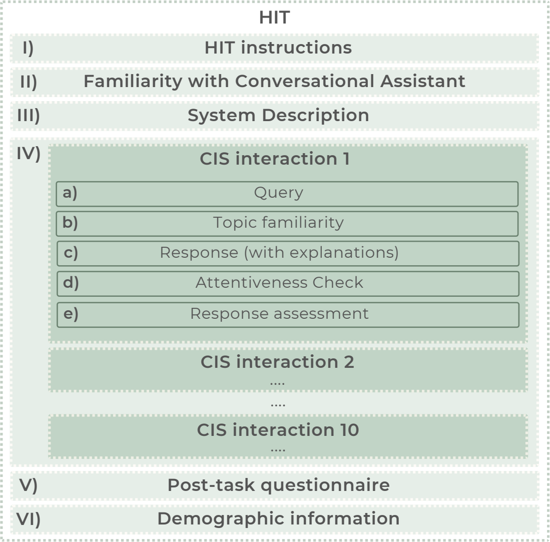


Figure 7.3: High-level design of the user study.

CIS interaction contains (a) a query; (b) a topic familiarity questionnaire; (c) a system response possibly enhanced with explanations; (d) a corresponding attentiveness check; and (e) a CIS response assessment. CIS interactions are followed by a post-task questionnaire (Part V) investigating workers' experience of interacting with the assistant in general, not concerning specific responses. The questionnaire contains indirect questions about all three types of explanations enhancing the system response (see Table 7.1). The HIT finishes with a short demographics questionnaire (Part VI) asking workers' age, education level, and gender.

**a) Query and b) Topic Familiarity** The query is followed by a short questionnaire asking about interest, familiarity, and likelihood of posing a similar query (Bolotova *et al.*, 2020) (see Table 7.1). In this user study, the worker's background knowledge and familiarity with the topic are dependent variables that we cannot control. Asking users to assess their familiarity with the topic enables us to condition the collected data on users' background knowledge (Krishna *et al.*, 2021).

**c) Response** The system response synthesizes the information nuggets identified in the top retrieved results. The response can be enhanced with explanations that can be presented in different formats.

**d) Attentiveness Check** We present workers with an attentiveness check for each query-response pair, to detect poorly performing workers, cheat submissions, or bots (Gadiraju *et al.*, 2015). Each attention check consists of three

Table 7.1: Questions used for collecting data about the user experience of using conversational agents, their involvement in the topic, and their rating for explanations.

| **Variable** | **Question used in the user study** |
|---|---|
| Conversational Agent Familiarity | How often do you use conversational assistants like Siri, Alexa, or Google Assistant? |
| Search with Agent Frequency | How often do you use conversational assistants to search for information? |
| Topic Familiarity | What is your level of familiarity with the topic of the question? |
| Interest in Topic | What is your level of interest in the question? |
| Similar Search Probability | What is the likelihood that you would search for this information? |
| Source Explanation | To what extent were the provided responses supported? |
| Limitation Explanation | To what extent did the assistant help you realize the potential limitations of the responses? |
| Confidence Explanation | To what extent are you aware of the assistant's confidence in the provided responses? |

sentences related to the topic of the query, one of them being a summary of the provided response. Sentences are provided in a random order and workers are asked to select the best summary (Bolotova-Baranova *et al.*, 2023). This simple quality check enables us to filter out responses from workers who are not performing the task attentively or reading the responses carefully. Submissions that failed on more than 3 out of 10 attentiveness questions were rejected.

**e) Response Assessment** In this part of the CIS interaction, workers are asked to evaluate different dimensions of the response variant presented for a given query. The question about each response dimension is answered on a four-point Likert scale. Explicitly asking users to report on its value is not helpful because they may have a different understanding of this concept (Kelly, 2007). Therefore, in our setup, user satisfaction is indirectly observable. To increase the ecological validity of our experiments, the questions do not use explicitly the names of the dimensions. Instead, we ask about each response dimension using an operational definition (see Table 7.2). This approach ensures a common understanding of the dimensions by all study participants. Both the response dimensions and the operational definitions are inspired by Cambazoglu *et al.* (2021)'s work investigating answer utility for non-factoid question answering.

Table 7.2: Operational definitions used in the response assessment questionnaire for all response dimensions. They followed a statement: *The provided assistant's response ...* and were answered by crowd workers on a four-point Likert scale.

| Response Dimension | Operational definition used in the user study |
|---|---|
| Usefulness | ...was useful for completing my task |
| Relevance | ...is about the subject of the question |
| Correctness | ...contains an accurate response to the question |
| Completeness | ...covers every aspect of the question |
| Comprehensiveness | ...contains detailed information |
| Conciseness | ...does not contain redundant information |
| Serendipity | ...contains some unexpected but positively surprising information |
| Coherence | ...does not contain inconsistent statement |
| Factuality | ...is based on things that are known to be true |
| Fairness | ...is free of any kind of bias |
| Readability | ...is fluently written |
| Satisfaction | ...is satisfying in terms of completing my information need |

## 7.3 User Study Execution

We used the Amazon Mechanical Turk (MTurk) crowdsourcing platform to collect responses from online workers.[2] Data collection was run between 20 December 2023 and 9 January 2024, divided into two stages: a pilot (Section 7.3.2) and a main study (Section 7.3.3).

### 7.3.1 Data

A critical element of the study is selecting query-response pairs and explanations enhancing the responses that enable us to answer our research questions. We use ten queries selected from the TREC CAsT'20 (Dalton *et al.*, 2020) and '22 (Owoicho *et al.*, 2022) datasets and two manually created responses for each query. Different variants of the responses (perfect and imperfect) and explanations (accurate and noisy) are created manually by the author of this thesis. The noise in responses and explanations is introduced manually using framing, i.e., distorting the information presented to the users (Kocielnik *et al.*, 2019). For each source, the specific information nuggets that contributed to the answer are highlighted, inspired by the CAsT-snippets dataset (see Section 5.4 for more details about the dataset).

**Queries and Responses** The query selection process takes into account the potential challenges of the query and the familiarity of crowd workers with the topic. We select a subset of 20 queries from the TREC CAsT datasets that are

[2] Our institution does not require ethics approval for this kind of study.

challenging in one of two aspects: (1) limited coverage of the topic in the corpus or lack of a full answer, resulting in factual errors; or (2) topic complexity or controversy resulting in an incomplete or biased response. By selecting these challenging queries we attempt to simulate scenarios where enhancing the system response with explanation can be beneficial for users. Additionally, queries selected in the first step are sorted according to the familiarity scores reported by crowd workers in a small crowdsourcing study that was set up to select the top ten queries that are deemed most well-known to users. This approach aims to ensure that users possess sufficient background to meaningfully assess responses and associated explanations. We consider two variants of the response for each query: perfect and imperfect. The perfect response, i.e., ground truth answer, is generated manually using the top retrieved results by the author of this thesis. The imperfect response is a manual modification of the ground truth answer to contain factual errors, be biased towards one point of view, or cover only one aspect of a complex problem. This way, we attempt to take into account significantly different versions of the responses in terms of their accuracy and quality.

**Explanations** We provide explanations related to (1) source, (2) system confidence, and (3) limitations, which are instantiated in two variants: accurate and noisy.

**(1) Sources** The "Source" component is an expandable element within the response, encompassing the complete text of the paragraph used for generating the response. It includes annotations of information nuggets, highlighting crucial pieces of information within the passage. Additionally, workers receive a link to the entire webpage from which the passage originates (Liu *et al.*, 2023a). This allows them to access the full text of the document, aiding in the assessment of its relevance, which is particularly beneficial for long, non-navigational queries (Kazai *et al.*, 2022). The URLs are anchored to the specific section of the webpage where the passage is located. Additionally, based on the URL, workers can assess the credibility or authority of the source. The noisy source pertains to the query's topic but lacks information that supports the provided response (Liu *et al.*, 2023a). It corresponds to the initial passage from the Wikipedia page related to the general query topic, allowing for an assessment of users' diligence in verifying the provided explanations.

**(2) System Confidence** Within conversational response generation, confidence can be assessed along several different dimensions:

- The confidence that the identified snippets contain the full, complete answer to the question, not only part of it.

- Given that the response is limited to only 3 sentences, the confidence that the top-$k$ snippets used in the response provide sufficient coverage of the retrieved information.

- The confidence that the response generated with LLM using the selected snippets is accurate; this accuracy is tied to the model's fluency in the topic, assessing how adept the model is in crafting content on a given subject, which can be influenced by the volume of data during LLM training or topic popularity.

System confidence is either communicated in textual form (*"the system confidence in the provided response is ... %"* appended at the end of the response) (Cau *et al.*, 2023) or through an additional UI element. Given that users best understand confidence displays inspired by well-known displays in other areas (Shani *et al.*, 2013), we decided to use a bar chart presentation that is often associated with cell phone connectivity. Adding noise to this component results in system confidence being reverted, i.e., although the provided response is correct, low system confidence is reported. We consider the confidence of 1–2 out of 5 for imperfect responses and 4–5 for perfect responses. We skip confidence of 3 as it is ambiguous and we skip confidence of 0 as it represents the situation when the system should not show any response, but state that an answer could not be found.[3]

**(3) Response Limitations** We have identified several key areas of potential challenges and problems that could impact the usefulness of provided responses. These issues, while not exhaustive, serve as a starting point for consideration in our user study. Among the challenges related to the topic, we recognize the potential issues related to controversy, leading to a lack of viewpoint diversification, and complexity, resulting in response incompleteness. Source-related challenges include the subjectivity of the source text used, possibly outdated source information, sources influenced by commercial interest, promoting specific products, or brands, and reliance on unverified or not reputable sources. Query-related issues encompass biases or ambiguities in queries, time-sensitive queries requiring current information, queries lacking sufficient context, privacy-sensitive queries involving private or confidential information, and speculative queries seeking insights into future events. Additionally, search and system issues may arise, such as rare topics insufficiently covered in the corpus, lack of credible sources supporting the response, or retrieved passages containing contradictory information.

The query-response pairs selected for this study contain factual errors, are incomplete, or rely on subjective sources. Additionally, challenges related to the topic, source, or query, not identified in the subset of query-passage pairs used in this study, are also considered to explore whether users can more easily identify these issues based on the presence and correctness of explanations provided by the system. Issues purely related to search or system failures, where the system is aware of its inability to find sources that answer the question, fall outside

[3] Users' reactions to such extreme confidence scores are not a subject of this study but could be explored in future work once it has been established that users find confidence explanations useful.

Table 7.3: Experimental conditions considered in the user study; components may be included without noise (+); with some inaccuracies (˜); or not provided in the system's output (–). (T) indicates textual and (V) visual presentation mode.

| | **EC1** | **EC2** | **EC3** | **EC4** | **EC5** | **EC6** | **EC7** | **EC8** | **EC9** | **EC10** |
|---|---|---|---|---|---|---|---|---|---|---|
| Response | +, T | +, T | ˜, T | ˜, T | +, T | +, T | ˜, T | ˜, T | +, T | ˜, T |
| Source | +, T | +, T | +, T | +, T | ˜, T | ˜, T | ˜, T | ˜, T | – | – |
| Confidence | +, V | +, T | +, V | +, T | ˜, V | ˜, T | ˜, V | ˜, T | – | – |
| Limitations | +, V | +, T | +, V | +, T | ˜, V | ˜, T | ˜, V | ˜, T | – | – |

the scope of this study. In such cases, the system should inform the user about no answer found without trying to produce a response. Response limitations are communicated either in a textual form by appending running text at the end of the system response (Rechkemmer and Yin, 2022; Costa *et al.*, 2018) or using an additional UI element resembling a warning message (inspired by fact-checking warning labels (Koch *et al.*, 2023)). Adding noise to limitation explanation results in communicating irrelevant limitations, i.e., if the topic is controversial, the system informs the user about query ambiguity or possibly outdated source information. We aim for the noisy limitations to be easily distinguishable after reading the query and the response carefully. The goal of this study is to investigate the limitation explanations rather than the detection of specific limitations, therefore the noise in the limitations is aimed to be easy to spot.

**Experimental Conditions (EC)** The subset of experimental conditions selected for this user study is summarized in Table 7.3. The conditions vary along three main dimensions: (1) response quality, (2) explanation quality, and (3) presentation of explanation. EC1 and EC2 represent a perfect system response with accurate explanations. More specifically, the explanations cover the source supporting the response, as well as the system's confidence score; the limitation explanation is not included because the response has no inaccuracies in this case. EC3 and EC4 correspond to imperfect responses that may contain some factual errors or be biased towards one specific point of view but are accompanied by accurate explanations related to source, confidence, and limitations. EC5 and EC6 represent the perfect response accompanied by noisy explanations. EC7 and EC8 correspond to imperfect responses that contain flaws and noisy explanations. The last group of conditions, EC9 and EC10, represents the response (either perfect or imperfect) without explanations.

### 7.3.2 Pilot Study

We ran a pilot study (MTurk; $N$=15; 3 HITs; US$3 per HIT, proportional to US minimum wage), where HITs corresponded to three experimental conditions selected from the 10 described in Table 7.3: EC3, EC4, and EC7. The selected conditions encompass border cases, featuring variations in both the presentation mode of explanations (EC3 vs. EC4) and the quality explanations (EC3 vs. EC7), and deliberately involve imperfect responses to simulate the most natural scenarios. In their overall feedback, crowd workers primarily expressed concerns about the length of the task and the payment which was accordingly increased in the large-scale data collection.

We performed a power analysis by employing one-way ANOVA with the experimental condition as an independent variable and user-reported response usefulness as a dependent variable (Sakai, 2018). The results indicate that 16 workers are required to observe a statistically significant effect of explanation quality on the perceived usefulness of system responses, whereas 56 workers are required for a statistically significant effect of the explanation presentation mode. Considering four additional pairs of experimental conditions with varying presentation modes, we expect that gathering data from 14 unique workers per HIT (56 from power analysis divided by 4 pairs of conditions) is adequate to observe a statistically significant effect of presentation mode across all ten experimental conditions. Based on this analysis, we decided to recruit 16 unique workers per HIT in our main study.

### 7.3.3 Main Study

Crowd workers with a greater than 97% approval rate, over 5,000 approved HITs, and located in the US were qualified to participate in the study. Workers were paid US$4 for successful HIT completion. Workers who failed 4 out of 10 attentiveness checks or more were rejected. Altogether we collected 273 submissions, out of which 113 were discarded due to failed attentiveness checks. Accepted tasks were submitted by 160 unique workers (16 per HIT), with the following user-reported demographics: 95 male, 60 female, 5 in "other" category (none in "prefer not to say"); age breakdown: 18–30 (39), 31–45 (76), 46–60 (41), 60+ (4); highest degree: Ph.D. or higher (3), master's (34), bachelor's (111), high school (12).

## 7.4 Results

To answer our research questions, we first analyze the sensitivity of our experiment. Tables 7.4, 7.5 and 7.6 show the results of the one- and two-way ANOVA tests for statistical significance on user-reported dimensions, using a significance level of $\alpha = 0.05$. Whenever applicable, the effect size of a given factor is classified based on the formula for the unbiased estimator and scales used by Culpepper *et al.* (2022). Given the large number of factors defining

Table 7.4: Results of one-way ANOVA for response dimensions: (Use)fulness, (Rel)evance, (Cor)rectness, (Compl)eteness, (Compr)ehensiveness, (Con)ciseness, (Ser)endipity, (Coh)erence, (Fac)tuality, (Fair)ness, (Read)ability, and (Sat)isfaction. Self-reported response dimensions (dependent variables) are in columns, and independent variables are in rows. Boldface indicates statistically significant effects ($p < 0.05$). Effect size: L=Large, M=Medium, S=Small.

|  | **Use.** | **Other Dimensions** |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | **Rel.** | **Cor.** | **Compl.** | **Compr.** | **Con.** | **Ser.** | **Coh.** | **Fac.** | **Fair.** | **Read.** | **Sat.** |
|  | *All conditions (EC1–EC10)* |  |  |  |  |  |  |  |  |  |  |  |
| Response Quality | 0.156 (S) | 0.176 (S) | **0.003 (S)** | 0.745 (–) | 0.846 (–) | 0.374 (S) | 0.093 (S) | 0.217 (S) | 0.265 (S) | 0.924 (–) | 0.881 (–) | 0.638 (S) |
| Explanation Quality | **0.0 (S)** | **0.0 (S)** | 0.508 (S) | **0.003 (S)** | **0.0 (S)** | **0.001 (S)** | 0.09 (S) | **0.002 (S)** | 0.713 (–) | **0.0 (S)** | **0.032 (S)** | **0.0 (S)** |
| Presentation Mode | **0.019 (S)** | **0.0 (S)** | 0.234 (S) | 0.347 (S) | 0.658 (–) | **0.001 (S)** | 0.149 (S) | 0.09 (S) | 0.842 (–) | **0.001 (S)** | 0.651 (–) | **0.0 (S)** |
| Query | 0.341 (S) | 0.911 (–) | 0.939 (–) | 0.84 (–) | 0.733 (–) | 0.449 (S) | 0.66 (–) | 0.543 (–) | 0.724 (–) | 0.098 (S) | 0.125 (S) | 0.254 (S) |
| Topic Familiarity | **0.017 (S)** | **0.0 (S)** | 0.285 (S) | **0.0 (S)** | **0.0 (S)** | **0.0 (S)** | **0.0 (M)** | **0.0 (S)** | **0.0 (S)** | **0.0 (S)** | **0.002 (S)** | **0.0 (S)** |
| Interest In Topic | **0.0 (S)** | **0.007 (S)** | **0.0 (S)** | **0.0 (S)** | **0.0 (S)** | 0.053 (S) | **0.0 (M)** | 0.115 (S) | **0.0 (S)** | **0.0 (S)** | **0.0 (S)** | **0.0 (S)** |
| Similar Search Prob. | **0.0 (S)** | **0.0 (S)** | **0.001 (S)** | **0.0 (M)** | **0.0 (S)** | **0.0 (S)** | **0.0 (M)** | **0.002 (S)** | **0.0 (S)** | **0.0 (S)** | **0.0 (S)** | **0.0 (S)** |
| Conv. Agent Fam. | 0.079 (S) | **0.0 (S)** | 0.077 (S) | **0.001 (S)** | **0.0 (S)** | 0.093 (S) | **0.0 (S)** | **0.003 (S)** | **0.0 (S)** | 0.079 (S) | **0.005 (S)** | **0.004 (S)** |
| Search w/ Agent Freq. | **0.0 (S)** | **0.002 (S)** | 0.351 (S) | **0.0 (S)** | **0.0 (S)** | **0.0 (S)** | **0.0 (M)** | **0.0 (S)** | 0.533 (–) | 0.426 (S) | **0.0 (S)** | **0.0 (S)** |
|  | *Only conditions with explanations (EC1–EC8)* |  |  |  |  |  |  |  |  |  |  |  |
| Explanation Quality | **0.0 (S)** | **0.006 (S)** | 0.256 (S) | **0.002 (S)** | **0.0 (S)** | 0.122 (S) | 0.319 (S) | **0.003 (S)** | 0.504 (S) | **0.0 (S)** | **0.014 (S)** | **0.007 (S)** |
| Presentation Mode | 0.872 (–) | 0.686 (–) | 0.096 (S) | 0.895 (–) | 0.38 (S) | 0.399 (S) | 0.86 (–) | 0.377 (S) | 0.739 (–) | 0.78 (–) | 0.771 (–) | 0.071 (S) |

