Can Users Detect Biases or Factual Errors in Generated Responses in Conversational Information-Seeking?

Weronika Łajewska
University of Stavanger
Stavanger, Norway
weronika.lajewska@uis.no

Krisztian Balog
University of Stavanger
Stavanger, Norway
krisztian.balog@uis.no

Damiano Spina
RMIT University
Melbourne, Australia
damiano.spina@rmit.edu.au

Johanne Trippas
RMIT University
Melbourne, Australia
j.trippas@rmit.edu.au

Abstract

Information-seeking dialogues span a wide range of questions, from simple factoid to complex queries that require exploring multiple facets and viewpoints. When performing exploratory searches in unfamiliar domains, users may lack background knowledge and struggle to verify the system-provided information, making them vulnerable to misinformation. We investigate the limitations of response generation in conversational information-seeking systems, highlighting potential inaccuracies, pitfalls, and biases in the responses. The study addresses the problem of *query answerability* and the challenge of *response incompleteness*. Our user studies explore how these issues impact user experience, focusing on users' ability to identify biased, incorrect, or incomplete responses. We design two crowdsourcing tasks to assess user experience with different system response variants, highlighting critical issues to be addressed in future conversational information-seeking research. Our analysis reveals that it is easier for users to detect response incompleteness than query answerability and user satisfaction is mostly associated with response diversity, not factual correctness.

CCS Concepts

• **Information systems** → **Users and interactive retrieval**.

Keywords

Conversational response generation; Answerability; Viewpoints

ACM Reference Format:
Weronika Łajewska, Krisztian Balog, Damiano Spina, and Johanne Trippas. 2024. 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), December 9–12, 2024, Tokyo, Japan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3673791.3698409

SIGIR-AP '24, December 9–12, 2024, Tokyo, Japan

ACM ISBN 979-8-4007-0724-7/24/12
https://doi.org/10.1145/3673791.3698409

1 Introduction

Conversational information-seeking (CIS) interactions enable users to fulfill complex information needs, navigate unknown domains, ask follow-up questions, and provide feedback via a series of natural language dialogues [55]. CIS research currently centers on retrieval components, such as passage retrieval, re-ranking, and query rewriting [32, 55]. 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* [38]. Unfortunately, responses generated by CIS systems are susceptible to limitations, including hallucinations when no answer is found [20], biased responses that only partially answer the question [17], or factual error presentation [45]. 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 [9], factually incorrect, unsupported, biased, or incomplete information may be easily overlooked.

This paper investigates users' ability to recognize pitfalls in CIS systems related to *query answerability* and *response incompleteness* (see Table 1). We hypothesize that untrained users cannot identify these problems in CIS interactions. More specifically, we aim to address the following research questions:

RQ1: Can users effectively recognize problems related to *query answerability* and *response incompleteness* in system responses?
RQ2: How do factually incorrect, 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 [13, 32] with manually injected inaccuracies or biases in a controlled manner. Query answerability can be defined at different levels, which includes determining whether 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 paper, we focus on (i) the consequences of generating response from passages that do not contain the answer, which result 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 [24] and the presence/validity of the information source [5, 29]. 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 paper, 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) [18] and balance in covering various viewpoints/facets in the response.

Results of the *answerability study* show that users cannot recognize factual errors in system responses. Additionally, a lack of source or an invalid source does not decrease their confidence in the response. On the other hand, according to the *viewpoints study*, it is easier for users to identify problems with viewpoint diversity and balance, and recognize if the response is biased or incomplete (**RQ1**). Moreover, the satisfaction ratings and comments from our user experience questionnaire reveal that overall system response satisfaction is associated with the investigated response dimensions (i.e., factual correctness and source presence/validity for the *answerability study*; diversity and balanced viewpoints/facets presentation for the *viewpoints study*), even though the satisfaction and response dimension ratings do not fully align with the comments (**RQ2**). The findings from our two studies reveal that the fluency of CIS responses may compromise users' capability to spot inaccurate information—even in responses attributed to the source.

In summary, the main contributions of this work include: (1) a novel methodology to study how users perceive query answerability and response incompleteness in CIS, (2) a manually curated dataset of CIS responses varying across multiple dimensions in terms of quality, (3) quantitative and qualitative analysis of how users recognize and perceive limitations in CIS responses. The developed resources (manually generated CIS responses, user studies results, data analysis scripts) and additional analysis are available at: https://github.com/iai-group/sigirap2024-resgen.

