

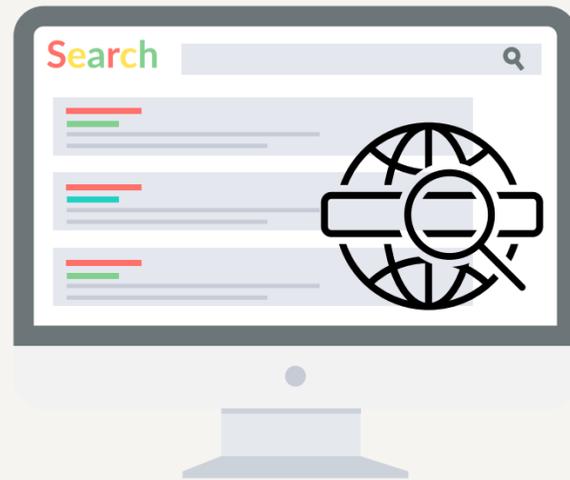
# Grounded and Transparent Response Generation for Conversational Information-Seeking Systems

Weronika Łajewska

EU InTech, Amazon

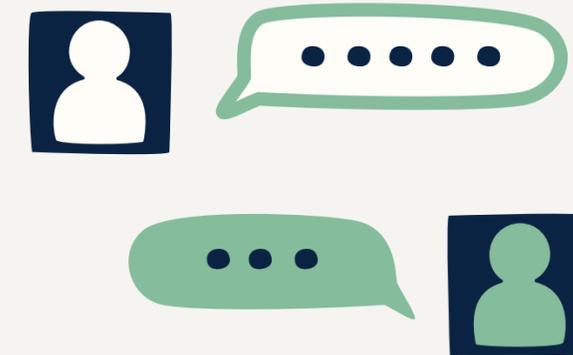
*Work done while at the University of Stavanger, Norway*

# Information Retrieval vs. Information Generation



## Traditional Search Engines

- Retrieve and rank existing web pages or documents based on relevance to the user's query
- The sources of information are displayed directly to the user
- Users often need to actively go through results



## LLM-based Conversational Assistant

- Generate responses by processing the input query and synthesizing information from the vast amount of data they've been trained on
- Responses do not explicitly cite the sources
- Immediate, cohesive answer is provided

# Conversational Information-Seeking (CIS) Queries

How can education systems be reformed to better prepare students for the challenges of the 21st century?

How has climate change influenced migration patterns globally in the last decade?

How can urban design contribute to mental health and well-being in densely populated cities?

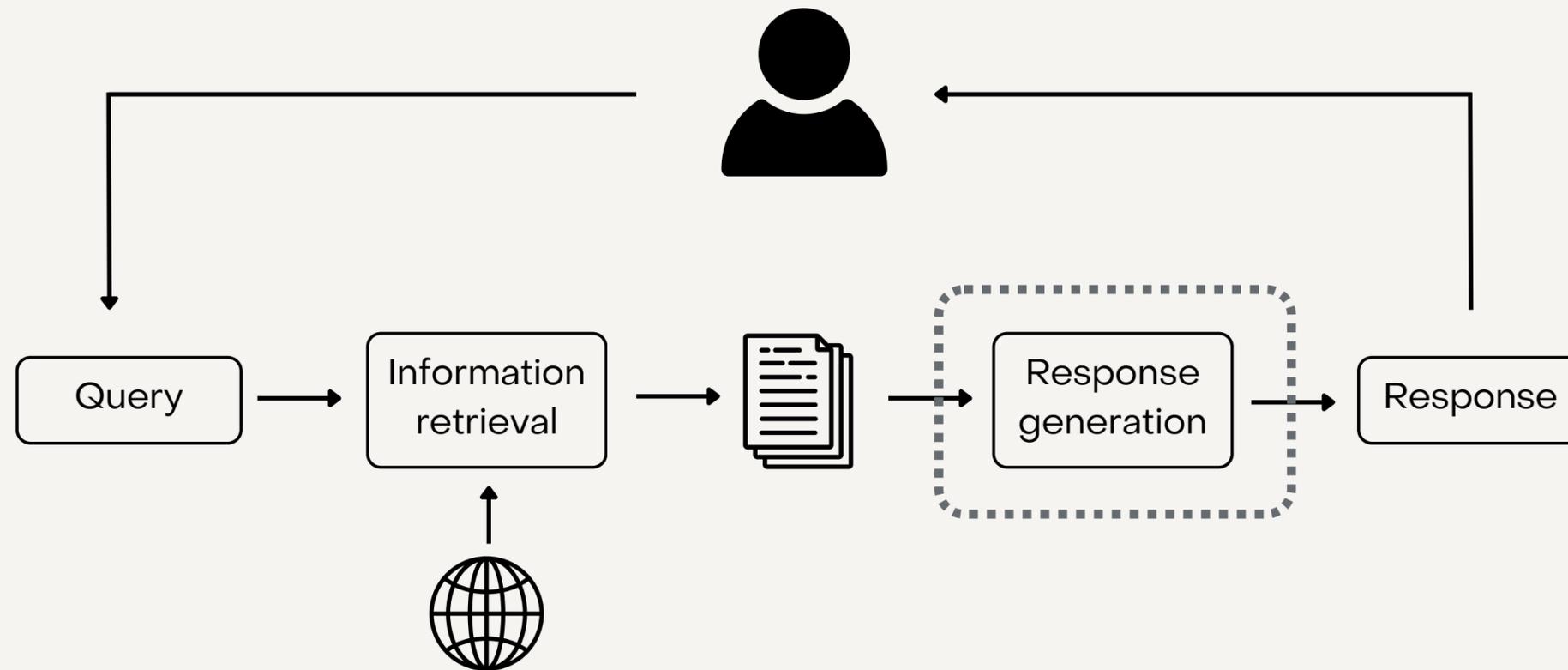
● They require covering multiple aspects or points of view