Table 7.5: Results of one-way ANOVA for explanations. Self-reported dimensions (dependent variables) are in columns, and independent variables are in rows. Boldface indicates statistically significant effects ($p < 0.05$). Effect size: L=Large, M=Medium, S=Small.

| | Explanation | | |
|---|---|---|---|
| | **Source** | **Confidence** | **Limititations** |
| | *All conditions (EC1–EC10)* | | |
| Response Quality | 0.697 (–) | 0.456 (S) | 0.445 (S) |
| Explanation Quality | **0.0 (S)** | **0.0 (S)** | 0.173 (S) |
| Presentation Mode | **0.0 (S)** | **0.0 (S)** | **0.0 (S)** |
| Query | 1.0 (–) | 1.0 (–) | 1.0 (–) |
| Topic Familiarity | **0.0 (M)** | **0.0 (S)** | **0.0 (S)** |
| Interest In Topic | **0.0 (M)** | **0.0 (S)** | **0.0 (S)** |
| Similar Search Prob. | **0.0 (M)** | **0.0 (S)** | **0.0 (S)** |
| Conv. Agent Familiarity | **0.0 (S)** | **0.0 (S)** | **0.0 (S)** |
| Search with Agent Freq. | **0.0 (M)** | **0.0 (S)** | **0.0 (M)** |
| | *Only conditions with explanations (EC1–EC8)* | | |
| Explanation Quality | 0.097 (S) | **0.0 (S)** | 0.088 (S) |
| Presentation Mode | **0.0 (S)** | 0.653 (–) | **0.0 (S)** |

each experimental condition, we treat response quality, quality of explanations, and their presentation mode as three separate independent variables to simplify the interpretation of the results. Each user-reported response dimension score and user rating for explanation is treated as a dependent variable. The analysis performed to answer RQ3.3a (Section 7.4.2) and RQ3.3b (Section 7.4.3) is based only on the results with the statistically significant effects discussed in Section 7.4.1.

### 7.4.1 User's Perception of Response and Explanations

**Response Quality.** Table 7.4 shows that *response quality* has a statistically significant effect only on user-reported correctness of the response. Completeness, factuality, and fairness are not influenced by the quality of the response, even though some responses contained manually injected errors related to these dimensions (e.g., bias towards one specific point of view, factual errors, or covering only one aspect of the topic). This insensitivity of user-reported response dimensions to the quality of provided information may suggest that users are not able to identify some of the problems with the response without expert knowledge about the topic.

Table 7.6: Results of two-way ANOVA. The boldface indicates statistically significant effects ($p < 0.05$). Effect size: S=Small.

| | Usefulness | Satisfaction | Explanation | | |
|---|---|---|---|---|---|
| | | | Source | Confidence | Limitations |
| *Interactions with Query* | | | | | |
| Response Quality | 0.069 (S) | 0.296 (S) | 1.0 (–) | 1.0 (–) | 1.0 (–) |
| Explanation Quality | 0.767 (–) | 0.993 (–) | 1.0 (–) | 1.0 (–) | 1.0 (–) |
| Presentation Mode | 0.94 (–) | 0.981 (–) | 1.0 (–) | 1.0 (–) | 1.0 (–) |
| Conv. Agent Fam. | 0.995 (–) | 0.887 (–) | 1.0 (–) | 1.0 (–) | 1.0 (–) |
| Search w/ Agent Freq. | 0.632 (–) | 0.215 (S) | 1.0 (–) | 1.0 (–) | 1.0 (–) |
| Topic Familiarity | 0.697 (–) | 0.489 (S) | **0.002 (S)** | 0.71 (–) | **0.001 (S)** |
| Interest in Topic | 0.087 (S) | 0.542 (–) | 0.063 (S) | 0.698 (–) | 0.234 (S) |
| Similar Search Prob. | **0.014 (S)** | **0.019 (S)** | 0.449 (S) | 0.922 (–) | 0.082 (S) |
| *Interactions with Topic Familiarity* | | | | | |
| Response Quality | 0.848 (–) | 0.42 (S) | 0.24 (S) | **0.005 (S)** | **0.0 (S)** |
| Explanation Quality | 0.155 (S) | 0.671 (–) | **0.0 (S)** | **0.0 (S)** | **0.0 (S)** |
| Presentation Mode | 0.663 (–) | 0.752 (–) | **0.0 (S)** | **0.0 (S)** | **0.0 (S)** |

**Explanations.** Our experiments include two experimental conditions where explanations are not provided (i.e., EC9–EC10). To understand the impact of quality and presentation mode of explanations, we conducted an additional analysis on the data from HITs representing only EC1–EC8 (reported in the bottom part of Table 7.4) and we focused our analysis on these results. In terms of *explanation quality*, we observe that introducing noise in explanations has a statistically significant effect on almost all user-reported response dimensions, suggesting that noisy explanations have a strong impact on user experience in general. However, the quality of explanations does not impact user assessment of correctness and factuality, dimensions related to factual errors in the response. It means that users seem to assess the factual correctness of the response independently of the quality of the explanations provided by the system. In terms of *presentation mode*, we observe a statistically significant effect only for the presence/absence of explanations on the usefulness of the response—a statistically significant effect is observed in the top part but not in the bottom part of the table. Similarly, the user-reported conciseness, fairness, and relevance of the response are impacted only by the presence/absence of explanations. This implies insensitivity of the response dimensions to the way explanations are presented.

**User Ratings for Source, Confidence, and Limitations.** The impact of noise in the source is solely tied to the presence or absence of the source—no statistically significant effect is observed for EC1–EC8 (see Table 7.5). However, the presentation mode of the source affects user ratings for the explanations independently of its presence or absence (statistical significance persists when excluding EC9 and EC10), even though the source is presented in the same format in both presentation modes. This may be due to the wording of the question about source explanation in the questionnaire—it does not explicitly mention sources, and therefore is open to other interpretations, especially when sources are not provided. In the case of noise in confidence explanation, it significantly affects user ratings. However, concerning presentation mode, we can only discern the effect of its presence or absence, not the specific mode of presentation. Regarding limitations, there is no statistically significant effect of noise in the corresponding explanation, but there is of presentation mode. User ratings for explanations related to limitations are influenced by the presentation mode, not the mere presence of this explanation. This implies that, in general, the impact of noise on explanations is only related to the confidence and the impact of the presentation mode only to the limitations. The effect of quality and presentation mode on other explanations—based on the user ratings—was not significant in this user study.

**Query.** We do not observe any statistically significant effects of the query on the user-reported response dimensions (see Table 7.5). This suggests that the results are topic-independent and generalizable. The proposed user study design mitigates the impact of the query on the results. Additionally, the interaction between response quality, quality of explanations, or presentation mode and the query does not have a statistically significant effect on user-reported scores for response satisfaction, usefulness, and explanations (see Table 7.6).

**Topic Familiarity.** Workers report that they are rather familiar with the query topics, which indicates that the process of query selection was successful. Following our hypothesis, users' background knowledge about the topic affects how they assess the response. It is visible in the effects reported for almost all response dimensions (see Table 7.4). Similar effects are observed for the user's interest in the topic and the likelihood of the user searching for a similar query. Additionally, we observe a statistically significant effect of all these three indicators of user involvement in the topic on the user ratings for explanations (see Table 7.5). It implies that these factors that we cannot control and are completely user-dependent directly impact the assessment of the responses we examined in this user study.

In terms of the results of two-way ANOVA (see Table 7.6), we observe a statistically significant effect of the interaction between the user's familiarity with the topic and the response quality on the user ratings for explanations related to limitation and the system's confidence. It confirms the intuitive relationship between the user's background knowledge and the quality of the response on their ability to correctly assess the explanations provided by the system and deem it useful or not. We do not observe a statistically significant effect of interaction between response quality and familiarity with the topic on the usefulness of the response or user satisfaction in general. The interaction between the noise in the explanations and the familiarity with the topic has a statistically significant effect on the user ratings for all three explanations. It can follow from the fact that a user unfamiliar with the topic needs high-quality explanations from the system to be able to verify and use the provided response. The user ratings for explanations are influenced by the interaction between the way explanations are presented and the user's familiarity with the topic. This suggests that depending on the user's background knowledge, the preferred way of receiving explanations from the system may differ.

**Familiarity with Conversational Agents.** We observe a statistically significant effect of the user-reported frequency of interacting with conversational assistants (in general and for search specifically) on some of the user-reported response dimensions (see Table 7.4). Interestingly, we observe a medium-size effect of the frequency of using the conversational search on the user ratings for explanations related to source and limitations (see Table 7.5). Additionally, we observe that higher values for familiarity with conversational agents are associated with explanations without noise and visual presentation mode. It indicates that the user's familiarity with the system impacts their assessment of its additional components.

### 7.4.2 Effect of the Explanation Quality (RQ3.3a)

**Effect on the User-reported Response Dimensions** The top-left plot in Figure 7.4 shows that user-reported values for the usefulness of the responses are concentrated around higher values (3 and 4). However, noise in the explanations results in slightly lower usefulness scores. It indicates that the high-quality source, system confidence score, and information about the response limitations make the response more useful from the user's perspective. Minor differences in usefulness scores between perfect and imperfect responses (second and third set of bars in the plot) suggest that when explanations are not provided ("None" variant), users are less likely to object to the usefulness of imperfect responses. In general, the explanations are meant to increase the reliability and transparency of the system. However, they require additional time and effort from the user and the cost of "processing" explanations may be higher than the actual gain. This situation is visible in the second and third set of bars in the top-left plot in Figure 7.4 where the highest usefulness is reported for the responses that do not contain any explanations ("None" variant), independent of their quality. It suggests that the explanations either pollute the response or make the user more critical of it, resulting in reduced usefulness.

**Effect on user ratings of explanations** Looking at the means of user ratings for source and confidence explanations (top-right plot in Figure 7.4), ratings are again skewed towards higher values, and scores for accurate explanations are slightly higher than for noisy explanations, especially for confidence. This suggests that users perceive noisy explanations as less useful in understand-

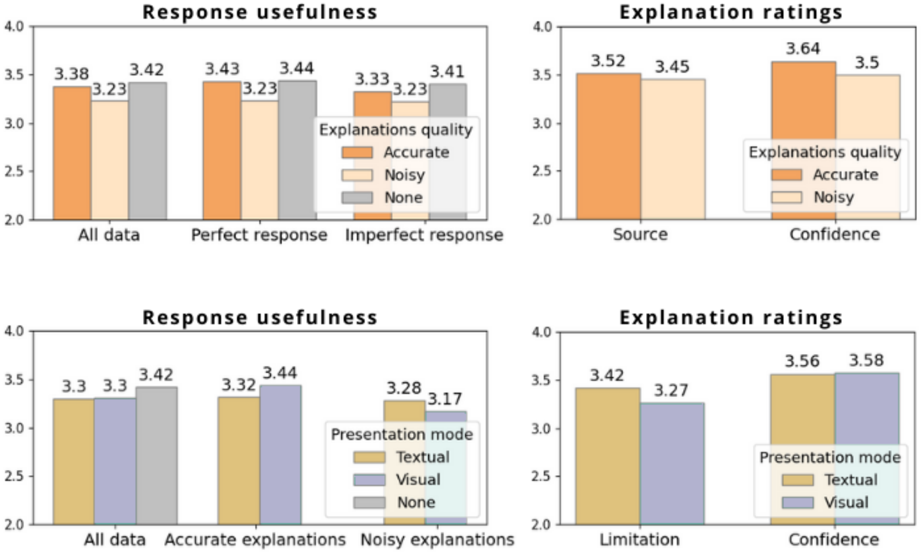


Figure 7.4: Mean scores for response usefulness and explanation ratings for different quality of the explanations (top) and presentation mode (bottom). All differences between the ratings within a given plot are statistically significant.

ing system confidence and attributed sources—we do not observe statistically significant differences for the explanations related to limitations.

### 7.4.3 Effect of the Presentation Mode (RQ3.3b)

**Effect on the user-reported response dimensions** On average, we do not observe differences in the usefulness scores between textual and visual modes, but usefulness scores are significantly higher when no explanations are provided (bottom-left plot in Figure 7.4). This is aligned with the one-way ANOVA results and suggests that the main issue is not the question of presentation mode but rather whether the explanations are necessary, hinting at the underlying trade-off between effort and gain. Nevertheless, we observe some differences in the user ratings for explanations when looking at responses accompanied by explanations with different quality. Namely, visual explanations result in higher usefulness scores for responses with accurate explanations, while in the case of noisy explanations workers find the textual format more useful.

**Effect on user ratings of explanations** Looking at the means of user ratings for explanations with respect to different presentation modes (bottom-right plot in Figure 7.4), the preferred presentation mode depends on the explained aspect of the response. (Note that user ratings for explanations related to the source are not informative in this case, as the source is always presented in the same way.) Namely, we observe slightly higher ratings for the textual presentation of limitations. In the case of confidence, the difference between presentation modes is very small with a slight preference towards visual presentation, which aligns with the results of one-way ANOVA. This suggests that further research is needed to better understand how to optimally integrate different aspects in the layout of transparent CIS responses.

### 7.4.4 Qualitative Analysis

We manually investigate the feedback given by crowd workers regarding their ratings for the source, confidence, and limitations explanations, seeking insights and suggestions to enhance their content and presentation (see Table 7.7). Many workers (18/160) pointed out that explanations related to limitations and confidence significantly enhanced their understanding of the constraints of both the system and the responses. The mention of encouragement towards information verification and critical thinking was consistent across various qualities (comments from EC1–EC8 HITs), and positive comments were also shared for noisy explanations (EC5–EC8). It suggests that workers may face challenges in identifying inaccuracies in the explanations. For instance, even though the provided sources did not align with the information in the response, none of the users mentioned these mismatches in their comments. Nevertheless, several crowd workers (4/160) emphasized the potential insufficiency of responses restricted to three sentences and a single source in certain situations. A few (3/160) crowd workers expressed uncertainty in interpreting explanations related to limitations and confidence scores, underscoring the need for additional explanations or tutorials describing the system interface before usage. For instance, some workers attempted to interpret the meaning of the confidence score on their own describing it as a *"transparency measure to indicate the system's level of certainty regarding the accuracy or relevance of the information shared"* or a *"model's estimate of the accuracy and reliability of its responses."* In terms of the presentation mode, one worker suggested that representing confidence score using percentages would be more precise and helpful than a *"wifi connection symbol."* This suggests that users might prefer a different display element, e.g., a fuel gauge (Shani *et al.*, 2013), and perhaps also a finer confidence scale (which would require a more precise estimation of confidence). In HITs with no explanations (EC9 and EC10), workers highlighted their lack of awareness regarding response limitations and confidence. Some workers attempted to gauge system confidence by searching for implicit confidence signals like *"I think"* or *"I believe"* in the responses (Radensky *et al.*, 2023). Overall, workers consistently emphasized that explanations enhance their understanding and encourage information verification and critical thinking. However, the comments reflect that workers are unlikely to identify flaws in the provided explanations.