2 Related Work

In search engine results pages (SERPs), users are presented with a ranked list of results, each accompanied by a source, providing a broader context for the requested information. In contrast, response generation in CIS systems offers limited information in the form of natural language responses, typically consisting of several sentences [38]. This shift has been facilitated by the widespread adoption of generative language models, which enable systems to produce fluent and coherent responses [56]. 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 [28]. User studies on CIS human-system interactions highlight desired response features, such as opinion aggregation, including information sources to ensure credibility, response verifiability, as well as balance between direct answers and expanded information to address the conciseness/completeness tradeoff [29, 49, 51]. CIS systems aim to provide users with responses that encapsulate relevant information from multiple documents, creating a more natural and user-friendly experience [10, 55]. Multi-document summary generation has been studied in the context of non-factoid question answering [5], generative search engine results verifiability [29], and generating summaries from retrieved results that was one of the tasks proposed in TREC CAsT'22 [32].

Table 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… 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.

However, transitioning from SERPs to conversational responses presents distinct challenges [41, 48, 53, 54]. Even though the ranking of top retrieved passages should ensure fairness [17] and viewpoint diversification [15, 17, 39], it is a non-trivial task to synthesize those passages into a reliable, trustworthy, and concise response. Additionally, cognitive biases, such as anchoring bias and confirmation bias, can impact user interactions with the search system, potentially disrupting the overall user experience [2, 22, 28, 31, 43, 52]. Response generation from retrieved passages faces additional challenges related to temporal considerations [7], biased queries [2, 22], source subjectivity, unanswerability [8, 33, 35, 37], the lack of expert knowledge , and additional issues related to text aggregation that may introduce hallucinations and factual errors [45]. Given the potential flaws that may result from these challenges, conversational response generation should involve system revealment and promote a more informed user experience [3, 26, 34]. In the proposed user studies, we aim to investigate to what extent users are unaware of the inaccuracies in the responses, and what the scale of the problem is in CIS. Nevertheless, providing users with an understanding of the search space and transparently conveying the system's certainty and potential pitfalls are essential for promoting user trust and informed interactions with the system.

Evaluating response quality in CIS systems presents unique challenges, as traditional offline evaluation measures like ROUGE [27] (commonly used for evaluating summaries) and NDCG [19] (for evaluating passage rankings) 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 [40], including trust and fairness [55], credibility [4], reliability [30, 36], verifiability [29], factual correctness, transparency (e.g., information sources, ranking, and consolidation process) [42], relevance, naturalness, conciseness [32], informativeness (supporting user in increasing their information literacy) [42], perceived satisfaction, and usefulness [6, 47, 57]. However, directly asking users to report on these metrics may not be reliable as users may interpret the concepts differently (the problem of indirect observables) [21]. To tackle this challenge, our research focuses on understanding the response dimensions that are (1) associated with user satisfaction and (2) affected by answerability and incompleteness issues.

3 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[1] employing a within-subject design that investigate 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 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 paper 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.

[1]This investigation was compliant with the ethics approval process of our institution.

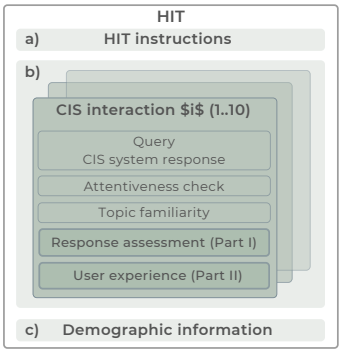

Figure 1: High-level design of the user studies.

3.1 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 variants for each query). The experimental conditions of manually crafted responses for both user studies are described in Section 4.1.2.

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 [21]. 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 [44].

3.2 Tasks

The design of the *answerability study* and the *viewpoints study* follows 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 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 [16]. 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) , and (4) a measurement of user experience (Part II).[2] The wording of the questions in all parts of the user studies follows questions proposed by Tang et al. [46] for evaluating the factual consistency of summaries (see Figure 2). Both studies

[2]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 2: Controlled vs. user-judged response dimensions.

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

finish with a short demographics questionnaire asking workers' age, education level, and gender.

3.2.1 ***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 [16]. 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 [5]. Submissions that failed on more than 3/10 attentiveness questions were rejected.

3.2.2 ***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 [23].

3.2.3 ***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 investigating different response dimensions for *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 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.