● In a conversational setting, responses are expected to be short and concise

● There is a trade-off between response completeness and succinctness

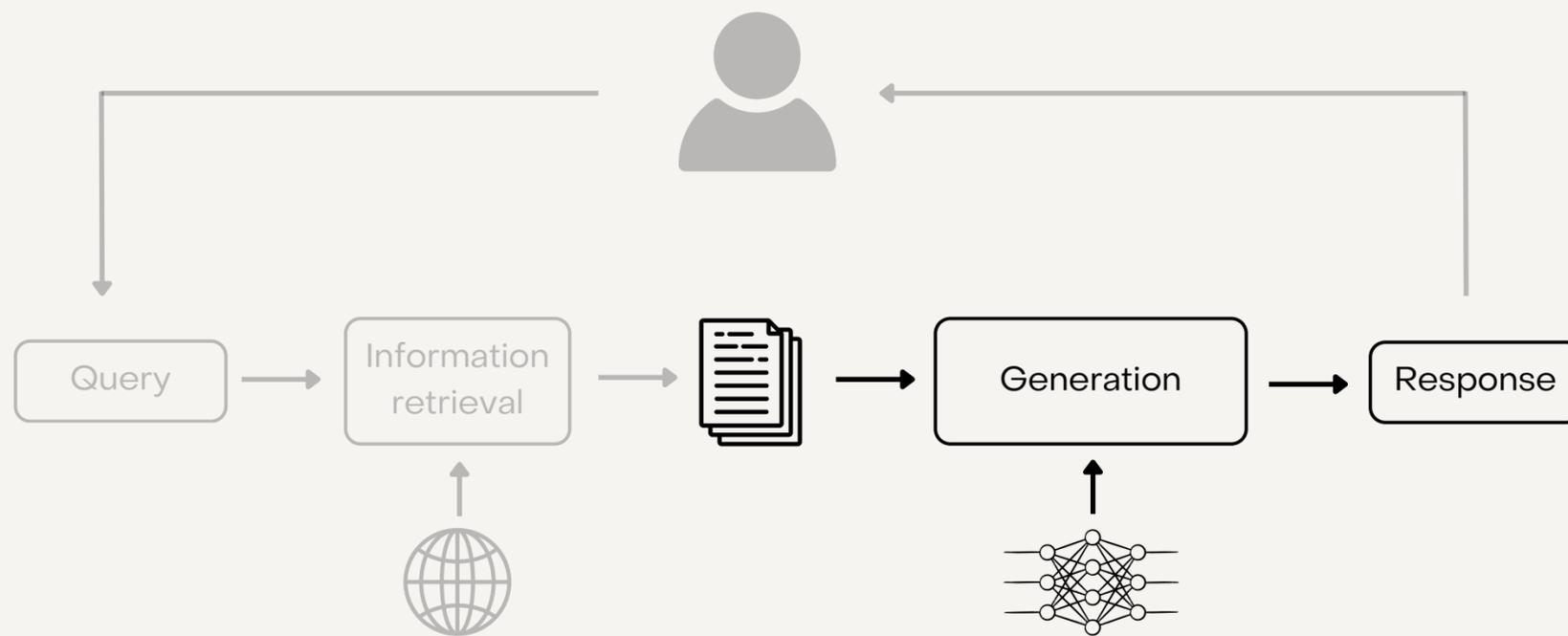
● The coverage of information in the response depends on user preferences, their background knowledge and previous interactions with the system

# Retrieval-Augmented Generation (RAG)



- Response grounding is not guaranteed as some facts from the generated response may be not supported by the provided evidence
- Redundant information and overly long contexts can lead to the “*lost in the middle*” problem<sup>[1]</sup>

# CIS Response Generation - Challenges



What can go wrong?

- System may fail to find the response
- The response may be biased
- Only part of the answer may be found
- Summarization with LLMs may introduce factual errors



# Objective

Conversational Information-Seeking (CIS) system capable of generating responses that enable users to navigate complex information needs successfully