Table 7.7: Selected comments provided by crowd workers as an explanation of their scores for additional information revealment.

| **Exp. Cond.** | **User's Feedback** | **Main Point** |
|---|---|---|
| EC1 | They all had a citation, but only one, and sometimes more perspectives were called for. | Limited sources |
| | I noticed the bars, but I wasn't sure how the assistant was coming up with its confidence level. | No confidence explanation |
| EC2 | The responses were supported by relevant and accurate information, addressing the questions adequately. | Relevant sources |
| | I understand it provided what it believed to be it's confidence in the answer, but knowing what that number is based on would be helpful. | No confidence explanation |
| EC3 | . . . a more explicit discussion of potential limitations or diverse perspectives could have been beneficial for a fuller understanding. | No limitations explanation |
| | This transparency in indicating confidence levels allows users to gauge the system's certainty in the provided responses. | Transparency |
| EC4 | The assistant is less certain about the accuracy of the information provided. This awareness is crucial for interpreting and verifying the information obtained from the assistant. | Critical evaluation |
| EC5 | . . . This acknowledgment highlighted the assistant's awareness of certain boundaries, privacy concerns, or potential ethical considerations related to specific topics. This helped me understand the constraints within which the assistant operates and the areas where it might not provide detailed information due to the nature of the query | Understanding system constraints |
| | The assistant consistently indicated its confidence level in each response. This feature provided transparency about the reliability of the information and the extent to which the assistant could guarantee accuracy. The confidence indicators were helpful in understanding the context and reliability of the provided information, allowing me to assess the responses with an awareness of the assistant's level of certainty. This transparency contributed to a more informed evaluation of the responses. | Transparency |
| EC6 | The assistant explicitly acknowledged potential limitations in several responses, mentioning factors such as bias in the questions. This helped me understand the potential constraints and caveats associated with the information provided. | Understanding response constraints |
| EC7 | I appreciated the symbol showing the confidence of the AI's response, but it seemed a bit abstract. I would have preferred a simple percentage confidence to the symbol that resembles a wifi connection symbol. I think that would have been more helpful and precise. | Preference towards textual confidence info |
| | The access to a confidence score provided by the system, which reflects the model's estimate of the accuracy and reliability of its responses. | Interpretation of confidence score |
| EC8 | . . . This transparency from the assistant helped me understand the potential biases or external influences that might affect the objectivity of the responses. This awareness of limitations enhances the user's critical evaluation of the information provided and promotes a more informed interaction with the assistant. | Encouraging information verification |
| | . . . The assistant frequently indicated its confidence levels, ranging from 80% to 100%, alongside certain responses. This information was provided as a transparency measure to indicate the system's level of certainty regarding the accuracy or relevance of the information shared. | Interpretation of confidence score |
| EC9 | The assistant's responses provided valuable information and insights on the respective topics, but it's essential to acknowledge potential limitations. | No info about limitations |
| EC10 | There were no qualifying statements like "I think" or "I believe" , it stated it's answers confidently as if they were facts. | Confidence indicated by NL statements |

## 7.5 Discussion

Our results show that high-quality explanations related to the source, system confidence, and response limitations increase the user-perceived usefulness of the response and user ratings for explanations. Additionally, noise in the explanations of the response provided by the system has a significant impact on user experience in general (almost all response dimensions are affected). These results align with previous research in AI-assisted decision-making claiming that confidence scores can help calibrate people's trust in the system model, but they are not sufficient to increase the success rate of interactions (Zhang *et al.*, 2020c). In our study, we observe a significant effect of familiarity with the topic on response assessment, indicating the need for the user's background knowledge to complement the system's errors (Zhang *et al.*, 2020c). In terms of user's sensitivity to inaccuracies in the responses and explanations, we show that users are not able to detect factual errors or biases in the provided information. The qualitative analysis shows that workers do not point out these inaccuracies explicitly. Similarly, they cannot identify flaws in the explanations related to response limitations. This aligns with previous research, demonstrating that explanations might cause users to follow the system's advice more often, even when it is wrong (van der Waa *et al.*, 2021). Our study is not conclusive about the preferred way of presenting explanations to the user. We find that limitations tend to receive higher user ratings when presented in a textual form, whereas, for confidence, we observe the opposite trend, which complies with the findings reported in the field of recommender systems (Shani *et al.*, 2013). Additionally, limited mentions of the presentation mode in the free-text feedback obtained from crowd workers may imply that the format of explanations is not a crucial factor in this setting.

Insights from this study about communicating explanations to facilitate users' assessment of the provided information need to be put in a broader context of system explainability and the associated effort/gain trade-off (Cheng *et al.*, 2019). While these explanations complement the system response with components that enable users to assess responses more objectively, they demand more time and effort than merely reading the provided response. Optimizing user gain is a complex task influenced by various factors. Firstly, the relationship between the user's gain and the effort associated with the amount of additional information is not linear; while more explanations generally increase gain, there is a tolerance threshold. Exceeding that threshold may overwhelm users, causing a drop in gain. Secondly, the overall quality of the system's response and explanations significantly impacts gain. This is evidenced by our findings: users struggle to detect flaws in provided responses when explanations contain noise or errors, and providing no explanations is more useful than providing noisy ones. Thirdly, the relevance of explanations depends on the topic's complexity and user familiarity, with more complex topics benefiting from adjusted and detailed information. Additionally, the optimal effort-gain trade-off is likely to be user-dependent, requiring personalized adjustments in the amount, level of detail, and presentation of the information, which is evidenced by various preferences for the confidence display we observed in the feedback. To our knowledge, investigating the adaptation of responses based on user preferences, previous interactions with CIS systems, and topic complexity has yet to be explored.

## 7.6 Conclusions

Response transparency has not received significant attention in a CIS setting. Our user study addresses this gap by examining various ways of explaining the source of the information provided by the system, the system's confidence in the response, and its limitations. In answer to **RQ3.3** *(How to generate responses transparent about the system's confidence and limitations?)*, we have explored the effect of noise and different presentation modes of these explanations on users' assessments of responses and explanations. Our findings show that high-quality explanations about source attribution, system confidence, and response limitations improve user-perceived usefulness and explanations ratings. Results reveal lower user-reported usefulness scores when explanations contain noise, although these scores seem insensitive to the quality of the response. In terms of presentation mode, we do not observe significant differences between visual and textual explanations—suggesting that the format of explanations may not be a critical factor in this setting—but users presented with no explanations, surprisingly, found the responses more useful. To our knowledge, this study is the first to examine response transparency in CIS, highlighting the need for further research to enhance transparency in CIS responses.

# Chapter 8

# Conclusions

*I stepped from Plank to Plank*
*So slow and cautiously*
*The Stars about my Head I felt,*
*About my Feet the Sea.*

*I knew not but the next*
*Would be my final inch —*
*This gave me that precarious Gait*
*Some call Experience.*

— **Emily Dickinson**

This chapter concludes the thesis by summarizing its key contributions, revisiting the main research questions, and discussing how they have been addressed through the proposed methods. It highlights the findings' implications for the development of transparent, reliable, and explainable conversational information-seeking systems (CIS). Additionally, the chapter acknowledges the limitations of the work and identifies avenues for future research.

## 8.1 Summary

In this section, we revisit the main findings of this thesis, guided by the research questions outlined in the introduction.

### 8.1.1 CIS System Baseline

To advance research in the area of transparent, factual, and grounded CIS systems, we introduced a competitive baseline for both the retrieval component, to collect the sources answering the user's query, and the generation component, to synthesize this information into a natural answer.

***RQ1a: What are strong baselines for passage retrieval in CIS systems?*** We first established a strong retrieval baseline by reproducing existing approaches for conversational passage retrieval in the context of TREC CAsT (see Chapter 3). Specifically, we replicated the top-performing TREC CAsT'22 submission and the organizers' baseline, both following a standard retrieve-then-rerank pipeline with a query rewriter. Our results align with previous research (Yan *et al.*, 2021), confirming that more advanced retrieval models consistently improve performance across metrics and datasets. We experimented with different query rewriting methods within an alternative retrieval pipeline, demonstrating that applying different methods at various stages can be beneficial. Based on our findings, we concluded that a combination of sparse and dense retrieval, enhanced with pseudo-relevance feedback in the first-pass retrieval and pointwise/pairwise reranking, coupled with a fine-tuned query rewriting component, represents a strong baseline for conversational passage retrieval.

***RQ1b: What are strong baselines for response generation in CIS systems?*** We adopted an approach inspired by retrieval-augmented open-domain QA using an off-the-shelf LLM without additional training (Ren *et al.*, 2025; Ram *et al.*, 2023; Muhlgay *et al.*, 2023) (see Section 6.3.2). Our prompting strategy is inspired by retrieval-augmented QA LLM instructions from Ren *et al.* (2025). As a second baseline, we tested Chain-of-Thought prompting (Wei *et al.*, 2022), incorporating a single in-context learning (ICL) demonstration manually curated from the TREC CAsT'22 dataset (Owoicho *et al.*, 2022). The performance of both baselines was evaluated on the TREC RAG'24 dataset using the AutoNuggetizer framework, confirming their competitiveness.

### 8.1.2 Understanding CIS Limitations

Recognizing that high retrieval performance does not necessarily guarantee useful responses, we explored factors limiting response quality and examined finer-grained information units beyond documents or passages to enhance the coverage, accuracy, and completeness of the responses.

***RQ2.1: Which limitations in the responses are detectable by users?*** We conducted two crowdsourcing experiments examining user perceptions of unanswerable questions and incomplete responses in a setting based on the TREC CAsT benchmark (Dalton *et al.*, 2020; Owoicho *et al.*, 2022) (see Chapter 4). Specifically, we explored users' ability to recognize factual inaccuracies, pitfalls, and biases related to viewpoint diversity by carefully controlling experimental conditions in manually crafted responses that simulate CIS interactions. Our findings indicated that users are more adept at detecting viewpoint diversity issues and response biases than factual errors or problems related to source validity. The insights gained from these experiments guided our efforts toward building transparent and reliable response mechanisms. Specifically, users' difficulty in detecting factual inaccuracies motivated our work on unanswerability detection and the development of more granular response generation techniques.

***RQ3.1: How to identify core information units in the relevant passages that need to be included in the response?*** We created CAsT-snippets, a high-quality dataset for conversational information seeking, featuring snippet-level annotations for all queries in TREC CAsT '20 (Dalton *et al.*, 2020) and '22 (Owoicho *et al.*, 2022) (see Chapter 5). To ensure data quality, we extensively explored different task designs and trade-offs for snippet annotation through crowdsourcing, experimenting with various interfaces and worker qualification criteria. Our approach was informed by a preliminary study evaluating multiple annotation setups, platforms, and worker pools. Based on these findings, we collaborated closely with a selected group of highly engaged crowd workers, releasing tasks in daily batches and providing continuous feedback. Compared to related datasets such as SaaC (Ren *et al.*, 2021) and QuaC (Choi *et al.*, 2018), the CAsT-snippets dataset provides a greater number of annotations per input text, with snippets that are longer on average. Our close collaboration with experienced annotators ensured high-quality data and yielded valuable insights to inform future response generation methods.

***RQ2.2: How to detect factors contributing to incorrect, incomplete, or biased responses?*** We proposed a mechanism for detecting unanswerable questions where the correct answer is either absent from the corpus or cannot be retrieved (see Chapter 5). Our baseline approach employs a sentence-level classifier to determine whether an answer is present, then aggregates these predictions at the passage level before producing a final answerability estimate across top-ranked passages. We evaluated multiple variations of this method with different configurations. To enable training and evaluation, we extended the CAsT-snippets dataset with answerability labels at the sentence, passage, and ranking levels, introducing the CAsT-answerability dataset. Despite their simplicity, our baseline models outperformed a state-of-the-art LLM in answerability prediction. By assessing whether a question can be at least partially answered using the top-ranked passages, we reduce the risk of generating responses based on irrelevant or nonexistent answers, thereby mitigating hallucinations (Ji *et al.*, 2023).

### 8.1.3 Addressing CIS Limitations

Ensuring grounding and transparency in CIS-generated responses is crucial for fostering user trust and enabling informed decision-making (Radlinski and

Craswell, 2017; Azzopardi *et al.*, 2018). Our goal was to generate responses that (1) synthesize the requested information, (2) ground it in specific facts from retrieved passages, (3) articulate the system's confidence, and (4) acknowledge its limitations. By explicitly communicating these limitations, we encourage users to examine the responses more critically.

***RQ3.2: How to ensure the grounding of responses in the retrieved sources?*** Grounding responses in specific facts from retrieved passages can be ensured by operating on fine-grained information snippets (see Chapter 6). We proposed `GINGER`, a multi-stage approach that constructs responses by leveraging information nuggets from top-ranked passages. `GINGER` employs clustering, reranking, summarization, and fluency enhancement to produce concise, informative, and non-redundant responses. This method offers several advantages: (1) maximizing information within length constraints, (2) providing source attribution for verifiability, (3) guiding users with relevant follow-up questions, and (4) allowing control over response completeness. Automatic evaluation demonstrated that `GINGER` outperforms the baseline in grounding and source entailment. Evaluation with the AutoNuggetizer framework further showed that `GINGER` achieves top performance in the Augmented Generation task at the TREC RAG'24 track. The human evaluation showed a clear preference for `GINGER` in terms of conciseness and confirmed that it generates more useful follow-up questions.

***RQ3.3: How to generate responses transparent about the system's confidence and limitations?*** We conducted a user study to explore different ways to communicate (1) the sources of information, (2) the system's confidence in its responses, and (3) the system's limitations (see Chapter 7). We examined how noise and different explanation formats affect users' assessments of both responses and explanations. Drawing inspiration from prior work on recommender systems, we experimented with various methods to convey this information, including natural language explanations (Rechkemmer and Yin, 2022), UI elements (Lu and Yin, 2021), and granular confidence scales (Shani *et al.*, 2013). Our findings showed that high-quality explanations about source attribution, system confidence, and response limitations improve user-perceived usefulness and ratings. However, noise in explanations negatively impacts the overall user experience. These results align with AI-assisted decision-making research, which suggests that while confidence scores can calibrate trust, they don't necessarily increase interaction success (Zhang *et al.*, 2020c). We also found that users' topic familiarity strongly influences response assessments, underscoring the need for background knowledge to complement system errors (Zhang *et al.*, 2020c). Additionally, users struggled to detect factual errors or biases in responses, as crowd workers rarely pointed out inaccuracies or flaws in explanations. This supports prior research showing that explanations can increase trust, even when the system is incorrect (van der Waa *et al.*, 2021). Our study did not provide a clear answer on the preferred explanation format. However, users tended to rate textual explanations of system limitations more favorably, while confidence indicators are better received in other formats, a trend consistent with recommender system research (Shani *et al.*, 2013). Free-text feedback from crowd workers suggests that presentation format may not be a critical factor in this context.

## 8.2 Limitations

We acknowledge several limitations that apply throughout this work.

### 8.2.1 Nugget Identification and Answerability Detection as Binary Tasks

In the CAsT-snippets dataset, we approach information nugget identification as a binary decision while in reality, it is a more granular task, as snippets can vary in relevance and complexity (see Chapter 5). They may contain exact facts fully answering the question, additional enriching details, or only address some aspects of the question. Similarly, we frame answerability as a binary decision: a question is considered answerable if at least one sentence in the retrieved passages contains the answer (or part of it). However, answerability is inherently more nuanced. A system might retrieve partial but insufficient information, leaving gaps in the response. A more realistic approach would be to model answerability on an ordinal scale (e.g., unanswerable, partially answerable, fully answerable). However, this would require complete answers as ground truth, explicitly specifying different facets of the response–data that is currently unavailable in any existing information-seeking dataset, conversational or otherwise that we are aware of.

### 8.2.2 Restricted Scope of Answerability

Our experiments define answerability based on the top-$n$ retrieved passages, as determined by ground truth relevance judgments from TREC CAsT (see Chapters 4 and 5). However, in practical CIS scenarios, answerability may need to be assessed across the entire corpus, multiple corpora, or external expert knowledge. Furthermore, our work considers answerability primarily in terms of factual correctness and source attribution. We attempted to introduce intermediate conditions in our user studies, such as factually correct responses without source attribution and factually incorrect responses with invalid sources, but these conditions still do not fully capture the spectrum of answerability (e.g., severity of factual errors, number of sources used, source credibility). Addressing this would require significantly more granular experimental conditions and extensive human annotations.

### 8.2.3 Operationalization of Viewpoint Diversity

In our user studies, viewpoint diversity is operationalized in a simplified manner: a response is considered diverse if it covers at least two different perspectives (see Chapter 4). However, this does not ensure that all possible viewpoints are represented. Due to response length constraints, some perspectives may be omitted. Additionally, not all viewpoints may be present in the top-ranked passages, and identifying them would require expert knowledge or a dataset with explicitly annotated answer facets and sources–which, to our knowledge, does not yet exist for information-seeking tasks.

### 8.2.4 Limitations of User Studies

To manage the complexity and cost of large-scale data collection, our user studies focus on single-turn interactions, despite the inherently multi-turn nature of CIS dialogues (see Chapters 4 and 7). While inspired by the TREC CAsT setting, our analysis is constrained to a limited number of representative queries. This means our experiments do not fully reflect real-world CIS interactions, where user needs evolve dynamically over multiple turns. Another limitation is our reliance on Amazon Mechanical Turk (MTurk) crowd workers, who may not fully represent the diversity of CIS users. Additionally, our studies do not explicitly control for participant biases, leaving this as an open area for future investigation. Lastly, our findings are inherently tied to the test collections used in our experiments. While they offer valuable insights, results may differ when applied to other datasets or real-world CIS systems.

## 8.3 Broader Context

As conversational systems increasingly take on responsibilities traditionally handled by users, the risk of misalignment with user needs grows. Without a deep understanding of user needs and accurate personalization, systems may generate responses that are misleading, incomplete, or suboptimal. As conversational agents continue to evolve, striking a balance between automation and user control will be critical to ensuring both reliability and user satisfaction. This challenge underscores the importance of explainability, ensuring that users remain informed and empowered rather than overly dependent on system outputs. A poorly designed CIS system could inadvertently reinforce biases, increase uncertainties, or harm user trust, ultimately reducing its effectiveness in real-world applications.

The findings of this thesis should be interpreted within the broader context of system explainability and the associated user effort/gain trade-off (Cheng *et al.*, 2019). While explanations enrich system responses by providing transparency, uncertainty estimates, or source attributions, they also require additional cognitive effort from users. As illustrated in Figure 8.1, explainability exists on a spectrum: at one extreme, systems offer no insights into their limitations or

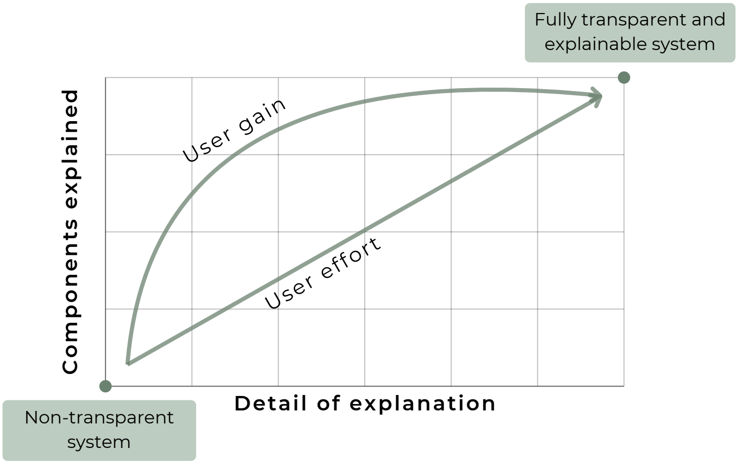


Figure 8.1: The trade-off between user effort and gain related to explanations provided by the system.

reasoning processes; at the other, they provide comprehensive details on response generation, potential pitfalls, and reliability signals. Optimizing user gain is a complex process—more explanations generally improve user trust and response assessment, but excessive details can overwhelm users, reducing overall benefit. Our findings highlight that explanations must be both accurate and appropriately tailored. Users may still struggle to detect errors, especially when explanations contain noise, and there is a risk that explanations may lead users to over-rely on incorrect AI-generated responses (van der Waa *et al.*, 2021). Additionally, the optimal balance between effort and gain is likely user-dependent, necessitating personalized adjustments in the amount, level of detail, and presentation of explanations.