3.2.4 ***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.

3.3 Data Analysis Methods

To address RQ1, 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 [21] 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 of the user-judged response dimensions (dependent variables) 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). The results of our user studies are reported in Section 5. We analyze the crowdsourced data with the Python `statsmodels` library.[3] We use significance level $\alpha = 0.05$ to report statistical significance. Whenever applicable, the ω^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. [11] ($\omega^2 \geq 0.14$: large effect size; 0.06–0.14: medium; 0.01–0.06: small; ≤ 0: no effect).

4 User Study Execution

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

4.1 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 [13] and '22 [32],[4] simulating everyday system interactions under various experimental conditions. The responses are curated by the authors of the paper to ensure accordance with defined response dimensions and high data quality.

4.1.1 ***Queries.*** For each user study, we select ten queries from the topics released in CAsT 2020 and 2022 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 [25] 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 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

[3] https://www.statsmodels.org/

[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.

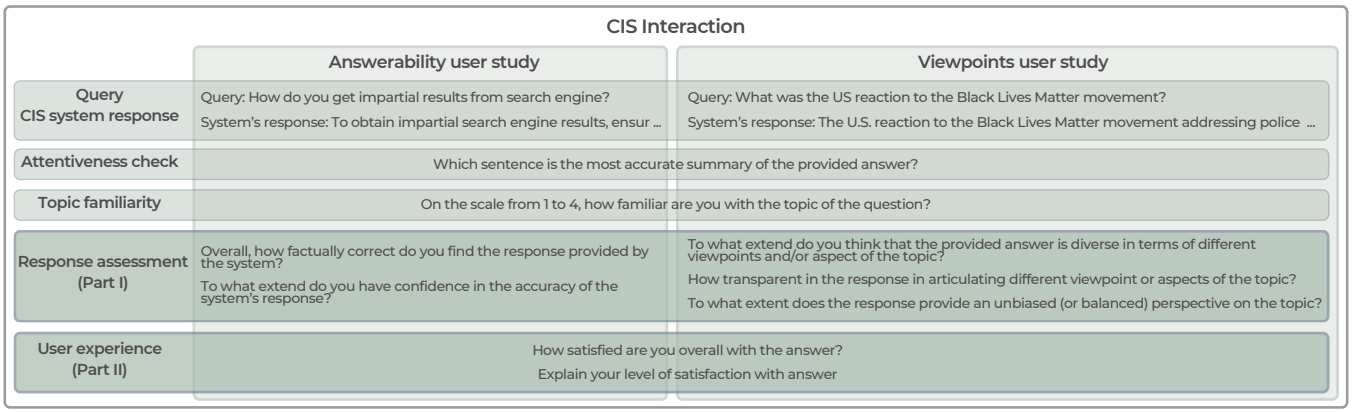

Figure 2: Questions provided to crowd workers in our user studies.

Table 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?

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 3.

Viewpoints Study. Open-ended queries about complex or contentious topics with multiple facets and/or viewpoints are specifically prone to incomplete responses [15]. To identify such queries in TREC CAsT collections, we: (1) manually select a subset of potential candidates and (2) ask crowdworkers 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 (see Section 4.1.2 for more details about the process). The final ten queries used in the *viewpoints study* are presented in Table 4.

[5]Note that answerability can be determined w.r.t. a document (e.g., SQuAD 2.0 [35]), corpus (e.g., TREC CAsT [12]), knowledge base [33], or external expert knowledge. In this paper, we consider answerability w.r.t. a particular set of retrieved passages.

Table 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.

4.1.2 ***Responses***. The responses were manually created by the authors of this paper 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 [45]. 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/unsupported information with an invalid source (EC_3^A), or without a source (EC_4^A);

Table 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 3).

	Experimental Condition	Response Dimension		CIS System Response	Source
		Factual Corr.	Source		
EC_1^A	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	Factually correct + no source	✓	×	Same as above	–
EC_3^A	Factually incorrect + invalid source	×	✓ (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/ (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	Factually incorrect + no source	×	×	Same as above	–

see Table 5. The flawed response may contain various factual inconsistencies, such as negation and number, entity, or antonym swaps [24], as well as fully hallucinated content not supported by any source information [20, 29]. 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 corrupted link. Following the setup proposed for evaluating the usefulness of supporting documents in the WikiHowQA benchmark [5], 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) [14]. 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 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 6.[6]

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

4.2 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 [44]. 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 *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 and previous research, which suggests that users are unlikely to detect hallucinations [26]. 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 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, 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

[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.