**USER:** What was the US reaction to the Black Lives Matter movement?



**LLM:** 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.



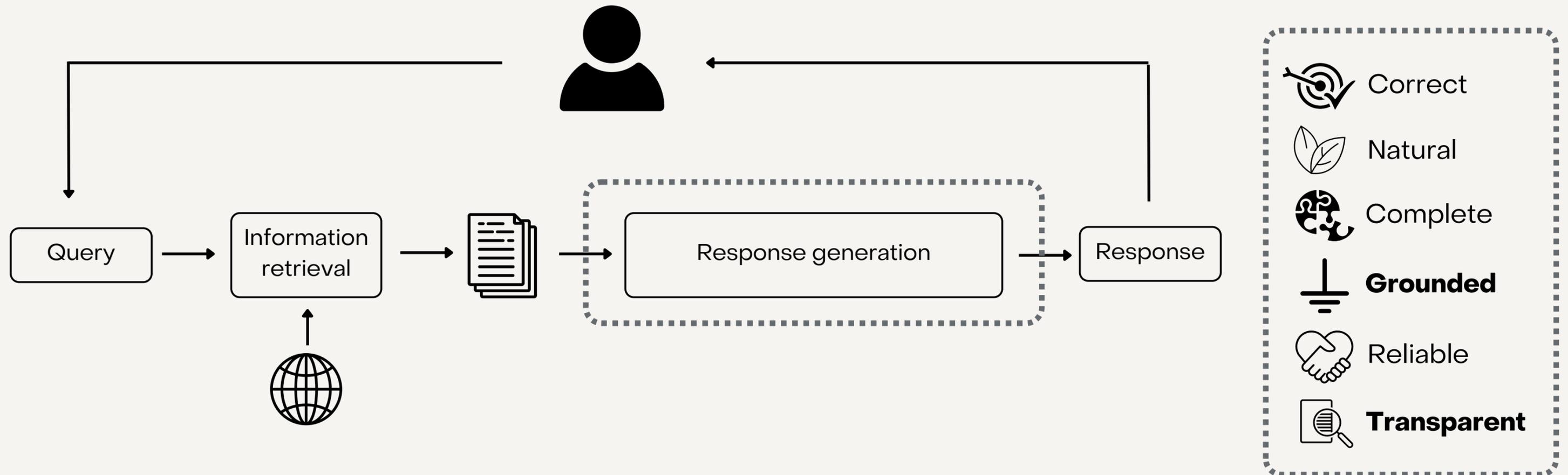
[Sources](#)

Response completeness



Controversial topic, multiple points of view possible!

# Explainable CIS System



# Part I – Understanding CIS Limitations

---

# Understanding CIS Limitations

Conversational search is a less transparent setting than SERP-based interface

Users are mostly not aware of the working mechanism of the system, its capabilities, and limitations

Detecting hallucinations, factual errors, and/or biases is extremely difficult for users without knowledge about the topic



Which limitations in the responses are detectable by users?

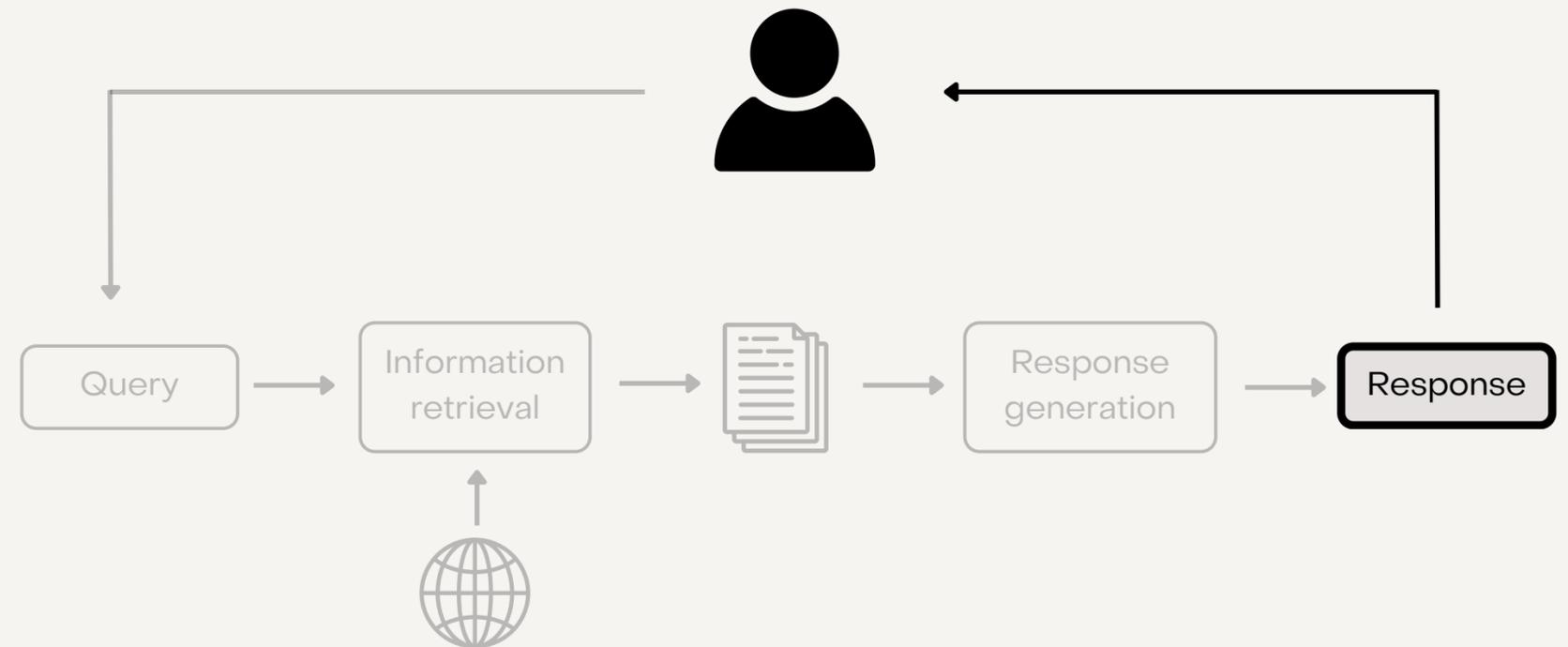
# Response Limitations in CIS



User-centric analysis of users' ability to effectively recognize the problems of:

- 1) query answerability resulting in hallucinations or invalid sources
- 2) multiple viewpoints leading to incomplete or biased response

.. as well as impact of inaccurate, incomplete, and/or biased responses on user experience



Łajewska, W., Balog, K., Spina, D., and Trippas, J. (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.

# User's Ability to Detect Response Inaccuracies

The second point of view only briefly mentioned

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?



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 ...



<https://www.winetourism.com/wine-tasting-in-penedes/>

Factual error  
+  
Invalid source

## Query Answerability

Studied by manipulating:

- Factual correctness
- Presence/validity of the information sources

What effects did the Watergate scandal have on President Nixon's legacy?



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.



## Response Incompleteness

Studied by manipulating:

- Diversity in terms of viewpoints and/or facets
- Balance in covering various viewpoints/facets in the response

# User Study Setup

Queries susceptible to one of the identified problems per study

Manually created response variants differing in terms of two controlled dimensions with isolated, easily detectable errors for each query

User asked to assess the provided responses and indicate their overall satisfaction

## Query Answerability

10 queries

4 responses per query

96 users

## Response Incompleteness

10 queries

3 responses per query

27 users

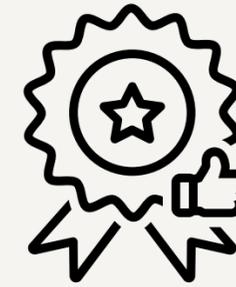
# Which response limitations are detectable by users?



Users find it easier to identify problems with diversity and balance of viewpoints rather than factual errors and source validity in the responses



Self-reported overall satisfaction scores are not necessarily associated with the main response dimensions



Credibility of the sources, as well as completeness, usefulness, and subjectivity of provided information impact the overall satisfaction of the users



- ⇒ CIS system responses cannot be limited to a simple synthesis of the retrieved information
- ⇒ Simple source attribution is not enough to ensure effective interaction with the system
- ⇒ CIS responses should explicitly inform users about potential inaccuracies and provide aid to assess the presented information objectively

## Part II – Addressing CIS Limitations

---

# Addressing CIS Limitations



How to generate grounded and transparent responses?