Finally, the increasing integration of LLMs into conversational systems brings significant ethical and practical concerns. LLMs are prone to hallucinations, generating plausible but incorrect information, which can mislead users in critical decision-making contexts. Additionally, their deployment involves high computational costs, which limit accessibility and concentrate advancements within resource-rich institutions, widening the gap between well-funded and lower-resourced research communities. The environmental impact of training and running large-scale models also raises ethical concerns, as these models require vast amounts of energy (Chien *et al.*, 2023). Furthermore, reliance on LLMs via APIs introduces data privacy and governance challenges, as organizations must carefully consider how user queries and interactions are processed, stored, and potentially used for further model training. Addressing these challenges requires a multi-faceted approach, balancing innovation with sustainability, accessibility, and ethical responsibility.

## 8.4 Future Directions

Our findings highlight the need for further research on response limitation identification and user-centered evaluation of responses. Additionally, exploration of explainability, personalization, and user biases will be crucial for enhancing user trust and system reliability.

### 8.4.1 Response Limitation Detection

Beyond answerability detection, future work should extend response limitation detection to capture a broader range of system constraints that impact the quality of generated responses. These include viewpoint diversity, particularly when handling controversial topics (Draws *et al.*, 2021b); partial unanswerability, where only some aspects of a question can be addressed; temporal validity, ensuring responses remain accurate within specific timeframes (Campos *et al.*, 2015); and bias in queries, which can shape both retrieval and response generation (Azzopardi, 2021). Additionally, subjectivity in source text and incomplete background information can impact a system's ability to generate reliable responses. Future models should integrate methods for detecting and explicitly communicating these limitations to users, allowing them to interpret responses more critically.

### 8.4.2 Experimental Scope

Future research should expand the experimental scope by incorporating a broader range of topics in user studies to increase result sensitivity. Additionally, alternative scales, such as magnitude estimation (Turpin *et al.*, 2015), could provide more granular insights into user satisfaction with system responses. The modular design of GINGER offers an opportunity to explore constraint-based response generation, including adapting response length based on user preferences and managing redundancy in multi-turn conversations by considering previously disclosed information. Future studies should also analyze explanations in CIS settings by examining the impact of response specificity, interactivity, and conversation history on user experience.

### 8.4.3 Personalization

Communicating through natural language dialogue enables the system to infer more about a specific user. By maintaining a model of the user's background, domain knowledge, goals, and capabilities, the overall effectiveness of a conversational search system is enhanced (Anand *et al.*, 2021). Since the information need is inherently tied to the user's context (Wilson, 1981, 1999), only a system that understands both the user's needs and their context can offer truly personalized experiences (Zhang *et al.*, 2018b). A critical aspect of personalization is gathering user preferences and storing them, alongside personal information, in a form that is easily accessible by the system. This could be in the form of sets of sentences in natural language (Zhang *et al.*, 2018a) or through the use of *personalized knowledge graphs* (PKGs) (Bernard *et al.*, 2024; Skjæveland *et al.*, 2024). A PKG is defined as a structured source of knowledge about entities personally related to the user and the relationships between them (Balog and Kenter, 2019). As a natural choice for personalized search systems, PKGs can support multiple components, such as query understanding, personalized ranking of results, and proactive system initiatives (Balog and Kenter, 2019). These graphs can store personal knowledge about a user's profession, hobbies, or preferences (Tigunova *et al.*, 2020). Despite advances, personalized CIS systems remain an underexplored area, particularly in balancing personalization with fairness or utilizing PKGs for storing user data. Another open challenge in the CIS system is the personalized presentation of results, which involves adjusting content, terminology, and response length to meet individual user preferences.

### 8.4.4 Addressing Cognitive Biases

Users engaging in information-seeking activities are susceptible to various cognitive biases, which hinder the absorption of information provided by the system. Cognitive biases significantly affect information seeking, retrieval behaviors, and outcomes (Azzopardi, 2021). A major concern is whether the impact of these biases is amplified by the vast amount of already biased information available on the Web, such as activity bias, data bias, algorithmic bias, and personalization (Baeza-Yates, 2018). Common biases in search settings include *confirmation bias* (a tendency to seek confirmatory information) and *projection bias* (projecting current thoughts onto past or future experiences) (Azzopardi, 2021). Identifying and addressing cognitive biases in conversations remains a significant challenge. Existing research on cognitive biases primarily focuses on traditional search engines, which offer less constrained response formats compared to CIS systems, typically providing short textual responses. One of the key open questions is how to distinguish between *bias* and *preference* in personalized CIS systems (Gerritse *et al.*, 2020).

# List of Figures

# List of Tables

# Appendix A

# Limitations of CIS Systems

This appendix contains additional details, results, and analysis for Chapter 4.

## A.1 User Studies Design

The design of the *answerability study* and the *viewpoints study* followed the same principle, where workers were asked to complete one HIT, consisting of ten query-response pairs. The task consisted of:

- HIT instructions
- Ten CIS interactions
- Demographics questionnaire

The instructions differ slightly between the studies. In the *answerability study*, we used the following instructions:

> *You are a search system user interested in specific topics. You pose a set of questions and get responses from the system in a form of short texts. Read carefully each question and the system's response and answer the questions below.*
>
> *Rely solely on you own judgment, restrain from using additional sources other than the ones referenced in the response.*
>
> ***Important note:*** *We kindly request your utmost attention and thoroughness in responding to the questions. Please be aware that*

Table A.1: Comparison of different datasets containing answerability labels. QPP indicates query-passage pairs.

| Demographic Information | Option | User study | |
|---|---|---|---|
| | | Answerability | Viewpoint |
| Age | 18-30 | 34 | 3 |
| | 31-45 | 35 | 12 |
| | 46-60 | 19 | 10 |
| | 60+ | 7 | 2 |
| | Prefer not to say | 1 | 0 |
| Education | High School | 19 | 8 |
| | Bachelor's Degree | 59 | 16 |
| | Master's Degree | 15 | 2 |
| | Ph.D. or higher | 2 | 0 |
| | Prefer not to say | 1 | 1 |
| Gender | Male | 44 | 15 |
| | Female | 52 | 12 |
| | Other | 0 | 0 |
| | Prefer not to say | 0 | 0 |

> *answers lacking sufficient justification that indicate a lack of attentiveness, may be subject to rejection.*

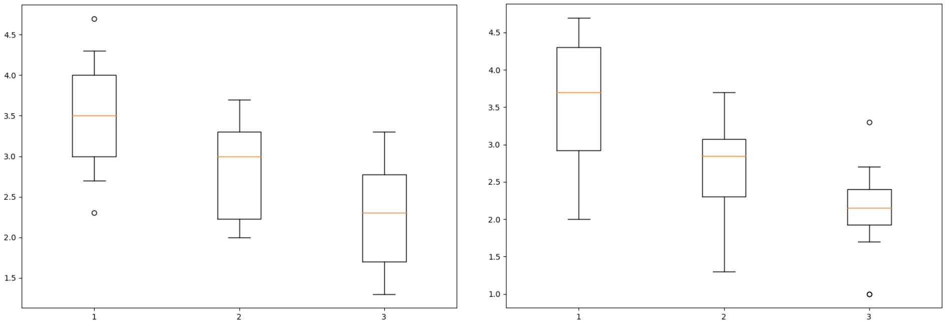
In the *viewpoints study*, we used the following instructions:

> *You are a search system user interested in specific topics. You pose a set of questions and get responses from the system in a form of short texts. Read carefully both the question and the system's response and answer the questions below.*
>
> ***Important note:*** *We kindly request your utmost attention and thoroughness in responding to the questions. Please be aware that answers lacking sufficient justification that indicate a lack of attentiveness, may be subject to rejection.*

Demographic information for both user studies is presented in Table A.1.

## A.2 Responses in the *viewpoints study*

In the *viewpoints study*, we focus on a widely understood diversity of viewpoints. It is left to the user to judge whether the expressed viewpoints are diverse enough or not. The accurate response equally covers various points of view and/or aspects of the topic. The flawed response only mentions one point of view and/or aspect of the topic or mentions several but elaborates only on one



Diversity

Balance

Figure A.1: Distribution of diversity and balance scores (on Y axis) for different variants of the responses (on X axis) collected in the survey preceding the *viewpoints study*.

of them. The last experimental condition with a lack of both diversity and balance makes no sense because the text discussing only one point of view can not be unbalanced.

### A.2.1 Answers Quality Assurance

We introduce an additional step for *viewpoints study* to validate the understanding of our proposed response dimensions. This additional step for optimizing queries and responses is introduced only for the *viewpoints study* because the problem of controversy and topic broadness is more subjective than the problem of answerability. This step will help us identify question-answer pairs that are not good representatives of the problem.

We select 12 questions and manually create 3 variants of the response for each of them. We create small surveys where expert annotators are presented with three topics and lists of recommended resources used to generate the answers. The expert annotators are asked to explore the topics to become familiar with basic concepts and problems related to each topic. We provide them with the links to the entire web pages or articles, without information about the specific passages that were used for generating the response. Then, they are presented with different variants of the answers to the questions about explored topics and asked to judge the diversity and balance of each of the provided question-answer pairs. For each question-answer pair used in the *answerability study*, we collect scores between 1-5 for diversity and balance from three expert annotators (12 annotators in total). We opted to release these surveys on the crowdsourcing platform because we wanted to have control over the time spent by the participants on actually exploring the topics. We employed PhD candidates for their academic skills in exploring new domains due to the nature of their work. One expert annotator's answers were excluded from the study due to an incorrect understanding of the task reported in the feedback after task submission.

The setup of the study encourages exploring the topics before assessing the questions. Therefore, the obtained scores can be assumed to come from users with high familiarity with the topics (experts). We focus on the outliers in the scores for $EC_1^V$ (in theory the best answer - multiple viewpoints, balanced) and $EC_3^V$ (in theory the worst answer - single viewpoints, imbalanced (see Figure A.1). We exclude questions for which the score for $EC_1^V$ is the furthest below the 1st quartile in terms of diversity and the question for which the score for $EC_3^V$ is the furthest above the 3rd quartile in terms of balance. These intuitively correspond to the question where $EC_1^V$ is not diverse enough, and $EC_3^V$ is too balanced. We exclude the query for which the response variant corresponding to the first experimental condition (multiple viewpoints covered to the same extent) is judged as not diverse enough and the query for which the response variant corresponding to the third experimental condition (single viewpoint mentioned and covered) is judged as too balanced. This additional step aims at improving the quality of input data for the user study, ensuring meaningful differences between answer variants and similar quality of answers between the questions.

# Appendix B

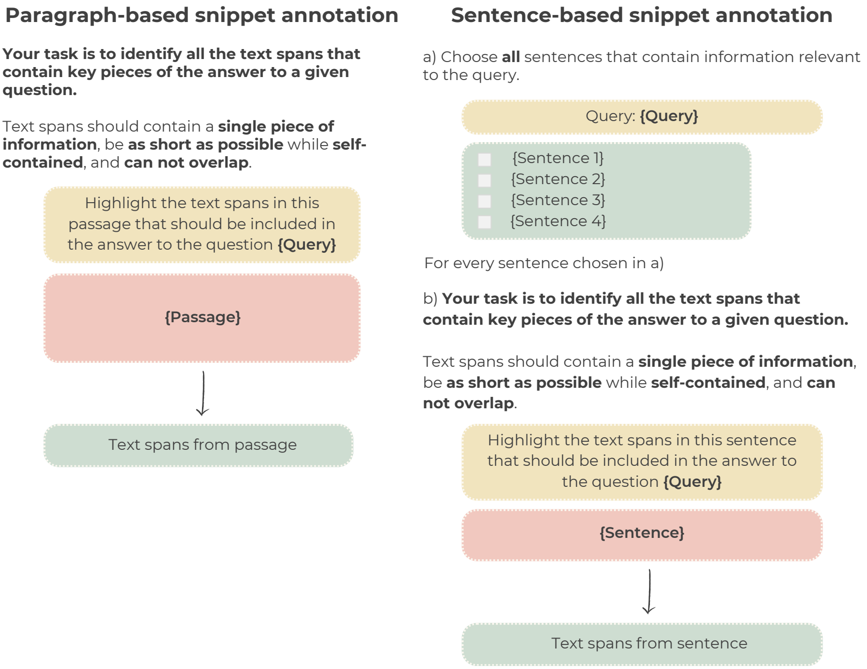
# Snippet-level Annotations for Predicting Query Answerability

This appendix contains additional details, results, and analysis for Chapter 5.

## B.1 CAsT-snippets Dataset

The crowdsourcing tasks for TREC CAsT'20 and '22 datasets were released on Amazon MTurk only for a small group of trained crowd workers. The qualification task has been completed by 20 MTurk workers with the results manually verified by experts. From the 20 workers who completed the qualification task, we chose 15 that provided results of the highest quality. Each worker received feedback on the provided responses and additional questions if needed. Several rounds of discussion that emerged from the qualification task resulted in extended guidelines with additional points addressing the challenging aspects identified in the annotation task:

- Task name: Snippet annotation
- General instruction: Identify all the text spans that contain key pieces of the answer to a given question
- Detailed instructions: Your task is to identify all the text spans that contain key pieces of the answer to a given question. Text spans should contain a single piece of information, be as short as possible while self-contained, and can not overlap.



**Paragraph-based snippet annotation**

**Your task is to identify all the text spans that contain key pieces of the answer to a given question.**

Text spans should contain a **single piece of information**, be **as short as possible** while **self-contained**, and **can not overlap**.

Highlight the text spans in this passage that should be included in the answer to the question **{Query}**

**Sentence-based snippet annotation**

a) Choose **all** sentences that contain information relevant to the query.

For every sentence chosen in a)

b) **Your task is to identify all the text spans that contain key pieces of the answer to a given question.**

Text spans should contain a **single piece of information**, be **as short as possible** while **self-contained**, and **can not overlap**.

Highlight the text spans in this sentence that should be included in the answer to the question **{Query}**

Figure B.1: Different task designs for snippet annotation.

- Extended guidelines: Your task is to identify all the text spans that contain key pieces of the answer to a given question. Text spans should contain a single piece of information, be as short as possible while self-contained, and can not overlap. The task of text spans annotation is non-trivial and requires a thorough reading of both questions and the accompanying passages. While performing the task keep in mind that:
  - An answer may be present in many different forms. It can be a short name, numerical value, or a longer explanation spread over several sentences. We want you to choose spans that are as short as possible while self-contained. Each chosen text span should make sense given the question, not necessarily as a standalone text.
  - Several different answers may be given in one passage. We want you to choose all the answers as separate spans. If the context of the answer changes its meaning or makes it different from other selected spans, the context should be included in the span as well.
  - Passages may contain noise in the form of links, and web page headlines. Try not to contain it in the spans, while remembering to keep spans self-contained.
  - The answer may not be present in the passage. Don't select any span if you think that there is no information in the passage that could

answer the question. However, we want you to select text spans that provide a partial answer even if full is not present.

- In the case of yes/no questions, it may be helpful to imagine how you would follow up the yes or no with some explanation and then find those pieces of information in the passage.
- In this task, we don't consider the problem of subjectivity of the statements in the passages. The text span that is obviously an opinion is as good as a text span that is perfectly objective.

We want to build your intuition about what is expected from you by providing you with a brief explanation of the problems that we want to solve with the collected data:

- Detecting unanswerability and partial answerability in questions given a passage → Given a question-passage pair we want to be able to say to what extent the question can be answered. Eventually, we also want to point out the missing pieces of information in the passage.
- Generating concise, informative answers grounded in statements from the passages → Given a question-passage pair we want to be able to detect the text spans that contain key pieces of information required to answer the question and generate the response by looking only at the question and the selected text spans.

The plan for data annotation:

- We will release one batch containing 25-85 question-passage pairs for every worker every day for the upcoming two weeks. The amount of HITs released will differ between the workers every day, but the overall amount of assigned HITs will be almost the same for everyone.
- HITs in each batch contain questions about one specific topic. Even though questions are related, they should be treated and answered independently of each other.
- You'll have 24h to annotate your batch. We'll let you know here every time a new set of tasks is released.
- We will try to be available on Slack every day so that you can ask questions.
- After obtaining all the annotations from all the workers we will do a manual verification of a sample of submissions and grant the 3 top-performing workers with the bonus.
- We need one person to spend a few minutes with us tomorrow early in the day to verify that there are no technical issues with the released tasks.