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

Table 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).

	Experimental Condition	Response Dimension: Diversity	Response Dimension: Balance	CIS System Response
EC_1^V	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	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	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.*

Table 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

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.

5 Results and Discussion

The analysis of data obtained from the crowdsourcing experiments is performed using the methods described in Section 3.3.

5.1 Users' Ability to Recognize Problems

Table 8 shows the results of the two-way ANOVA performed to answer RQ1 (*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 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 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.

Effect of the interaction between query and controlled response dimensions on user ratings. The three-away ANOVA results in Table 10 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 [1, 11]. 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.

Table 8: Results of two-way ANOVA. Statistically significant effects are in bold. Effect size: L=Large, M=Medium, S=Small.

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	-
Overal 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	1.000	-
	Contr. Diversity * Contr. Balance	**0.0**	**M**
Transparency	**Contr. Diversity**	**0.0**	**M**
	Contr. Balance	1.000	-
	Contr. Diversity * Contr. Balance	**0.0**	**M**
Balance	**Contr. Diversity**	**0.0**	**S**
	Contr. Balance	1.000	-
	Contr. Diversity * Contr. Balance	**0.0**	**S**
Overall Satisfaction	**Contr. Diversity**	**0.0**	**S**
	Contr. Balance	1.000	-
	Contr. Diversity * Contr. Balance	**0.0**	**M**

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.[8]

[8] Detailed results are in the online repository.

Table 9: Results of one-way ANOVA. Statistically significant effects are in bold. Effect size: L=Large, M=Medium, S=Small.

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	-

Table 10: Results of three-way ANOVA. Stat. significant effects are in bold. Effect size: L=Large, M=Medium, S=Small.

Dependent Variable (User-Judged)	Independent Variable(s) (Controlled)	p-value	Effect 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	1.000	-
	Contr. Diversity * Contr. Balance * Query	**0.016**	**S**
Transparency	Query	0.350	-
	Contr. Diversity * Query	0.582	-
	Contr. Balance * Query	1.000	-
	Contr. Diversity * Contr. Balance * Query	0.689	-
Balance	**Query**	**0.012**	**S**
	Contr. Diversity * Query	0.559	-
	Contr. Balance * Query	1.000	-
	Contr. Diversity * Contr. Balance * Query	0.316	-
Overall Satisfaction	**Query**	**0.001**	**M**
	Contr. Diversity * Query	0.599	-
	Contr. Balance * Query	1.000	-
	Contr. Diversity * Contr. Balance * Query	**0.034**	**S**

5.2 User Experience

This section discusses the results to answer RQ2 (*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 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 overall user experience. Correlations for the *answerability study* are lower than for the *viewpoints study*. As we discussed in Section 5.1, 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 their 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.

Table 11: Pearson correlation between user-reported response dimensions and their overall satisfaction with 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 query and response quality on overall satisfaction. In both studies, the query significantly affects overall satisfaction (see Table 9). 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 8). The opposite observation is made in the *viewpoints study*, implying that workers can spot response inaccuracies. The three-way ANOVA (Table 10) 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 *answerability study*, may increase the sensitivity of the experiment.

5.3 Further Analysis

Rating distributions for response dimensions. In the *answerability study*, the ratings for user-judged response dimensions, topic familiarity, and overall satisfaction per query are concentrated around higher values (3 and 4) for all response dimensions apart from familiarity.[9] 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 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).

[9]Data distribution can be found here: https://github.com/iai-group/sigirap2024-resgen/blob/main/results/quantitative_analysis.

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 9), 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 9). 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.

6 Discussion

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.[10] Users may also associate their satisfaction with response fluency, that can be easily ensured by existing generative search engines. However, it does not guarantee the accuracy or proper citation of all statements [29].

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 [26]. 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.

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 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 [50]). 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.

7 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. Our findings provide evidence that: (i) CIS system responses cannot be limited to a simple synthesis of the retrieved information; and (ii) 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 results presented in this paper can be regarded as guidelines for designing CIS solutions and conducting a more comprehensive analysis of the problems in the future.

Acknowledgments

This research was supported by the Norwegian Research Center for AI Innovation, NorwAI (Research Council of Norway, nr. 309834), and by the Australian Research Council (DE200100064, CE200100005).