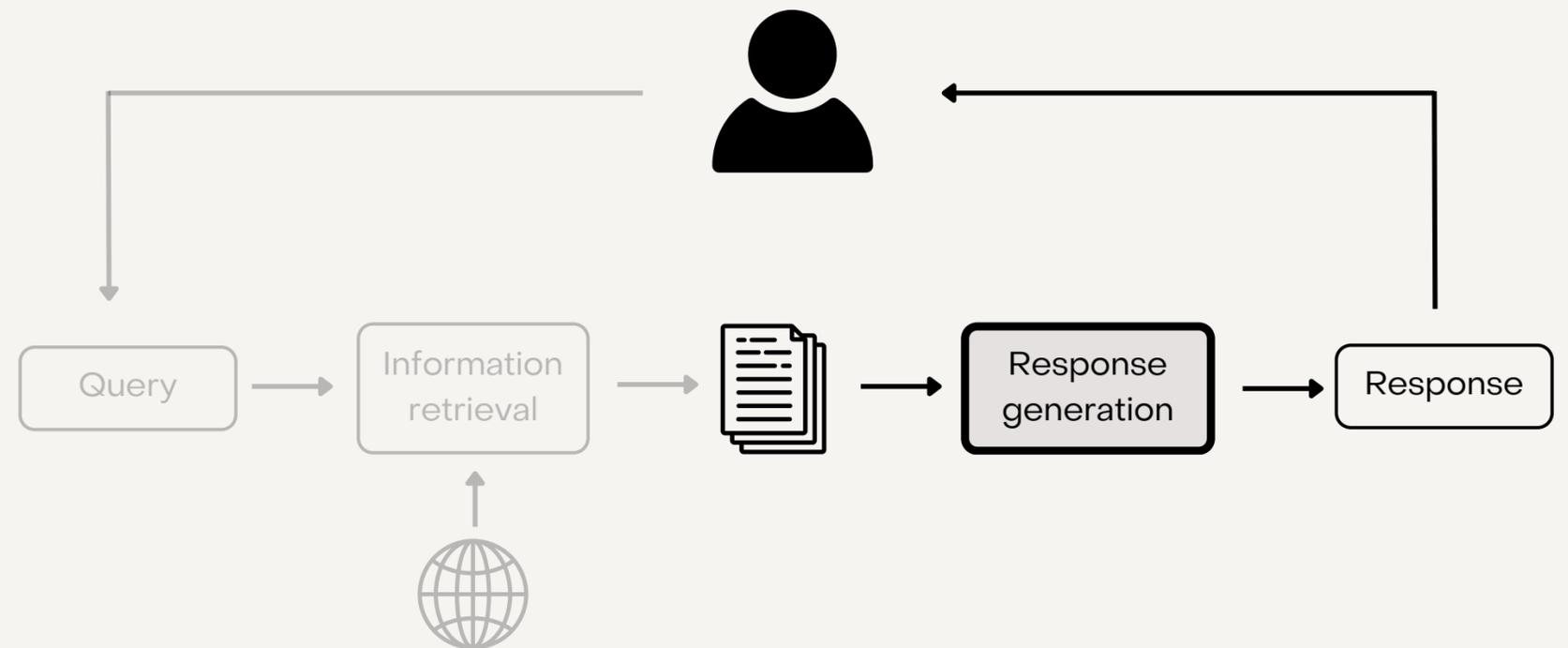


1. Detecting specific facts answering the question

2. Detecting response limitations

3. Synthesizing most important facts into a final response

4. Providing users with transparent response



# Addressing CIS Limitations



How to generate grounded and transparent responses?

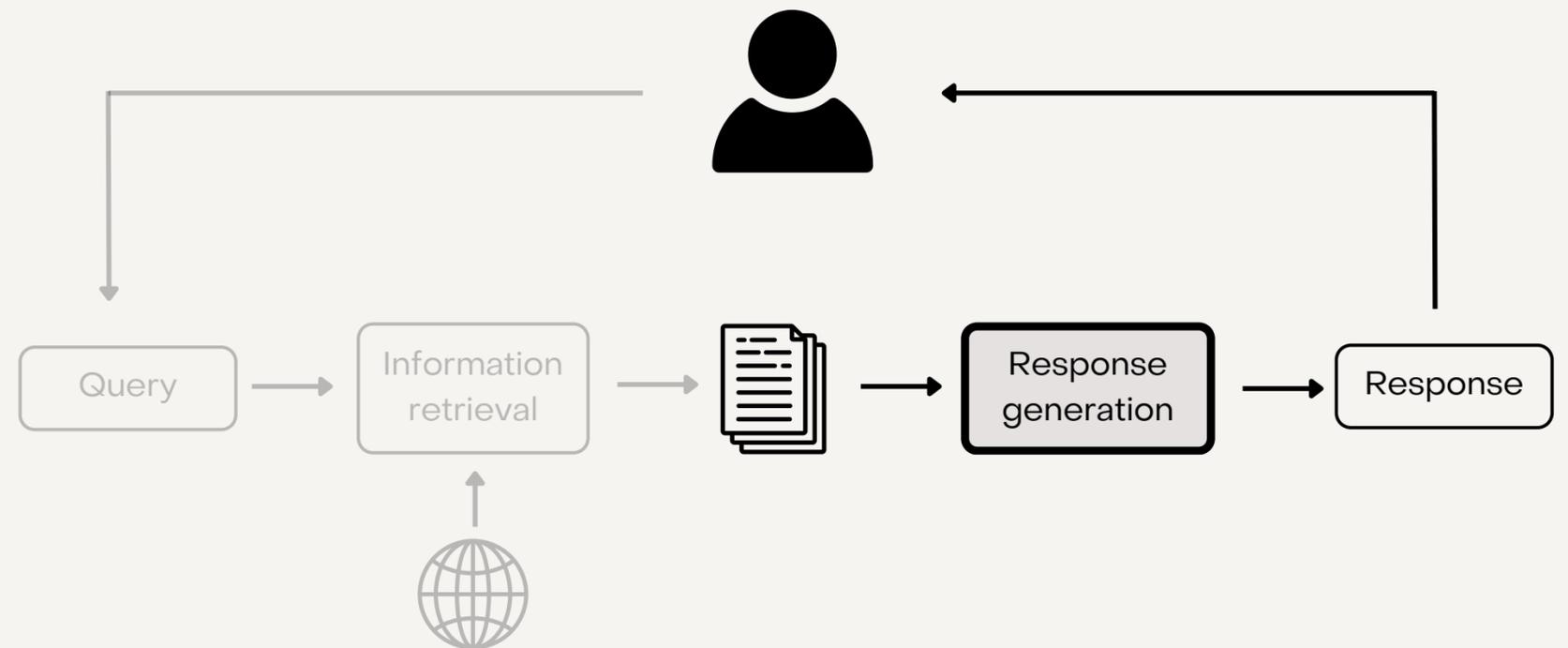


**1.** Detecting specific facts answering the question

**2.** Detecting response limitations

**3.** Synthesizing most important facts into a final response

**4.** Providing users with transparent response

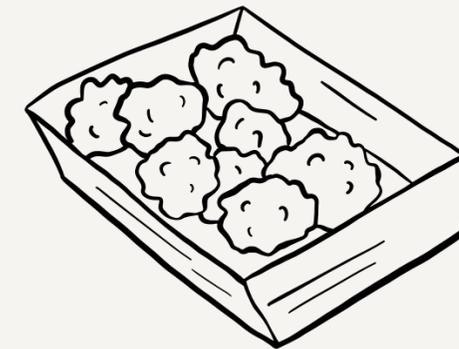


# Detecting Information Nuggets



Crowdsourcing task design and protocol to collect high-quality snippet-level annotations of relevant passages, to enable:

- 1) training of response generation models that are able to ground answers in actual statements
- 2) automatic evaluation of the generated responses in terms of completeness



Łajewska, W. and Balog, K. (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.