We will release a new batch for everybody every day before 6 am EST. The HITs from each batch will be available for you for 3 days. However,

Table B.1: Comparison of different datasets containing answerability labels. QPP indicates query-passage pairs.

| | **SQuAD 2.0** | **CAsT-snippets** | **CAsT-unanswerable** |
|---|---|---|---|
| # questions | 142,192 | 371 | 371 |
| # ans. questions | 92,749 | 365 | 0 |
| # unans. questions | 49,443 | 6 | 371 |
| # QPP | 142,192 | 1,855 | 1,855 |
| # ans. QPP | 92,749 | 1778 | 0 |
| # unans. QPP | 49,443 | 77 | 1,855 |
| # sent. w/ answers in ans. QPP | 106,146 | 6395 | 0 |
| # sent. w/o answers in ans. QPP | 369,025 | 5839 | 0 |
| # sent. w/o answers in unans. QPP | 250,365 | 453 | 12,751 |

if you manage to complete all HITs assigned to you within 24 hours (till 6 am EST the following day), you will be granted an additional bonus. We still want you to complete all the HITs from every batch assigned to you. If you failed to do so within 3 days, your HITs will be released for someone else in the group.
We are not expecting you to find anything in the passage. Going for the "no answer" option is perfectly fine. Basically, there are three options in this task: no answer, one span/multiple spans containing the answer, and one span/multiple spans that contain the partial answer. The third case is the most tricky. There, we are interested only in spans that actually contain a part of the answer but don't answer the question entirely (some information is still missing). It is not enough for the span to be somehow relevant, it needs to actually answer some part of the question. For example, if you have a question about how is a carrot cake made but the passage lists only ingredients. Ingredients are still a partial answer and they should be selected.

- Context example: Question: *"How many words must a man type, if a man is to type words?"*. If the passage is *"It's good to type words. Words are useful. A man must type 100 words if a man is to type words."* the span would be *"100"*. If the passage is *"It's good to type words. Words are useful. A man must type 100 words if he is a dancer."* the span would be *"100 words if he is a dancer"*. But if the passage is *"It's good to type words. Words are useful. A man must type 100 words if he is a dancer. A man must type 101 words is he is a writer."* the spans would be: *"100 words if he is a dancer"* and *"101 words is he is a writer"*.

The preliminary task study of snippet annotation was considered in two different task designs: paragraph-based and sentence-based. The annotation task designs are presented in Figure B.1.

Table B.2: Answerability annotations from the CAsT-answerability dataset on sentence (S), passage (P), and ranking (R) levels for the following query: *What's important for me to know about the safety of smart garage door openers?*

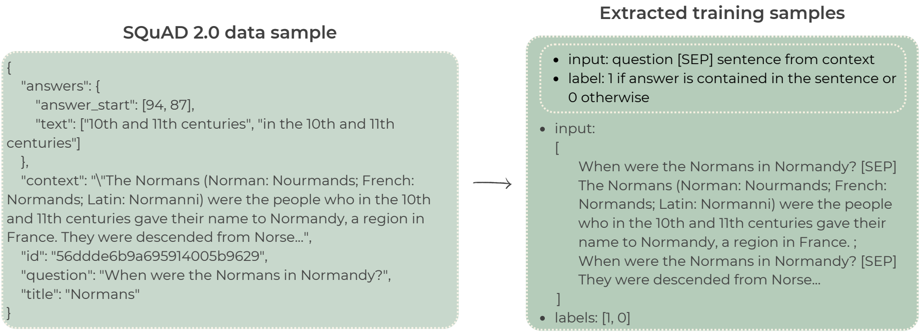
| Passage ID | Sentence from the Passage | Answer. S | P | R |
|---|---|---|---|---|
| MARCO_7107975 | If you're looking to get a little more creative with ... | 0 | 1 | 1 |
| | Echo can connect with this device to tell you if you've ... | 1 | | |
| | You can even say, Alexa, tell Garageio to close my ... | 1 | | |
| MARCO_8270733 | The Good The Chamberlain MyQ Garage is one of ... | 0 | 0 | |
| | The Bad It works with a growing list of other smart ... | 0 | | |
| | The Bottom Line Chamberlain's MyQ Garage should ... | 0 | | |
| | The MyQ isn't a garage door opener as it says in ... | 0 | | |
| | It works well and does exactly what you'd expect. | 0 | | |
| MARCO_8270733 | The LiftMaster MyQ Home and Property Control ... | 0 | 1 | |
| | Imagine receiving an alert if you left your garage or ... | 1 | | |
| | *** Note: Requires LiftMaster MyQ hardware and a ... | 1 | | |
| | Learn more about compatible products and find a ... | 0 | | |

## B.2 CAsT-answerability Dataset

The CAsT-snippets dataset is built on the top-relevant passages and it is highly imbalanced in terms of answerable and unanswerable query-passage pairs. To address this issue we build a synthetic unanswerable CAsT dataset, referred to as *CAsT-unanswerable*. Namely, for each query in the CAsT-snippets dataset, we add 5 random non-relevant passages according to ground truth judgment as passages without an answer. Choosing passages with low relevance scores instead of any random passages from the corpus increases the difficulty of unanswerability detection as passages in the pool are taken from the top of rankings submitted by participants to the TREC CAsT. The resulting dataset which is a concatenation of CAsT-snippets and CAsT-unanswerable, named CAsT-answerability, contains around 1.8k answerable and 1.9k unanswerable question-passage pairs. Statistics comparing different datasets considered in this work can be found in Table B.1. A sample from the CAsT-answerability dataset can be found in Table B.2. An example of extracting answerability labels on sentence level from SQuAD 2.0 samples is presented in Figure B.2.

Answerability score prediction is performed with ChatGPT. We consider two settings to predict the answerability of a question: given a passage (analogous to the passage-level setup) and given a set of passages as input (analogous to the ranking-level setup). We prompt the model to verify whether the question is answerable in the provided passage(s) and return 0 or 1 accordingly.

In the passage-level setup, the passage-level predictions returned by ChatGPT are aggregated using fixed thresholds (0.33 or 0.66) to obtain a ranking-



SQuAD 2.0 data sample

```
{
   "answers": {
      "answer_start": [94, 87],
      "text": ["10th and 11th centuries", "in the 10th and 11th
centuries"]
   },
   "context": "\"The Normans (Norman: Nourmands; French:
Normands; Latin: Normanni) were the people who in the 10th
and 11th centuries gave their name to Normandy, a region in
France. They were descended from Norse...",
   "id": "56ddde6b9a695914005b9629",
   "question": "When were the Normans in Normandy?",
   "title": "Normans"
}
```

Extracted training samples

- input: question [SEP] sentence from context
- label: 1 if answer is contained in the sentence or 0 otherwise

- input:
  [
  When were the Normans in Normandy? [SEP] The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. ;
  When were the Normans in Normandy? [SEP] They were descended from Norse...
  ]
- labels: [1, 0]

Figure B.2: An example of extracting answerability labels on sentence level from SQuAD 2.0 samples.

level prediction. In the passage-level answerability prediction, the data is generated only in the zero-shot setting using the following prompt:

- role: `system`
- content: `You are an assistant verifying whether the question is answerable in the provided passage. Return 1 if the answer or partial answer to the question is provided in the passage and 0 otherwise. Return only a number without explanation.`

In the ranking-level setup, we experiment with both a zero-shot setting, where neither examples nor context is given to the model, and a two-shot setting, where two examples (one positive and one negative) containing a question followed by two sentences extracted from the passage are provided. We use the following prompt:

- role: `system`
- content: `You are an assistant verifying whether the answer to the question is included in the provided text. Return 0 if the answer is not given in the text or 1 if the text contains an answer to the question. Return only a number without explanation.`
- role: `user`
- content: `Question: Why does waste compaction slow the biodegradation of organic waste? Text: Introduction. It is illegal to burn household or garden waste at home or in your garden. Burning waste is not only a nuisance to neighbours, it can release many harmful chemicals into the air you breathe.`
- role: assistant
- content: `0`
- role: `user`
- content: `Question: I remember Glasgow hosting COP26 last year, but unfortunately I was out of the loop. What was the conference about? Text: The 2021 United Nations Climate Change Conference, also known as COP26, is the 26th United Nations Climate Change conference. This conference will be the most important intergovernmental meeting on the climate crisis since the Paris agreement was passed in 2015.`

- role: `assistant`
- content: `1`

# Appendix C

# Grounded Response Generation

This appendix contains additional details, results, and analysis for Chapter 6.

## C.1 Pilot Study

We run a pilot to validate the design of the response evaluation study and ensure that the statements related to diversity and specificity capture the dimensions we attempt to measure. We select ten queries from TREC CAsT'20 and '22 that are related to controversial topics or topics not broadly covered in the corpus. For each query, we manually create two pairs of responses. The first pair contains a ground truth response that is based on the top 5 passages according to the relevance judgments and its reformulation generated by GPT-3 and verified manually by the author of this thesis (*rephrasing* pair). The second pair of responses contains a diverse response briefly covering different facets of the topic but missing details about different aspects and a very detailed response that focuses on one aspect of the answer and discusses it in detail at a cost of diversity (*coverage manipulation* pair). Both responses are created manually with the support of GPT-3 in polishing and/or summarizing the text. The first pair of responses verifies whether the study is sensitive to responses of the same quality. Ideally, crowd workers should not present preference towards any of the responses in this pair. The second pair of the responses representing coverage manipulation verifies whether crowd workers associate differences in diversity and specificity with the statements included in the task. It works as a validation of the formulation of response dimensions.

Table C.1: Statements used in 3 different formulations of the pilot study for the human evaluation of breath and depth of the responses. Each statements starts with "The response ...".

| # | **Diversity Question** | **Specificity Question** |
|---|---|---|
| 1 | ... covers diverse information | ... offers detailed information |
| 2 | ... encompasses a broad range of information | ... provides specific details and explanations |
| 3 | ... covers various aspects or perspectives | ... provides in-depth information |

Table C.2: The results of three independent pilot studies for the human evaluation of breath and depth of the responses that differ in the formulation of questions about response diversity and specificity. Rephr. indicates a *rephrasing* pair of responses. Cov. man. indicates a *coverage manipulation* pair of responses.

| # | **Response** | **Diversity votes** | | **Specificity votes** | | **ANOVA** | |
|---|---|---|---|---|---|---|---|
| | **pair variant** | **R1** | **R2** | **R1** | **R2** | **Diversity** | **Specificity** |
| 1 | Rephr. | 19 | 11 | 14 | 16 | 0.039 (S) | 0.613 |
| | Cov. man. | 26 | 4 | 15 | 15 | 0.0 (L) | 1.0 |
| 2 | Rephr. | 25 | 5 | 18 | 12 | 0.0 (L) | 0.125 (S) |
| | Cov. man. | 20 | 10 | 13 | 17 | 0.009 (M) | 0.31 (S) |
| 3 | Rephr. | 19 | 11 | 12 | 18 | 0.039 (S) | 0.125 (S) |
| | Cov. man. | 22 | 8 | 12 | 18 | 0.0 (L) | 0.125 (S) |

Altogether, we release 20 unique tasks in the pilot study and each is completed by 3 unique crowd workers. Crowd workers with a greater than 97% approval rate, over 10,000 approved tasks, and located in the US were qualified to participate in the study. Workers were paid US$ 0.25 for successful task completion. Workers who failed to correctly classify 4 out of 8 aspects or more were rejected. The acceptance rate was around 68%.

We run three independent pilot studies that differ in the formulation of questions about response diversity and specificity (see Table C.1). We observe the smallest difference in votes for different response variants in the rephrasing pair for both diversity and specificity for the first formulation of the pilot study (see Table C.2). In the coverage manipulation response pair we observe the biggest difference in votes between responses for 1 in diversity votes and in 3 for specificity votes. We also perform one-way ANOVA for each response pair variant and question set with response variant (either 1 or 2) as an independent variable and scores for diversity and specificity are dependent variables (1 if the given response variant has been selected and 0 otherwise). The results confirm our previous observations. Namely, for rephrasing the response pair variant we observe the smallest effect of the response variant on diversity and specificity scores for 1. In terms of coverage manipulation, we observe the largest effect of

Table C.3: Evaluation of automatic nugget detection using ROUGE F1.

| **Information nuggets** | **ROUGE F1** |
|---|---|
| Detected by GPT-4 | 0.43 |
| MTurk master workers | 0.54 |
| MTurk regular workers | 0.36 |

Table C.4: Statistics about the number of information nuggets for different inputs and nugget detection methods. *Avg len* indicates the average length of nuggets in characters. "Gold" nuggets are from the CAsT-snippets dataset.

| **Input** | **Nugget** | **Nuggets** | | | |
|---|---|---|---|---|---|
| | **det.** | **Avg** | **Med** | **Max** | **Avg len** |
| relevant | Gold | 11.25 | 10.50 | 22.00 | 165.97 |
| | GPT-4 | 10.61 | 9.00 | 29.00 | 87.55 |
| irrelevant | GPT-4 | 8.00 | 6.00 | 25.00 | 76.45 |
| rephrased | GPT-4 | 10.98 | 10.00 | 31.00 | 105.40 |
| retrieved | GPT-4 | 8.89 | 8.00 | 23.00 | 86.30 |

the response variant on the diversity score of 1 and on the specificity score of 3. Given the results, in the final response evaluation study, we use the statement for diversity from the first question set and the statement for specificity from the third question set.

## C.2 Additional Results

### C.2.1 Automatic Nugget Detection

Automatic nugget detection with GPT-4 is evaluated using similarity to ground-truth annotations from the CAsT-snippets dataset (ROUGE F1) (see Section 5.2). The agreement between ground-truth annotations and snippets detected by GPT-4 is higher than the similarity against reference (expert) annotations reported for two control topics in the original paper for regular MTurk workers (see Table C.3). Even though nugget detection with GPT-4 does not reach the performance of best-performing MTurk master workers, it still presents good performance compared to human annotation variants considered in the CAsT-snippets dataset development (see Chapter 5).

### C.2.2 Nugget Detection and Clustering

We observe that GPT-4 detects a similar number of snippets compared to CAsT-snippets ground truth annotations (see Table C.4). Interestingly, GPT-4 detects a relatively large amount of snippets also in the passages annotated as irrelevant

Table C.5: Statistics about the number of information clusters for different inputs and clustering methods. *Avg len* indicates the average number of nuggets per cluster.

| **Input** | **Nugget det.** | **Clustering** | **Clusters** | | | |
|---|---|---|---|---|---|---|
| | | | **Avg** | **Med** | **Max** | **Avg len** |
| relevant | Gold | BERTopic | 3.48 | 3 | 8 | 3.26 |
| | | LSA | 5.09 | 5 | 10 | 2.22 |
| | GPT-4 | BERTopic | 3.32 | 3 | 10 | 3.21 |
| | | LSA | 4.89 | 4 | 12 | 2.18 |
| irrelevant | GPT-4 | BERTopic | 3.00 | 2 | 10 | 2.67 |
| | | LSA | 3.89 | 3 | 12 | 2.06 |
| rephrased | GPT-4 | BERTopic | 3.16 | 3 | 9 | 3.48 |
| | | LSA | 4.43 | 4 | 11 | 2.48 |
| retrieved | GPT-4 | BERTopic | 2.93 | 3 | 7 | 3.04 |

(based on relevance scores provided in TREC CAsT'20 and '22 datasets). Snippets detected by GPT-4 are significantly shorter than the ground-truth information nuggets from the CAsT-snippets dataset. This implies higher granularity of the automatically detected nuggets and as a consequence information nuggets that may be not self-contained. Insufficient context for detected information may result in hallucinations and unsupported statements in the summarization step.

Regarding clustering, we observe that the CAsT-snippets dataset yields more clusters, indicating greater diversity in the ground-truth annotated information compared to nuggets detected by GPT-4 (see Tables C.5). The number of nuggets per cluster is similar independently of the input and methods used. The number of topics in LSA cannot be statistically determined, so by default, we set the number of clusters to 50% of the information nuggets. Since BERTopic automatically detects the number of clusters, a direct comparison between LSA and BERTopic is not feasible. We observe similar amount of clusters for 5 irrelevant passages compared to other inputs, which can be caused by the fact that nuggets detected in irrelevant passages are likely discussing various topics. The lowest number of clusters is observed for 5 retrieved passages resulting either from a lower number of nuggets compared to other inputs or the fact that the top retrieved passages cover very similar information. The average number of nuggets assigned to one cluster is similar independently of the input type and nugget detection method.

### C.2.3 Final Response Evaluation

Results for additional system variants calculated using RAGAs framework can be found in Table C.6. Completeness scores for additional system variants are

Table C.6: Automatic evaluation of responses using RAGAs framework.

| Input | Method | Faithfulness | Answer rel. |
|---|---|---|---|
| relevant | Baseline | 0.79±0.24 | 0.94±0.04 |
| | GINGER w/ GTnuggets | 0.78±0.25 | 0.87±0.14 |
| | GINGER -fluency w/ GTnugget | 0.81±0.24 | 0.86±0.14 |
| | GINGER -fluency w/ GTnuggets+BM25 | 0.81±0.25 | 0.85±0.14 |
| | GINGER -fluency w/ GTnuggets+LSA | 0.76±0.26 | 0.86±0.14 |
| | GINGER -fluency w/ GTnuggets+BM25+LSA | 0.79±0.28 | 0.86±0.14 |
| | GINGER | 0.69±0.30 | 0.87±0.14 |
| | GINGER -fluency | 0.72±0.29 | 0.86±0.14 |
| | GINGER -fluency w/ BM25 | 0.72±0.29 | 0.86±0.11 |
| | GINGER -fluency w/ LSA | 0.75±0.26 | 0.88±0.05 |
| | GINGER -fluency w/ BM25+LSA | 0.72±0.27 | 0.87±0.06 |
| retrieved | Baseline | 0.71±0.36 | 0.92±0.15 |
| | GINGER | 0.71±0.28 | 0.88±0.06 |
| | GINGER -fluency | 0.70±0.25 | 0.87±0.05 |
| irrelevant | Baseline | 0.48±0.36 | 0.73±0.37 |
| | GINGER | 0.47±0.29 | 0.75±0.27 |
| | GINGER -fluency | 0.56±0.28 | 0.76±0.21 |
| rephrased | Baseline | 0.74±0.26 | 0.87±0.20 |
| | GINGER | 0.75±0.26 | 0.88±0.05 |
| | GINGER -fluency | 0.84±0.24 | 0.86±0.05 |

presented in Table C.7. The results of the additional analysis that compares human judgments with the corresponding automatic metrics are presented in Table C.8. We calculate Kendall's correlation between coherence and answer relevance, correctness and faithfulness, and sufficiency and faithfulness. The correlation is calculated on the response level. Response dimensions evaluated in the human study are not strongly correlated with the automatic measures that we use. It means that automatic and human evaluation of the response are complementary.