[10] Additional qualitative analysis of the impact of response inaccuracies and biases on user experience based on free-text comments is in the online repository: https://github.com/iai-group/sigirap2024-resgen/tree/main/results/qualitative_analysis.

References

[1] Marwah Alaofi, Luke Gallagher, Dana Mckay, Lauren L. Saling, Mark Sanderson, Falk Scholer, Damiano Spina, and Ryen W. White. 2022. Where Do Queries Come From?. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. 2850–2862.

[2] Leif Azzopardi. 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)*. 27–37.

[3] Leif Azzopardi, Mateusz Dubiel, Martin Halvey, and Jeffrey Dalton. 2018. Conceptualizing Agent-Human Interactions During the Conversational Search Process. In *2nd International ACM SIGIR Workshop Conference on Conversational Approaches to IR (CAIR '18)*.

[4] Markus Bink, Steven Zimmerman, and David Elsweiler. 2022. Featured Snippets and their Influence on Users' Credibility Judgements. In *Proceedings of the 2022 Conference on Human Information Interaction & Retrieval (CHIIR '22)*. 113–122.

[5] Valeriia Bolotova-Baranova, Vladislav Blinov, Sofya Filippova, Falk Scholer, and Mark Sanderson. 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)*. 5291–5314.

[6] B. Barla Cambazoglu, Valeriia Baranova, Falk Scholer, Mark Sanderson, Leila Tavakoli, and Bruce Croft. 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)*. 75–84.

[7] Ricardo Campos, Gaël Dias, Alípio M. Jorge, and Adam Jatowt. 2015. Survey of Temporal Information Retrieval and Related Applications. *Comput. Surveys* 47, 2 (2015), 1–41.

[8] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP '18)*.

[9] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 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)*. 7282–7296.

[10] J. Shane Culpepper, Fernando Diaz, and Mark D. Smucker. 2018. Research Frontiers in Information Retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018). *ACM SIGIR Forum* 52, 1 (2018), 34–90.

[11] J. Shane Culpepper, Guglielmo Faggioli, Nicola Ferro, and Oren Kurland. 2022. Topic Difficulty: Collection and Query Formulation Effects. *ACM Transactions on Information Systems* 40, 1 (2022), 1–36.

[12] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2019. TREC CAsT 2019: The Conversational Assistance Track Overview. In *The Twenty-Eighth Text REtrieval Conference Proceedings (TREC '19)*. arXiv:2003.13624

[13] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. CAsT 2020: The Conversational Assistance Track Overview. In *The Twenty-Ninth Text REtrieval Conference Proceedings (TREC '20)*.

[14] Tim Draws, Oana Inel, Nava Tintarev, Christian Baden, and Benjamin Timmermans. 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)*. 135–145.

[15] Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. 2021. 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)*.

[16] Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. 2015. Human Beyond the Machine: Challenges and Opportunities of Microtask Crowdsourcing. *IEEE Intelligent Systems* 30, 4 (2015), 81–85.

[17] Ruoyuan Gao and Chirag Shah. 2020. Toward creating a fairer ranking in search engine results. *Information Processing & Management* 57, 1 (2020), 102–138.

[18] Natali Helberger, Kari Karppinen, and Lucia D'Acunto. 2018. Exposure diversity as a design principle for recommender systems. *Information, Communication & Society* 21, 2 (2018), 191–207.

[19] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20, 4 (2002), 422–446.

[20] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[21] Diane Kelly. 2007. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval* 3, 1—2 (2007), 1–224.

[22] Johannes Kiesel, Damiano Spina, Henning Wachsmuth, and Benno Stein. 2021. The Meant, the Said, and the Understood: Conversational Argument Search and Cognitive Biases. In *Proceedings of the 3rd Conference on Conversational User Interfaces (CUI '21)*. 1–5.

[23] Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 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)*. 4940–4957.

[24] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: EMNLP 2020 (EMNLP '20)*. 9332–9346.

[25] Weronika Łajewska and Krisztian Balog. 2023. 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)*. 5326–5330.

[26] Weronika Łajewska, Damiano Spina, Johanne Trippas, and Krisztian Balog. 2024. Explainability for Transparent Conversational Information-Seeking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. 1040–1050.

[27] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Association for Computational Linguistics (ACL '04)*. 74–81.

[28] Jiqun Liu. 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)*. 236–246.