# Information Nuggets – Example



Information nugget is a minimal, atomic units of relevant information

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

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 & gmunoزابوگابیر 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.

# *CAsT-snippets* Dataset

## 1. Qualification Task

Task consisted of: a detailed description of the problem, examples of correct annotations, a quiz, and 10 query-passage pairs to be annotated

20 workers completed/15 passed

Initial guidelines

## 2. Discussion

Feedback on qualification task

Extended guidelines

## 3. Data Collection

Performed in daily batches (1 topic/batch ≈46 HITs)

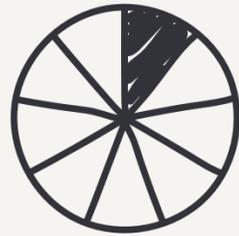
Individual feedback after each submitted batch

General comments/suggestions on a common Slack channel

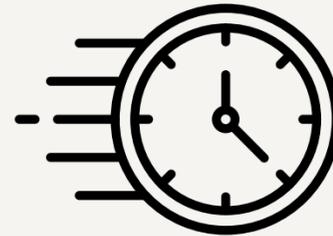
\$0.3 per HIT +\$2 bonus for completing within 24h

**371 queries, top 5 passages per query ⇒ 1855 query-passage pairs (each annotated by 3 crowd workers)**

# Information Nuggets Detection – Challenges



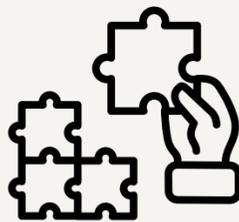
Handling snippets that only contain a partial answer to the query



Temporal validity of snippets and time constraints indicated in the query



Subjectivity of the passages originating from blogs or comments



Indirect answers that require reasoning and background knowledge



Determining the appropriate amount of context to include in each snippet



Including evidence or additional information to make the snippet valid

# Addressing CIS Limitations



How to generate grounded and transparent responses?