## C.3 Prompts

This section presents prompts used by different components of the system.

### C.3.1 GINGER

Prompt used for nugget detection:

```
Given a query and a passage, annotate information nuggets that contain
the key information answering the query.  Copy the text of the passage
```

Table C.7: Response completeness scores in terms of how many ground-truth nuggets (from CAsT-snippets dataset) are entailed by the final response.

| Input | Method | Completeness |
|---|---|---|
| relevant | Baseline | 0.25 |
| | GINGER w/ GTnuggets | 0.46 |
| | CAsT-snippets/BERTopic/DuoT5/GPT-4 | 0.48 |
| | GINGER -fluency w/ GTnuggets+BM25 | 0.45 |
| | GINGER -fluency w/ GTnuggets+LSA | 0.44 |
| | CAsT-snippets/LSA/BM25/GPT-4 | 0.44 |
| | GINGER | 0.29 |
| | GINGER -fluency | 0.28 |
| | GPT-4/BERTopic/BM25/GPT-4 | 0.30 |
| | GINGER -fluency w/ LSA | 0.29 |
| | GINGER -fluency w/ BM25+LSA | 0.29 |
| | GINGER -fluency (deep) | 0.17 |
| | GINGER -fluency (broad) | 0.31 |
| irrelevant | Baseline | 0.07 |
| | GINGER | 0.03 |
| | GINGER -fluency | 0.04 |
| retrieved | Baseline | 0.17 |
| | GINGER | 0.13 |
| | GINGER -fluency | 0.16 |
| rephrased | Baseline | 0.20 |
| | GINGER | 0.24 |
| | GINGER -fluency | 0.25 |
| | GINGER -fluency w/ LSA | 0.27 |

```
and put the annotated information nuggets between <IN> and </IN>.
Do NOT modify the content of the passage.  Do NOT add additional
symbols, spaces, etc.  to the text.  Question:  query Passage:  passage
```

Prompt used for information cluster summarization:

```
Summarize the provided information into one sentence (approximately
35 words).  Generate a one-sentence long summary that is short,
concise and only contains the information provided.  text
```

Prompt used for improving response fluency:

```
Rephrase the response given a query.  Do not change the information
included in the response.  Do not add information not mentioned
in the response.  Query:  query Response:  response
```

Prompt used for response generation from one cluster:

Table C.8: Kendall's correlation between the human scores and corresponding automatic measures. Answer relevance and faithfulness are computed using the RAGAs framework. Nugget entailment and contradiction are computed using NLI. Completeness is the fraction of ground-truth information nuggets included in the response.

| **Human score** | **Automatic metric** | **Correlation** |
|---|---|---|
| Coherence | Answer relevance | 0.17* |
| Correctness | Faithfulness | -0.07 |
| Correctness | Nugget entailment | -0.00 |
| Correctness | Nugget contradiction | -0.11 |
| Sufficiency | Completeness | -0.01 |
| Breadth | Completeness | 0.08 |
| Depth | Completeness | 0.13 |

```
Generate the answer to a query that is 3 sentences long (approximately 100 words in total) using the provided information. Use only the provided information. You can expand the provided information but do not add any additional information. text
```

## C.3.2 Baselines

Prompt used by zero-shot one-step response generation baseline:

```
Generate the answer to a query that is 3 sentences long (approximately 100 words in total) using the provided information. Use only the provided information and do not add any additional information.
Question: query Passage: passages
```

Prompt used by Chain-of-Thought two-shot response generation baseline:

```
TASK

You are an assistant generating responses to user questions based on the provided information. Your response should rely on the context passage and it should not incorporate any additional information.

SPECIFIC STEPS

Follow a structured step-by-step approach to ensure relevance and coherence in the generated response:
- Step 1: extract key pieces of information relevant to the question from the provided passage
- Step 2: group related pieces if information based on the aspect of the topic they discuss or the point of view they represent
- Step 3: rank groups of information based on their relevance to the query
```

- Step 4: use the top groups of relevant information to write a coherent and concise response

The final response should be three sentences long (approximately 100 words).

EXAMPLE

Question: Tell me more about the Blue Lives Matter movement.

Passage: [passage]The internet facilitates the spread of the message 'All Lives Matter' as a response to the Black Lives Matter hashtag as well as the 'Blue Lives Matter' hashtag as a response to Beyonce's halftime performance speaking out against police brutality. Following the shooting of two police officers in Ferguson and in response to BLM, the hashtag [Blue Lives Matter or hashtag BlueLivesMatter] was created by supporters of the police. Following this, Blue Lives Matter became a pro-police movement in the United States. It expanded after the killings of American police officers. On December 20, 2014, in the wake of the killings of officers Rafael Ramos and Wenjian Liu, a group of law enforcement officers formed Blue Lives Matter to counter media reports that they perceived to be anti-police. Blue Lives Matter is made up of active and retired law enforcement officers. The current national spokesman for Blue Lives Matter is retired Las Vegas Metropolitan Police Department Lieutenant Randy Sutton. Originating in New York City in December 2014, Blue Lives Matter NYC is an organization and current nationwide movement that was created to help law enforcement officers and their families during their times of need. Sergeant Joey Imperatrice, along with Officers Chris Brinkley and Carlos Delgado, felt 'compelled to show support for their brothers and sisters in blue' and handed out bracelets that stated 'Blue Lives Matter'. They decided to create an organization, which became the non-profit Blue Lives Matter NYC. This organization's mission is to raise awareness and encourage public aid for the needs of police officers, to help police officers assist one another, and to provide a police officer's family with comfort and support during difficult times. This campaign is designed to raise both awareness and money for families in need. In order to increase nationwide awareness, over three hundred billboards have been posted with the slogan 'Blue Lives Matter'. Many of these billboards are also accompanied by the hashtag hashtag thankublu, which individuals use to show their support for police officers. Blue Lives Matter is a pro-police movement in the United States. It was started after the killings of NYPD officers Rafael Ramos and Wenjian Liu in Brooklyn, New York, on December 20, 2014, after they were ambushed in their patrol car. Blue Lives Matter was formed in reaction to the Black Lives Matter movement, which seeks to end police brutality against the African American community.[/passage]

Step-by-step response generation process:

```
Step 1: Extract key pieces of information relevant to the query
'Blue Lives Matter' hashtag as a response to Beyonce's halftime
performance speaking out against police brutality
[Blue Lives Matter or hashtag BlueLivesMatter] was created by supporters
of the police
Blue Lives Matter became a pro-police movement in the United States
a group of law enforcement officers formed Blue Lives Matter to
counter media reports that they perceived to be anti-police
made up of active and retired law enforcement officers
help law enforcement officers and their families during their times
of need
raise awareness and encourage public aid for the needs of police
officers, to help police officers assist one another, and to provide
a police officer's family with comfort and support as they go through
hard times raise both awareness and money for the families in need
over three hundred billboards have been posted with the slogan 'Blue
Lives Matter'
hashtag thankublu, a hashtag many individuals use to show their
support for police officers
It was started after the killings of NYPD officers Rafael Ramos
and Wenjian Liu in Brooklyn, New York, on December 20, 2014, after
they were ambushed in their patrol car
which seeks to end police brutality against the African American
community

Step 2: Group related pieces of information
Group A - Purpose & Support for Officers:
help law enforcement officers and their families during their times
of need
raise awareness and encourage public aid for the needs of police
officers, to help police officers assist one another, and to provide
a police officer's family with comfort and support as they go through
hard times
raise both awareness and money for the families in need
Group B - Origin & Formation:
'Blue Lives Matter became a pro-police movement in the United States
a group of law enforcement officers formed Blue Lives Matter to
counter media reports that they perceived to be anti-police
made up of active and retired law enforcement officers
It was started after the killings of NYPD officers Rafael Ramos
and Wenjian Liu in Brooklyn, New York, on December 20, 2014, after
they were ambushed in their patrol car
which seeks to end police brutality against the African American
community
Group C- Broader Context & Media Reaction:
'Blue Lives Matter' hashtag as a response to Beyonce's halftime
performance speaking out against police brutality
[Blue Lives Matter or hashtag BlueLivesMatter] was created by supporters
of the police
```

hashtag thankublu, a hashtag many individuals use to show their support for police officers
over three hundred billboards have been posted with the slogan 'Blue Lives Matter'

Step 3: Rank groups based on relevance to the query
Group B - Origin & Formation (Most relevant)
Group A - Purpose & Support for Officers
Group C - Broader Context & Media Reaction

Step 4: Generate a coherent, concise response
Response: Blue Lives Matter was founded by active and retired law enforcement officers following the targeted killings of NYPD officers Rafael Ramos and Wenjian Liu on December 20, 2014. It emerged as a pro-police movement aimed at countering what it perceived as anti-police media narratives while supporting officers and their families through fundraising and awareness campaigns. The movement also engages in public outreach through billboards, social media hashtags like hashtag thankublu, and nonprofit initiatives dedicated to aiding law enforcement personnel in times of need.

NOW PERFORM THE TASK ON THE FOLLOWING INPUT

# Bibliography

Adams, G., Fabbri, A., Ladhak, F., Lehman, E., and Elhadad, N. (2023). From sparse to dense: GPT-4 summarization with chain of density prompting. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, ACL '23, pages 68–74.

Alaofi, M., Gallagher, L., Mckay, D., Saling, L. L., Sanderson, M., Scholer, F., Spina, D., and White, R. W. (2022). Where do queries come from? In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 2850–2862.

Aliannejadi, M., Zamani, H., Crestani, F., and Croft, W. B. (2019). Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 475–484.

Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H. R., Bertino, E., and Dustdar, S. (2013). Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing*, **17**(2), 76–81.

Allein, L., Augenstein, I., and Moens, M.-F. (2021). Time-aware evidence ranking for fact-checking. *Journal of Web Semantics*, **71**, 100663.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., and Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–13.

Anand, A., Cavedon, L., Hagen, M., Joho, H., Sanderson, M., and Stein, B. (2019). Conversational search (Dagstuhl seminar 19461). *Dagstuhl Reports*, **9**(11).

Anand, A., Cavedon, L., Hagen, M., Joho, H., Sanderson, M., and Stein, B. (2021). Dagstuhl seminar 19461 on conversational search: seminar goals and working group outcomes. *SIGIR Forum*, **54**(1).

Anantha, R., Vakulenko, S., Tu, Z., Longpre, S., Pulman, S., and Chappidi, S. (2021). Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '21, pages 520–534.

Armstrong, T. G., Moffat, A., Webber, W., and Zobel, J. (2009). Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 601–610.

Asai, A. and Choi, E. (2021). Challenges in information-seeking QA: Unanswerable questions and paragraph retrieval. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, ACL-IJNLP '21, pages 1492–1504.

Azzopardi, L. (2021). Cognitive biases in search: A review and reflection of cognitive biases in information retrieval. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, CHIIR '21, pages 27–37.

Azzopardi, L., Dubiel, M., Halvey, M., and Dalton, J. (2018). Conceptualizing agent-human interactions during the conversational search process. In *2nd International ACM SIGIR Workshop Conference on Conversational Approaches to IR*, CAIR '18.

Baeza-Yates, R. (2018). Bias on the web. *Commun. ACM*, **61**(6), 54–61.

Baheti, A., Ritter, A., and Small, K. (2020). Fluent response generation for conversational question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 191–207.

Bai, Y., Miao, Y., Chen, L., Li, D., Ren, Y., Xie, H., Yang, C., and Cai, X. (2024). Pistis-RAG: A scalable cascading framework towards trustworthy retrieval-augmented generation.

Balog, K. and Kenter, T. (2019). Personal knowledge graphs: A research agenda. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, pages 217–220.

Balog, K., Metzler, D., and Qin, Z. (2025). Rankers, judges, and assistants: Towards understanding the interplay of llms in information retrieval evaluation. *arXiv*, **cs.IR/2503.19092**.

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, **13**(5), 407–424.

Belkin, N. (1995). Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications*, **9**(3), 379–395.

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian journal of information science*, **5**(1), 133–143.

Bernard, N., Kostric, I., Łajewska, W., Balog, K., Galusčáková, P., Setty, V., and Skjæveland, M. G. (2024). PKG API: A tool for personal knowledge graph management. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, pages 1051–1054.

Bhandari, M., Gour, P. N., Ashfaq, A., Liu, P., and Neubig, G. (2020). Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '2020, pages 9347–9359.

Bink, M., Zimmerman, S., and Elsweiler, D. (2022). Featured snippets and their influence on users' credibility judgements. In *Proceedings of the 2022 Conference on Human Information Interaction & Retrieval*, CHIIR '22, pages 113–122.

Bohnet, B., Tran, V. Q., Verga, P., Aharoni, R., Andor, D., Soares, L. B., Ciaramita, M., Eisenstein, J., Ganchev, K., Herzig, J., Hui, K., Kwiatkowski, T., Ma, J., Ni, J., Saralegui, L. S., Schuster, T., Cohen, W. W., Collins, M., Das, D., Metzler, D., Petrov, S., and Webster, K. (2023). Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv*, **cs.CL/2212.08037**.

Bolotova, V., Blinov, V., Zheng, Y., Croft, W. B., Scholer, F., and Sanderson, M. (2020). Do people and neural nets pay attention to the same words: Studying eye-tracking data for non-factoid QA evaluation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pages 85–94.

Bolotova-Baranova, V., Blinov, V., Filippova, S., Scholer, F., and Sanderson, M. (2023). WikiHowQA: A comprehensive benchmark for multi-document non-factoid question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '23, pages 5291–5314.

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, **54**(10), 913–925.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, pages 632–642.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20.

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. (2018). MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 5016–5026.

Cambazoglu, B. B., Baranova, V., Scholer, F., Sanderson, M., Tavakoli, L., and Croft, B. (2021). Quantifying human-perceived answer utility in non-factoid question answering. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, CHIIR '21, pages 75–84.

Campos, D. F., Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L., and Mitra, B. (2016). MS MARCO: A human generated machine reading comprehension dataset. *arXiv*, **cs.CL/1611.09268**.

Campos, R., Dias, G., Jorge, A. M., and Jatowt, A. (2015). Survey of temporal information retrieval and related applications. *ACM Computing Surveys*, **47**(2), 1–41.

Cao, Z., Li, W., Li, S., Wei, F., and Li, Y. (2016). AttSum: Joint learning of focusing and summarization with neural attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, COLING '16, pages 547–556.

Cau, F. M., Hauptmann, H., Spano, L. D., and Tintarev, N. (2023). Supporting high-uncertainty decisions through AI and logic-style explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, pages 251–263.

Chang, C.-Y., Chen, N., Chiang, W.-T., Lee, C.-H., Tseng, Y.-H., Wang, C.-J., Chen, H.-H., and Tsai, M.-F. (2020). Query expansion with semantic-based ellipsis reduction for conversational IR. In *The Tweenty-Ninth Text REtrieval Conference Proceedings*, TREC '20.

Chaudhry, A., Thiagarajan, S., and Gorur, D. (2024). Finetuning language models to emit linguistic expressions of uncertainty. *arXiv*, **cs.CL/2409.12180**.

Chen, H., Liu, X., Yin, D., and Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, **19**(2), 25–35.

Chen, V., Liao, Q. V., Wortman Vaughan, J., and Bansal, G. (2023). Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-Computer Interaction*, **7**(CSCW2).

Chen, X., Chen, F., Meng, F., Li, P., and Zhou, J. (2021). Unsupervised knowledge selection for dialogue generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, ACL-IJNLP '21, pages 1230–1244.

Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., and Zhu, H. (2019). Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12.

Chiang, C.-W. and Yin, M. (2022). Exploring the effects of machine learning literacy interventions on laypeople's reliance on machine learning models. In *Proceeding of the 27th International Conference on Intelligent User Interfaces*, IUI '22, pages 148–161.

Chien, A. A., Lin, L., Nguyen, H., Rao, V., Sharma, T., and Wijayawardana, R. (2023). Reducing the carbon impact of generative AI inference (today and in 2035). In *Proceedings of the 2nd Workshop on Sustainable Computer Systems*, HotCarbon '23, pages 1–7.

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., Liang, P., and Zettlemoyer, L. (2018). QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18.

Chomsky, N. (1957). *Syntactic Structures*. Martino Publishing.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. (2023). PaLM: scaling language modeling with pathways. *The Journal of Machine Learning Research*, **24**(1).

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4302–4310.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16**(1), 22–29.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., and Smith, N. A. (2021). All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, ACL-IJNLP '21, pages 7282–7296.

Cormack, G. V., Clarke, C. L. A., and Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 758–759.

Costa, F., Ouyang, S., Dolog, P., and Lawlor, A. (2018). Automatic generation of natural language explanations. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, IUI '18, pages 1–2.

Craswell, N., Mitra, B., Yilmaz, E., and Campos, D. (2020). Overview of the TREC 2020 deep learning track. In *The Twenty-Ninth Text REtrieval Conference Proceedings*, TREC '20.

Cuconasu, F., Trappolini, G., Siciliano, F., Filice, S., Campagnano, C., Maarek, Y., Tonellotto, N., and Silvestri, F. (2024). The power of noise: Redefining retrieval for RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24.

Culpepper, J. S., Diaz, F., and Smucker, M. D. (2018). Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in Lorne (SWIRL 2018). *SIGIR Forum*, **52**(1), 34–90.

Culpepper, J. S., Faggioli, G., Ferro, N., and Kurland, O. (2022). Topic difficulty: Collection and query formulation effects. *ACM Transactions on Information Systems*, **40**(1), 1–36.

Dalton, J., Xiong, C., and Callan, J. (2019). TREC CAsT 2019: The conversational assistance track overview. In *The Twenty-Eighth Text REtrieval Conference Proceedings*, TREC '19.