[29] Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023 (EMNLP '23)*. 7001–7025.

[30] Zhuoran Lu and Ming Yin. 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)*.

[31] Dana McKay, Kaipin Owyong, Stephann Makri, and Marisela Gutierrez Lopez. 2022. Turn and Face the Strange: Investigating Filter Bubble Bursting Information Interactions. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval (CHIIR '22)*. 233–242.

[32] Paul Owoicho, Jeffrey Dalton, Mohammad Aliannejadi, Leif Azzopardi, Johanne R Trippas, and Svitlana Vakulenko. 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)*.

[33] Sachin Pathiyan Cherumanal, Lin Tian, Futoon M. Abushaqra, Angel Felipe Magnossão De Paula, Kaixin Ji, Halil Ali, Danula Hettiachchi, Johanne R. Trippas, Falk Scholer, and Damiano Spina. 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)*. 401–405.

[34] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. 117–126.

[35] Pranav Rajpurkar, Robin Jia, and Percy Liang. 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)*.

[36] Amy Rechkemmer and Ming Yin. 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 System (CHI '22)*.

[37] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics* 7 (2019), 249–266.

[38] Pengjie Ren, Zhumin Chen, Zhaochun Ren, Evangelos Kanoulas, Christof Monz, and Maarten De Rijke. 2021. Conversations with Search Engines: SERP-based Conversational Response Generation. *ACM Transactions on Information Systems* 39, 4 (2021).

[39] Ryoma Sakaeda and Daisuke Kawahara. 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)*.

[40] Tetsuya Sakai. 2023. SWAN: A Generic Framework for Auditing Textual Conversational Systems. arXiv:2305.08290 [cs.IR]

[41] Tal Schuster, Adam Lelkes, Haitian Sun, Jai Gupta, Jonathan Berant, William Cohen, and Donald Metzler. 2024. 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: Long Papers) (NAACL-HLT'24)*. 1363–1381.

[42] Chirag Shah and Emily M. Bender. 2022. Situating Search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '22)*. 221–232.

[43] Damiano Spina, Danula Hettiachchi, and Anthony McCosker. 2024. *Quantifying and Measuring Bias and Engagement in Automated Decision-Making*. Technical Report. ARC Centre of Excellence for Automated Decision-Making and Society, RMIT University, Melbourne, Australia.

[44] Julius Steen and Katja Markert. 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 (ACL '21)*. 1861–1875.

[45] Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 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)*. 11626–11644.

[46] Xiangru Tang, Alexander Fabbri, Haoran Li, Ziming Mao, Griffin Adams, Borui Wang, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 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)*.

[47] Maartje Ter Hoeve, Julia Kiseleva, and Maarten Rijke. 2022. What Makes a Good and Useful Summary? Incorporating Users in Automatic Summarization Research. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'22)*. 46–75.

[48] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the Design of Spoken Conversational Search: Perspective Paper. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. 32–41.

[49] Johanne R. Trippas, Damiano Spina, Paul Thomas, Mark Sanderson, Hideo Joho, and Lawrence Cavedon. 2020. Towards a model for spoken conversational search. *Information Processing & Management* 57, 2 (2020), 102–162.

[50] Andrew Turpin, Falk Scholer, Stefano Mizzaro, and Eddy Maddalena. 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)*. 565–574.

[51] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles L. A. Clarke. 2017. Exploring Conversational Search With Humans, Assistants, and Wizards. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI '17)*. 2187–2193.

[52] Ryen W. White. 2014. Belief dynamics in web search: Belief Dynamics in Web Search. *Journal of the Association for Information Science and Technology* 65, 11 (2014), 2165–2178.

[53] Mei-Mei Wu. 2005. Understanding patrons' micro-level information seeking (MLIS) in information retrieval situations. *Information Processing & Management* 41, 4 (2005), 929–947.

[54] Mei-Mei Wu and Ying-Hsang Liu. 2003. Intermediary's information seeking, inquiring minds, and elicitation styles. *Journal of the American Society for Information Science and Technology* 54, 12 (2003), 1117–1133.

[55] Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. 2023. Conversational Information Seeking. *Foundations and Trends in Information Retrieval* 17, 3-4 (2023), 244–456.

[56] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. 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)*. 270–278.

[57] Qingxiao Zheng, Yiliu Tang, Yiren Liu, Weizi Liu, and Yun Huang. 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)*. 1–24.