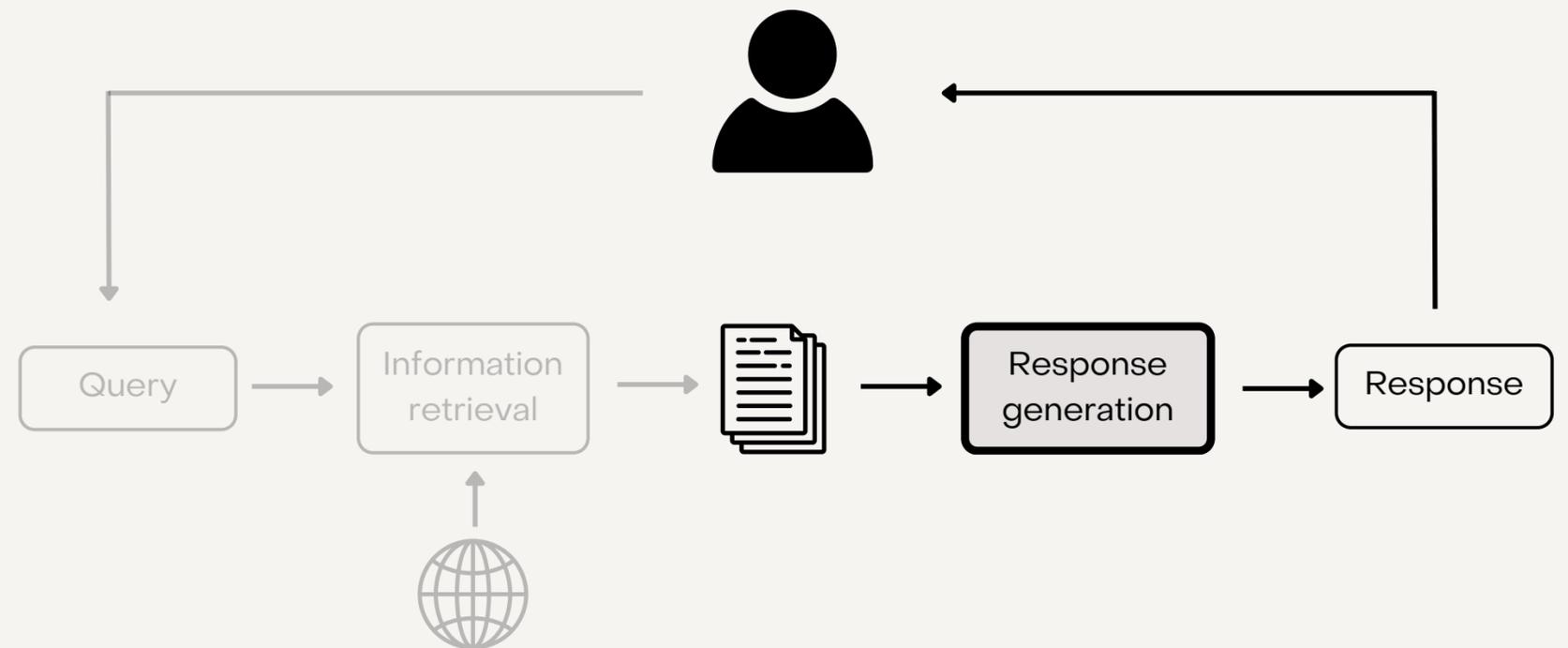


1. Detecting specific facts answering the question

2. Detecting response limitations

3. Synthesizing most important facts into a final response

4. Providing users with transparent response



# Detecting Response Limitations – Unanswerability



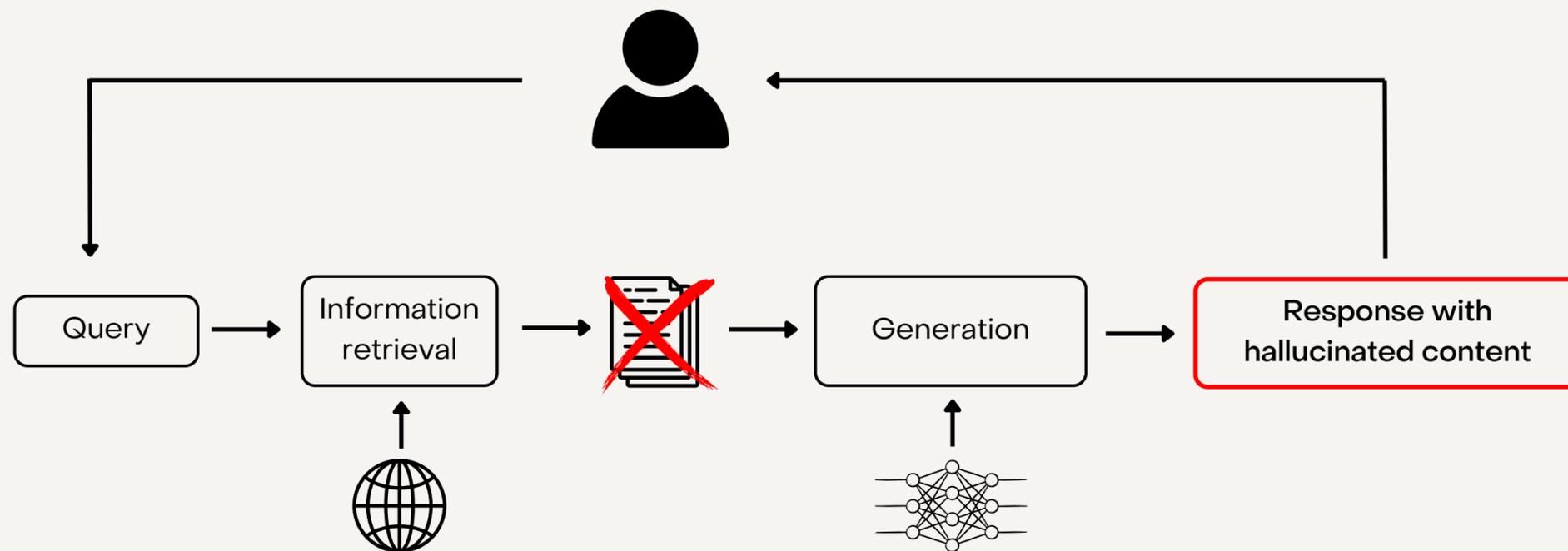
Mechanism for detecting unanswerable questions for which the correct answer is not present in the corpus or could not be retrieved including:

- 1) *CAsT-snippets* dataset extended with answerability labels on the sentence, passage, and ranking levels
- 2) a baseline approach for predicting answerability based on the top retrieved results



Łajewska, W. and Balog, K. (2024). *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*.

# Answerability Prediction – Motivation



Answerability extends beyond the general notion of relevance and asks for the presence of a specific answer

The answer to the user's question may not always be contained in the top retrieved passage

Response generated from passages not containing the answer may result in hallucinations