Dalton, J., Xiong, C., and Callan, J. (2020). CAsT 2020: The conversational assistance track overview. In *The Twenty-Ninth Text REtrieval Conference Proceedings*, TREC '20.

Dalton, J., Xiong, C., and Callan, J. (2021). TREC CAsT 2021: The conversational assistance track overview. In *The Thirtieth Text REtrieval Conference Proceedings*, TREC '21.

Dang, H. T. and Lin, J. (2007). Different structures for evaluating answers to complex questions: Pyramids won't topple, and neither will human assessors. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, ACL '07, pages 768–775.

Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., and Allahbakhsh, M. (2019). Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys*, **51**(1), 1–40.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, NAACL-HLT '19, pages 4171–4186.

Dietz, L. (2024). A workbench for autograding retrieve/generate systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 1963–1972.

Dietz, L., Verma, M., Radlinski, F., and Craswell, N. (2018). TREC complex answer retrieval overview. In *The Twenty-Seventh Text REtrieval Conference Proceedings*, TREC '18.

Draws, T., Tintarev, N., and Gadiraju, U. (2021a). Assessing viewpoint diversity in search results using ranking fairness metrics. *ACM SIGKDD Explorations Newsletter*, **23**(1), 50–58.

Draws, T., Tintarev, N., Gadiraju, U., Bozzon, A., and Timmermans, B. (2021b). This is not what we ordered: Exploring why biased search result rankings affect user attitudes on debated topics. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21.

Draws, T., Inel, O., Tintarev, N., Baden, C., and Timmermans, B. (2022). Comprehensive viewpoint representations for a deeper understanding of user interactions with debated topics. In *Proceedings of the 2022 Conference on Human Information Interaction & Retrieval*, CHIIR '22, pages 135–145.

Dubiel, M., Halvey, M., Azzopardi, L., Anderson, D., and Daronnat, S. (2020). Conversational strategies: Impact on search performance in a goal-oriented task. In *ACM CHIIR 3rd Conversational Approaches to Information Retrieval Workshop (CAIR)*, CAIR '20.

Dumais, S., Jeffries, R., Russell, D. M., Tang, D., and Teevan, J. (2014). *Understanding User Behavior Through Log Data and Analysis*. Springer.

Dziri, N., Kamalloo, E., Mathewson, K., and Zaiane, O. (2019). Augmenting neural response generation with context-aware topical attention. In *Proceedings of the First Workshop on NLP for Conversational AI*, NLPConAI '19, pages 18–31.

Eickhoff, C. (2018). Cognitive biases in crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 162–170.

Ekstrand, M. D., Das, A., Burke, R., and Diaz, F. (2022). Fairness in information access systems. *Foundations and Trends® in Information Retrieval*, **16**(1-2), 1–177.

Elgohary, A., Peskov, D., and Boyd-Graber, J. (2019). Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, EMNLP-IJCNLP '19, pages 5918–5924.

Es, S., James, J., Espinosa Anke, L., and Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, EACL '24, pages 150–158.

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D. (2021). SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, **9**, 391–409.

Faggioli, G., Dietz, L., Clarke, C. L. A., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., and Wachsmuth, H. (2023). Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '23, pages 39–50.

Falke, T., Ribeiro, L. F. R., Utama, P. A., Dagan, I., and Gurevych, I. (2019). Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, pages 2214–2220.

Farzi, N. and Dietz, L. (2024a). EXAM++: LLM-based answerability metrics for IR evaluation. In *LLM4Eval: The First Workshop on Large Language Models for Evaluation in Information Retrieval*, SIGIR '24.

Farzi, N. and Dietz, L. (2024b). TREMA-UNH at TREC: RAG systems and RUBRIC-style evaluation. In *The Thirty-Third Text REtrieval Conference Proceedings*, TREC '24.

Ferreira, R., Leite, M., Semedo, D., and Magalhaes, J. (2022). Open-domain conversational search assistants: the transformer is all you need. *Information Retrieval*, **25**(2), 123–148.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis, Special Volume/Blackwell*, (1952-59), 1–32.

Fisher, R. A. (1992). *Statistical Methods for Research Workers. Breakthroughs in Statistics: Methodology and Distribution*. Springer New York.

Fröbe, M., Gienapp, L., Scells, H., Schmidt, E. O., Wiegmann, M., Potthast, M., and Hagen, M. (2024). Webis at TREC 2024: Biomedical generative retrieval, retrieval-augmented generation, and tip-of-the-tongue tracks. In *The Thirty-Third Text REtrieval Conference Proceedings*, TREC '24.

Gabburo, M., Jedema, N. P., Garg, S., Ribeiro, L. F. R., and Moschitti, A. (2024). Measuring retrieval complexity in question answering systems. In *Findings of the Association for Computational Linguistics: ACL 2024*, ACL '24, pages 14636–14650.

Gadiraju, U., Demartini, G., Kawase, R., and Dietze, S. (2015). Human beyond the machine: Challenges and opportunities of microtask crowdsourcing. *IEEE Intelligent Systems*, **30**(4), 81–85.

Gao, J., Galley, M., and Li, L. (2019). Neural approaches to conversational AI. *Found. Trends Inf. Retr.*, **13**(2-3), 127–298.

Gao, J., Xiong, C., Bennett, P., and Craswell, N. (2023a). *Neural Approaches to Conversational Information Retrieval*. Springer Cham.

Gao, R. and Shah, C. (2020). Toward creating a fairer ranking in search engine results. *Information Processing & Management*, **57**(1), 102–138.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. (2023b). Retrieval-augmented generation for large language models: A survey. **cs.CL/2312.10997**.

Gemmell, C. and Dalton, J. (2020). Glasgow representation and information learning lab (GRILL) at the conversational assistance track 2020. In *The Twenty-Ninth Text REtrieval Conference Proceedings*, TREC '20.

Gerritse, E. J., Hasibi, F., and de Vries, A. P. (2020). Bias in conversational search: The double-edged sword of the personalized knowledge graph. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, ICTIR '20, pages 133–136.

Gienapp, L., Scells, H., Deckers, N., Bevendorff, J., Wang, S., Kiesel, J., Syed, S., Fröbe, M., Zuccon, G., Stein, B., Hagen, M., and Potthast, M. (2024). Evaluating generative ad hoc information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 1916–1929.

Godin, F., Kumar, A., and Mittal, A. (2019). Learning when not to answer: a ternary reward structure for reinforcement learning based question answering. In *North American Chapter of the Association for Computational Linguistics*, NAACL '19, pages 122–129.

Gospodinov, M., MacAvaney, S., and Macdonald, C. (2023). Doc2Query–: When less is more. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval*, ECIR '23, pages 414–422.

Goyal, T. and Durrett, G. (2021). Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '21, pages 1449–1462.

Goyal, T., Li, J. J., and Durrett, G. (2023). News summarization and evaluation in the era of GPT-3. *arXiv*, **cs.CL/2209.12356**.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv*, **cs.CL/2203.05794**.

Guo, J., Fan, Y., Ai, Q., and Croft, W. B. (2016). A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 55–64.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. (2020). REALM: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML '20*, pages 3929–3938.

Harris, Z. S. (1954). *Distributional Structure. Papers on Syntax*. Springer.

He, G., Buijsman, S., and Gadiraju, U. (2023a). How stated accuracy of an AI system and analogies to explain accuracy affect human reliance on the system. *Proceedings of the ACM on Human-Computer Interaction*, **7**(CSCW2), 1–29.

He, G., Kuiper, L., and Gadiraju, U. (2023b). Knowing about knowing: An illusion of human competence can hinder appropriate reliance on AI systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–18.

Helberger, N., Karppinen, K., and D'Acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, **21**(2), 191–207.

Hersh, W., Turpin, A., Price, S., Chan, B., Kramer, D., Sacherek, L., and Olson, D. (2000). Do batch and user evaluations give the same results? In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 17–24.

Hewitt, J. and Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, NAACL-HLT '19, pages 4129–4138.

Horvitz, E. (1999). Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, CHI '99, pages 159–166.

Hu, M., Wei, F., Peng, Y., Huang, Z., Yang, N., and Li, D. (2019). Read + verify: Machine reading comprehension with unanswerable questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI '19, pages 6529–6537.

Huang, K., Tang, Y., Huang, J., He, X., and Zhou, B. (2019a). Relation module for non-answerable predictions on reading comprehension. In *Proceedings of the 23rd Conference on Computational Natural Language Learning*, CoNLL '19, pages 747–756.

Huang, K., Tang, Y., Huang, J., He, X., and Zhou, B. (2019b). Relation module for non-answerable predictions on reading comprehension. In *Proceedings of the 23rd Conference on Computational Natural Language Learning*, CoNLL '19, pages 747–756.

Huang, P.-S., He, X., Gao, J., Deng, L., Acero, A., and Heck, L. (2013). Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, CIKM '13, pages 2333–2338.

Huang, Y. and Huang, J. (2024). A survey on retrieval-augmented text generation for large language models. *arXiv*, **cs.IR/2404.10981**.

Iskender, N., Schaefer, R., Polzehl, T., and Möller, S. (2021). Argument mining in tweets: Comparing crowd and expert annotations for automated claim and evidence detection. In *Natural Language Processing and Information Systems: 26th International Conference on Applications of Natural Language to Information Systems*, NLDB '21, pages 275–288.

Izacard, G. and Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, EACL '21, pages 874–880.

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., and Riedel, S. (2023). Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, (24), 1–43.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, **20**(4), 422–446.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, **55**(12), 1–38.

Jiang, H., Wu, Q., Luo, X., Li, D., Lin, C.-Y., Yang, Y., and Qiu, L. (2024). LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '24, pages 1658–1677.

Jiang, Z., Dou, Z., Zhao, W. X., Nie, J.-Y., Yue, M., and Wen, J.-R. (2018). Supervised search result diversification via subtopic attention. *IEEE Transactions on Knowledge and Data Engineering*, **30**(10), 1971–1984.

Ju, J.-H., Yeh, C.-T., Lin, C.-W., Tsao, C.-Y., Ding, J.-E., Wang, C.-J., and Tsai, M.-F. (2021). An exploration study of multi-stage conversational passage retrieval: Paraphrase query expansion and multi-view point-wise ranking. In *The Thirtieth Text REtrieval Conference Proceedings*, TREC '21.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv*, **cs.LG/2001.08361**.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 6769–6781.

Kazai, G., Kamps, J., and Milic-Frayling, N. (2011). Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 1941–1944.

Kazai, G., Mitra, B., Dong, A., Craswell, N., and Yang, L. (2022). Less is less: When are snippets insufficient for human vs machine relevance estimation? In *Advances in Information Retrieval: 44th European Conference on IR Research*, ECIR '22, pages 153–162.

Kekäläinen, J. and Järvelin, K. (2002). Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, **53**(13), 1120–1129.

Kelly, D. (2007). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval*, **3**(1—2), 1–224.

Kim, Y. and Allan, J. (2019). Unsupervised explainable controversy detection from online news. In *Advances in Information Retrieval: 41th European Conference on IR Research*, ECIR '19, pages 836–843.

Koch, T. K., Frischlich, L., and Lermer, E. (2023). Effects of fact-checking warning labels and social endorsement cues on climate change fake news credibility and engagement on social media. *Journal of Applied Social Psychology*, **53**(6), 495–507.

Kocielnik, R., Amershi, S., and Bennett, P. N. (2019). Will you accept an imperfect AI? exploring designs for adjusting end-user expectations of AI systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–14.

Koopman, B. and Zuccon, G. (2023). Dr ChatGPT tell me what i want to hear: How different prompts impact health answer correctness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, EMNLP '23, pages 15012–15022.

Kostric, I., Balog, K., Aresvik, T. A., Bernard, N., Dørheim, E. T., Hantula, P., Havn-Sørensen, S., Henriksen, R., Hosseini, H., Khlybova, E., Lajewska, W., Mosand, S. E., and Orujova, N. (2022). DAGFiNN: A conversational conference assistant. In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys '22, pages 628–631.

Krishna, K., Roy, A., and Iyyer, M. (2021). Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '21, pages 4940–4957.

Kryscinski, W., McCann, B., Xiong, C., and Socher, R. (2020). Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 9332–9346.

Kumar, V. and Callan, J. (2020). Making information seeking easier: An improved pipeline for conversational search. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, EMNLP '20, pages 3971–3980.

Kurland, O. and Culpepper, J. S. (2018). Fusion in information retrieval: SIGIR 2018 half-day tutorial. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 1383–1386.

Ladhak, F., Durmus, E., He, H., Cardie, C., and McKeown, K. (2022). Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '22, pages 1410–1421.

Lafferty, J. and Zhai, C. (2003). *Probabilistic Relevance Models Based on Document and Query Generation. Language Modeling for Information Retrieval*. Springer.

Łajewska, W. (2024). Grounded and transparent response generation for conversational information-seeking systems. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, pages 1142–1144.

Łajewska, W. and Balog, K. (2023a). From baseline to top performer: A reproducibility study of approaches at the TREC 2021 conversational assistance track. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval*, ECIR '23, page 177–191.

Łajewska, W. and Balog, K. (2023b). Towards filling the gap in conversational search: From passage retrieval to conversational response generation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, pages 5326–5330.

Łajewska, W. and Balog, K. (2024a). Towards reliable and factual response generation: Detecting unanswerable questions in information-seeking conversations. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval*, ECIR '24, page 336–344.

Łajewska, W. and Balog, K. (2024b). The University of Stavanger (IAI) at the TREC 2024 retrieval-augmented generation track. In *The Thirty-Third Text REtrieval Conference Proceedings*, TREC '24.

Łajewska, W. and Balog, K. (2025). GINGER: Grounded information nugget-based generation of responses. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25.

Łajewska, W., Bernard, N., Kostric, I., Sekulic, I., and Balog, K. (2022). The University of Stavanger (IAI) at the TREC 2022 conversational assistance track. In *The Thirty-First Text REtrieval Conference Proceedings*, TREC '22.

Łajewska, W., Balog, K., Spina, D., and Trippas, J. (2024a). Can users detect biases or factual errors in generated responses in conversational information-seeking? In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP '24, pages 92–102.

Łajewska, W., Spina, D., Trippas, J., and Balog, K. (2024b). Explainability for transparent conversational information-seeking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 1040–1050.

Łajewska, W., Hardalov, M., Aina, L., John, N. A., Su, H., and Màrquez, L. (2025). Understanding and improving information preservation in prompt compression for LLMs. *arXiv*, **cs.CL/2503.19114**.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6405–6416.

Lavrenko, V. and Croft, W. B. (2001). Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 120–127.

Le, J., Edmonds, A., Hester, V., and Biewald, L. (2010). Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, SIGIR '10.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 9459–9474.

Liao, J., Zhao, X., Zheng, J., Li, X., Cai, F., and Tang, J. (2022). PTAU: Prompt tuning for attributing unanswerable questions. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 1219–1229.

Liao, Q. and Sundar, S. S. (2022). Designing for responsible trust in AI systems: A communication perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1257–1268.

Liao, Q. V. and Vaughan, J. W. (2024). AI transparency in the age of LLMs: A human-centered research roadmap. *Harvard Data Science Review*, (5).

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004*, ACL '04, pages 74–81.

Lin, S.-C., Yang, J.-H., Nogueira, R., Tsai, M.-F., Wang, C.-J., and Lin, J. (2021). Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting. *ACM Transactions on Information Systems*, **39**(4).

Lippe, P., Ren, P., Haned, H., Voorn, B., and de Rijke, M. (2020). Diversifying task-oriented dialogue response generation with prototype guided paraphrasing. *arXiv*, **cs.CL/2008.03391**.

Liu, J. (2023). Toward a two-sided fairness framework in search and recommendation. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, CHIIR '23, pages 236–246.

Liu, N., Zhang, T., and Liang, P. (2023a). Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, EMNLP '23, pages 7001–7025.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, **12**, 157–173.

Liu, T.-Y. (2010). Learning to rank for information retrieval. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, page 904.

Liu, Y., Fabbri, A., Liu, P., Zhao, Y., Nan, L., Han, R., Han, S., Joty, S., Wu, C.-S., Xiong, C., and Radev, D. (2023b). Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '23, pages 4140–4170.

Lu, Z. and Yin, M. (2021). Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21.

Luan, Y., Eisenstein, J., Toutanova, K., and Collins, M. (2021). Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, **9**.

MacAvaney, S. and Soldaini, L. (2023). One-shot labeling for automatic relevance estimation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pages 2230–2235.

MacCartney, B. and Manning, C. D. (2008). Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics*, COLING '08, pages 521–528.

Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM*, **49**(4), 41–46.

McDonald, R., Brokos, G., and Androutsopoulos, I. (2018). Deep relevance ranking using enhanced document-query interactions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 1849–1860.

Mele, I., Muntean, C. I., Nardini, F. M., Perego, R., Tonellotto, N., and Frieder, O. (2020). Topic propagation in conversational search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 2057–2060.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. **cs.CL/1301.3781**.

Mombaerts, L., Ding, T., Banerjee, A., Felice, F., Taws, J., and Borogovac, T. (2024). Meta knowledge for retrieval augmented large language models. **cs.IR/2408.09017**.

Monroe, D. (2018). Ai, explain yourself. *Communications of the ACM*, **61**(11), 11–13.

Muhlgay, D., Ram, O., Magar, I., Levine, Y., Ratner, N., Belinkov, Y., Abend, O., Leyton-Brown, K., Shashua, A., and Shoham, Y. (2023). Generating benchmarks for factuality evaluation of language models. In *Conference of the European Chapter of the Association for Computational Linguistics*, EACL '23.

Nalisnick, E., Mitra, B., Craswell, N., and Caruana, R. (2016). Improving document ranking with dual word embeddings. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16, pages 83–84.

Nenkova, A. and Passonneau, R. (2004). Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL-HLT '04, pages 145–152.

Nenkova, A., Passonneau, R., and McKeown, K. (2007). The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, **4**(2).

Niehues, J. and Pham, N.-Q. (2019). Modeling confidence in sequence-to-sequence models. In *Proceedings of the 12th International Conference on Natural Language Generation*, INLG '19, pages 575–583.

Nogueira, R. and Cho, K. (2019). Passage re-ranking with BERT. *arXiv*, **cs.IR/1901.04085**.

Nogueira, R., Yang, W., Cho, K., and Lin, J. J. (2019). Multi-stage document ranking with BERT. *arXiv*, **cs.IR/1910.14424**.

Nogueira, R., Jiang, Z., Pradeep, R., and Lin, J. (2020). Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, EMNLP '20, pages 708–718.

Nunes, I. and Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, **27**(3-5).

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, Ł., Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Kondraciuk, Ł., Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., Peres, F. d. A. B., Petrov, M., Pinto, H. P. d. O., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C. J., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. (2024). GPT-4 technical report. *arXiv*, **cs.CL/2303.08774**.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Proceedings of the 33rd International Conference on Neural Information*

*Processing Systems*, NIPS '19.

Owoicho, P., Dalton, J., Aliannejadi, M., Azzopardi, L., Trippas, J. R., and Vakulenko, S. (2022). TREC CAsT 2022: Going beyond user ask and system retrieve with initiative and response generation. In *The Thirty-First Text REtrieval Conference Proceedings*, TREC '22.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311.

Pathiyan Cherumanal, S., Tian, L., Abushaqra, F. M., Magnossão De Paula, A. F., Ji, K., Ali, H., Hettiachchi, D., Trippas, J. R., Scholer, F., and Spina, D. (2024). Walert: Putting conversational information seeking knowledge into action by building and evaluating a large language model-powered chatbot. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '24, pages 401–405.

Pavlu, V., Rajput, S., Golbus, P. B., and Aslam, J. A. (2012). IR system evaluation using nugget-based test collections. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, page 393–402.

Pei, J., Ren, P., Monz, C., and de Rijke, M. (2020). Retrospective and prospective mixture-of-generators for task-oriented dialogue response generation. In *European Conference on Artificial Intelligence*, ECAI '20, pages 2148–2155.

Peng, B., Galley, M., He, P., Brockett, C., Liden, L., Nouri, E., Yu, Z., Dolan, B., and Gao, J. (2022). GODEL: Large-scale pre-training for goal-directed dialog. *arXiv*, **cs.CL/2206.11309**.

Penha, G. and Hauff, C. (2021). On the calibration and uncertainty of neural learning to rank models for conversational search. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, EACL, pages 160–170.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, NAACL-HLT '18, pages 2227–2237.

Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., Plachouras, V., Rocktäschel, T., and Riedel, S. (2021). KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT 2021, pages 2523–2544.

Pirolli, P. and Card, S. (2015). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, ICIA '15, pages 2–4.

Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281.

Pradeep, R., Nogueira, R., and Lin, J. (2021). The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *arXiv*, **cs.IR/2101.05667**.

Pradeep, R., Sharifymoghaddam, S., and Lin, J. (2023). RankVicuna: Zero-shot listwise document reranking with open-source large language models. **cs.IR/2309.15088**.

Pradeep, R., Thakur, N., Upadhyay, S., Campos, D., Craswell, N., and Lin, J. (2024). Initial nugget evaluation results for the TREC 2024 RAG track with the AutoNuggetizer framework. *arXiv*, **cs.IR/2411.09607**.

Qin, Z., Jagerman, R., Hui, K., Zhuang, H., Wu, J., Yan, L., Shen, J., Liu, T., Liu, J., Metzler, D., Wang, X., and Bendersky, M. (2024). Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, NAACL-HLT '24, pages 1504–1518.

Rackauckas, Z., Câmara, A., and Zavrel, J. (2024). Evaluating RAG-fusion with RAGElo: an automated elo-based framework. *arXiv*, **cs.IR/2406.14783**.

Radensky, M., Séguin, J. A., Lim, J. S., Olson, K., and Geiger, R. (2023). "I think you might like this": Exploring effects of confidence signal patterns on trust in and reliance on conversational recommender systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 792–804.

Rader, E., Cotter, K., and Cho, J. (2018). Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–13.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI*.

Radlinski, F. and Craswell, N. (2017). A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 117–126.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, EMNLP '16, pages 2383–2392.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL '18.

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., and Shoham, Y. (2023). In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, **11**, 1316–1331.

Rashkin, H., Nikolaev, V., Lamm, M., Aroyo, L., and Collins, M. (2021). Measuring attribution in natural language generation models. *Computational Linguistics*, **49**(4), 777–840.

Rechkemmer, A. and Yin, M. (2022). When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–14.

Reddy, S., Chen, D., and Manning, C. D. (2019). CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, **7**, 249–266.

Ren, P., Chen, Z., Ren, Z., Kanoulas, E., Monz, C., and De Rijke, M. (2021). Conversations with search engines: SERP-based conversational response generation. *ACM Transactions on Information Systems*, **39**(4).

Ren, R., Wang, Y., Qu, Y., Zhao, W. X., Liu, J., Tian, H., Wu, H., Wen, J.-R., and Wang, H. (2025). Investigating the factual knowledge boundary of large language models with retrieval augmentation. In *Proceedings of the 31st International Conference on Computational Linguistics*, COLING '25, pages 3697–3715.

Robertson, S. (1977). The probability ranking principle in IR. *Journal of Documentation*, **33**(4), 294–304.

Robertson, S. and Zaragoza, H. (2009). The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, **3**(4), 333–389.

Sakaeda, R. and Kawahara, D. (2022). Generate, evaluate, and select: A dialogue system with a response evaluator for diversity-aware response generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '22.

Sakai, T. (2018). *Laboratory Experiments in Information Retrieval*. Springer Singapore.

Sakai, T. (2023). SWAN: A generic framework for auditing textual conversational systems. *arXiv*, **cs.IR/2305.08290**.

Salton, G. (1968). *Automatic Information Organization And Retrieval*. McGraw Hill Text.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, **18**(11), 613–620.

Samarinas, C., Dharawat, A., and Zamani, H. (2022). Revisiting open domain query facet extraction and generation. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '22, pages 43–50.

Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval*, **4**(4), 247–375.

Schneider, P., Afzal, A., Vladika, J., Braun, D., and Matthes, F. (2023). Investigating conversational search behavior for domain exploration. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval*, ECIR '23, pages 608–616.

Schuster, T., Lelkes, A. D., Sun, H., Gupta, J., Berant, J., Cohen, W. W., and Metzler, D. (2023). SEMQA: Semi-extractive multi-source question answering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 of *NAACL-HLT '23*, pages 1363–1381.

Sekulić, I., Aliannejadi, M., and Crestani, F. (2021). Towards facet-driven generation of clarifying questions for conversational search. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '21, pages 167–175.

Sekulić, I., Aliannejadi, M., and Crestani, F. (2022). Exploiting document-based features for clarification in conversational search. In *Advances in Information Retrieval: 44th European Conference on IR Research*, ECIR '22, pages 413–427.

Sekulić, I., Łajewska, W., Balog, K., and Crestani, F. (2024). Estimating the usefulness of clarifying questions and answers for conversational search. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval*, ECIR '24, page 384–392.

Shah, C. and Bender, E. M. (2022). Situating search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '22, pages 221–232.

Shani, G., Rokach, L., Shapira, B., Hadash, S., and Tangi, M. (2013). Investigating confidence displays for top-N recommendations. *Journal of the American Society for Information Science and Technology*, **64**(12), 2548–2563.

Shapira, O., Gabay, D., Gao, Y., Ronen, H., Pasunuru, R., Bansal, M., Amsterdamer, Y., and Dagan, I. (2019). Crowdsourcing lightweight pyramids for manual summary evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, NAACL-HLT '19, pages 682–687.

Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., and Yih, W.-t. (2024). REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, NAACL-HLT '24, pages 8371–8384.

Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, EMNLP '21, pages 3784–3803.

Skjæveland, M. G., Balog, K., Bernard, N., Łajewska, W., and Linjordet, T. (2024). An ecosystem for personal knowledge graphs: A survey and research roadmap. *AI Open*, **5**, 55–69.

Steen, J. and Markert, K. (2021). How to evaluate a summarizer: Study design and statistical analysis for manual linguistic quality evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, EACL '21, pages 1861–1875.

Subbiah, M., Zhang, S., Chilton, L. B., and McKeown, K. (2024). Reading subtext: Evaluating large language models on short story summarization with writers.

Sulem, E., Hay, J., and Roth, D. (2022). Yes, no or IDK: The challenge of unanswerable yes/no questions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '22, pages 1075–1085.

Sun, W., Yan, L., Ma, X., Wang, S., Ren, P., Chen, Z., Yin, D., and Ren, Z. (2023). Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, EMNLP '23, pages 14918–14937.

Sun, Z., Wang, X., Tay, Y., Yang, Y., and Zhou, D. (2022). Recitation-augmented language models. In *The Eleventh International Conference on Learning Representations*, ICLR '23.

Tan, C., Wei, F., Yang, N., Du, B., Lv, W., and Zhou, M. (2018). S-Net: From answer extraction to answer synthesis for machine reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI '18.

Tang, L., Goyal, T., Fabbri, A., Laban, P., Xu, J., Yavuz, S., Kryscinski, W., Rousseau, J., and Durrett, G. (2023). Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '23, pages 11626–11644.

Tang, X., Fabbri, A., Li, H., Mao, Z., Adams, G., Wang, B., Celikyilmaz, A., Mehdad, Y., and Radev, D. (2022). Investigating crowdsourcing protocols for evaluating the factual consistency of summaries. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '22.

ter Hoeve, M., Kiseleva, J., and de Rijke, M. (2022). What makes a good summary? reconsidering the focus of automatic summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '22, pages 46–75.

Tian, Z., Bi, W., Li, X., and Zhang, N. L. (2019). Learning to abstract for memory-augmented conversational response generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, pages 3816–3825.

Tigunova, A., Yates, A., Mirza, P., and Weikum, G. (2020). CHARM: Inferring personal attributes from conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 5391–5404.

Toader, D.-C., Boca, G., Toader, R., Măcelaru, M., Toader, C., Ighian, D., and Rădulescu, A. T. (2019). The effect of social presence and chatbot errors on trust. *Sustainability*, **12**(1), 256.

Tombros, A. and Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 2–10.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). LLaMA: Open and efficient foundation language models. *ArXiv*.

Trippas, J. R., Spina, D., Cavedon, L., Joho, H., and Sanderson, M. (2018). Informing the design of spoken conversational search: Perspective paper. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, pages 32–41.

Trippas, J. R., Spina, D., Thomas, P., Sanderson, M., Joho, H., and Cavedon, L. (2020). Towards a model for spoken conversational search. *Information Processing & Management*, **57**(2), 102–162.

Tsai, C.-H., You, Y., Gui, X., Kou, Y., and Carroll, J. M. (2021). Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–17.

Turney, P. D. (2005). Measuring semantic similarity by latent relational analysis. In *Proceedings of the 19th international joint conference on Artificial intelligence*, IJCAI '05, pages 1136–1141.

Turpin, A., Scholer, F., Mizzaro, S., and Maddalena, E. (2015). The benefits of magnitude estimation relevance assessments for information retrieval evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 565–574.

Vakharia, D. and Lease, M. (2013). Beyond AMT: An analysis of crowd work platforms. *arXiv*, **cs.CY/1310.1672**.

Vakulenko, S., Voskarides, N., Tu, Z., and Longpre, S. (2021a). A comparison of question rewriting methods for conversational passage retrieval. In *Advances in Information Retrieval: 43rd European Conference on IR Research*, ECIR '21, pages 418–424.

Vakulenko, S., Voskarides, N., Tu, Z., and Longpre, S. (2021b). Leveraging query resolution and reading comprehension for conversational passage retrieval. In *The Tweenty-Ninth Text REtrieval Conference Proceedings*, TREC '20.

Vakulenko, S., Longpre, S., Tu, Z., and Anantha, R. (2021c). Question rewriting for conversational question answering. In *WSDM '21*.

van der Waa, J., Nieuwburg, E., Cremers, A., and Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, **291**.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010.

Voorhees, E. M. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 315–323.

Voorhees, E. M. (2004). Overview of the TREC 2003 question answering track. In *The Twelfth Text Retrieval Conference Proceedings*, TREC '03.

Voskarides, N., Li, D., Panteli, A., and Ren, P. (2019). ILPS at TREC 2019 conversational assistant track. In *The Twenty-Eighth Text REtrieval Conference Proceedings*, TREC '19.

Voskarides, N., Li, D., Ren, P., Kanoulas, E., and de Rijke, M. (2020). Query resolution for conversational search with limited supervision. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 921–930.

Wang, X., Shou, L., Gong, M., Duan, N., and Jiang, D. (2020). No answer is better than wrong answer: A reflection model for document level machine reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, EMNLP, pages 4141–4150.

Wang, Zhenduo, Tu, Yuancheng, Rosset, Corby, Craswell, Nick, Wu, Ming, and Ai, Qingyao (2023). Zero-shot clarifying question generation for conversational search. In *Proceedings of the ACM Web Conference 2023*, WWW '23.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 24824–24837.

White, R. W. (2014). Belief dynamics in web search: Belief dynamics in web search. *Journal of the Association for Information Science and Technology*, **65**(11), 2165–2178.

Williams, J. P. (2008). Emergent themes. *The Sage encyclopedia of qualitative research methods*, **1**, 248–249.

Wilson, T. (1981). On user studies and information needs. *Journal of Documentation*, **37**(1), 3–15.

Wilson, T. D. (1999). Models in information behaviour research. *Journal of Documentation*, **55**(3), 249–270.

Xing, C., Wu, W., Wu, Y., Liu, J., Huang, Y., Zhou, M., and Ma, W.-Y. (2017). Topic aware neural response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31 of *AAAI '17*.

Xiong, L., Xiong, C., Li, Y., Tang, K.-F., Liu, J., Bennett, P. N., Ahmed, J., and Overwijk, A. (2020). Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv*, **cs.IR/2007.00808**.

Xu, F., Shi, W., and Choi, E. (2023). RECOMP: Improving retrieval-augmented lms with compression and selective augmentation.

Xu, J., Xia, L., Lan, Y., Guo, J., and Cheng, X. (2017). Directly optimize diversity evaluation measures: A new approach to search result diversification. *ACM Transactions on Intelligent Systems and Technology*, **8**(3), 1–26.

Xu, Y. C. and Chen, Z. (2006). Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology*, **57**(7), 961–973.

Yan, X., Clarke, C., and Arabzadeh, N. (2021). WaterlooClarke at the TREC 2021 conversational assistant track. In *The Thirtieth Text REtrieval Conference Proceedings*, TREC '21.

Yang, H., Li, Z., Zhang, Y., Wang, J., Cheng, N., Li, M., and Xiao, J. (2023). PRCA: Fitting black-box large language models for retrieval question answering via pluggable reward-driven contextual adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, EMNLP '23.

Yang, J.-H., Lin, S.-C., Wang, C.-J., Lin, J. J., and Tsai, M.-F. (2019). Query and answer expansion from conversation history. In *The Twenty-Eighth Text REtrieval Conference Proceedings*, TREC '19.

Yang, W., Li, Y., Fang, M., and Chen, L. (2024). Enhancing temporal sensitivity and reasoning for time-sensitive question answering.

Yates, A., Nogueira, R., and Lin, J. (2021). Pretrained transformers for text ranking: BERT and beyond. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2666–2668.

Yilmaz, Z. A., Wang, S., and Lin, J. (2019). H2oloo at TREC 2019: Combining sentence and document evidence in the deep learning track. In *The Twenty-Eighth Text REtrieval Conference Proceedings*, TREC '19.

Yoon, C., Lee, T., Hwang, H., Jeong, M., and Kang, J. (2024). CompAct: Compressing retrieved documents actively for question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, EMNLP '24, pages 21424–21439.

Yu, S., Liu, J., Yang, J., Xiong, C., Bennett, P., Gao, J., and Liu, Z. (2020). Few-shot generative conversational query rewriting. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 1933–1936.

Yu, S., Liu, Z., Xiong, C., Feng, T., and Liu, Z. (2021). Few-shot conversational dense retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 829–838.

Zamani, H., Dumais, S., Craswell, N., Bennett, P., and Lueck, G. (2020). Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, WWW '20, pages 418–428.

Zamani, H., Trippas, J. R., Dalton, J., and Radlinski, F. (2023). Conversational information seeking. *Foundations and Trends® in Information Retrieval*, **17**(3-4), 244–456.

Zhai, C. and Massung, S. (2016). *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Association for Computing Machinery and Morgan & Claypool.

Zhang, H., Zhang, R., Guo, J., de Rijke, M., Fan, Y., and Cheng, X. (2024). Are large language models good at utility judgments? In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1941–1951.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. (2020a). PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML '20*, pages 11328–11339.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018a). Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '18, pages 2204–2213.

Zhang, Y. and Chen, X. (2020). Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, **14**(1), 1–101.

Zhang, Y., Chen, X., Ai, Q., Yang, L., and Croft, W. B. (2018b). Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pages 177–186.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2020b). DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL '20, pages 270–278.

Zhang, Y., Liao, Q. V., and Bellamy, R. K. E. (2020c). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT '20, pages 295–305.

Zhang, Z., Yang, J., and Zhao, H. (2021). Retrospective reader for machine reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI '21, pages 14506–14514.

Zheng, Q., Tang, Y., Liu, Y., Liu, W., and Huang, Y. (2022). UX research on conversational human-ai interaction: A literature review of the acm digital library. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–24.

Zheng, X., Che, F., Wu, J., Zhang, S., Nie, S., Liu, K., and Tao, J. (2024). KS-LLM: Knowledge selection of large language models with evidence document for question answering. *Clinical Orthopaedics and Related Research.*

Zhong, L., Cao, J., Sheng, Q., Guo, J., and Wang, Z. (2020). Integrating semantic and structural information with graph convolutional network for controversy detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 515–526.

Zhou, Q., Yang, N., Wei, F., Huang, S., Zhou, M., and Zhao, T. (2018). Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '18, pages 654–663.