# CAsT-answerability Dataset

← Answerability →

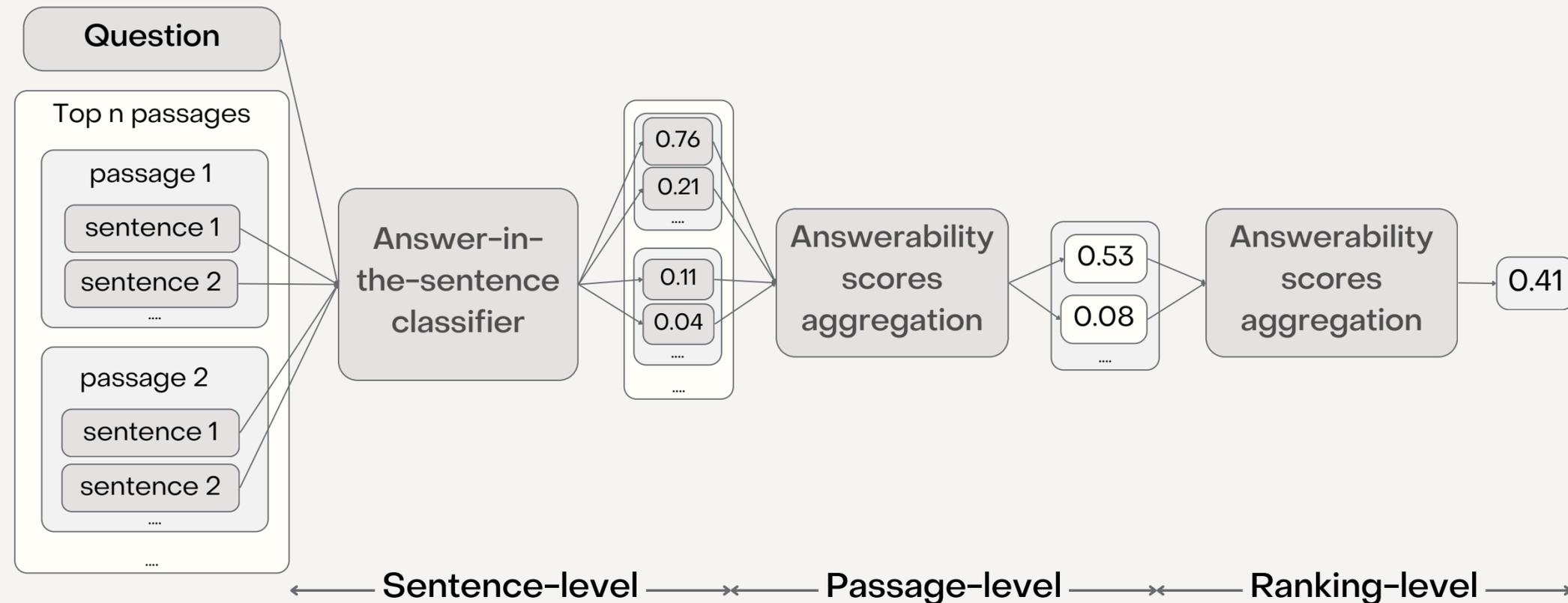
	Sentence	Passage	Ranking
<b>What's important for me to know about the safety of smart garage door openers?</b>			
<b>MARCO_7107975</b>			
If you're looking to get a little more creative with your smart home ...	0	1	1
Echo can connect with this device to tell you if you've left your ...	1		
You can even say, Alexa, tell Garageio to close my garage door, and ...	1		
<b>MARCO_8270733</b>			
The Good The Chamberlain MyQ Garage is one of the most ...	0	0	
The Bad It works with a growing list of other smart home products, ...	0		
The Bottom Line Chamberlain's MyQ Garage should be the first on ...	0		

Answerability labels on three levels: 1) sentence, 2) passage, and 3) ranking

For ranking-level answerability all possible 3-element subsets of passages available for this question are considered

5 relevant (based on snippet-level answer annotations CAsT-snippets dataset) and 5 non-relevant passages for each query

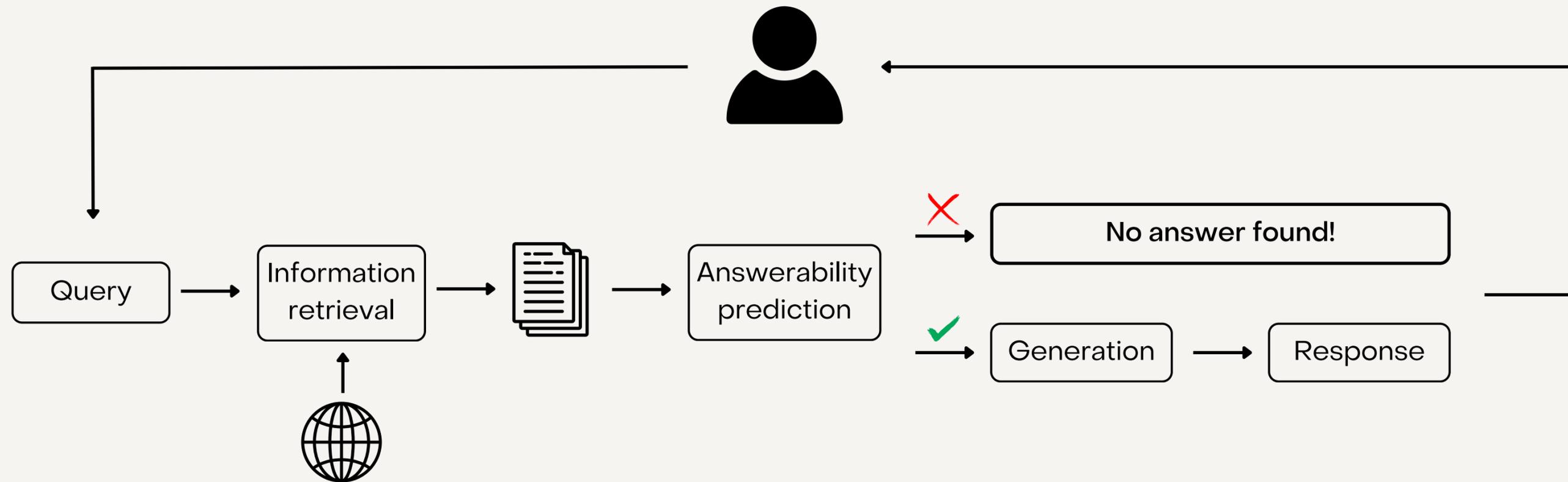
# Answerability Prediction Baseline



Max aggregation on the passage level followed by *mean* aggregation on the ranking level gives the best results

LLMs have a limited ability to detect answerability without additional guidance

# Query Answerability Disclosure



# Addressing CIS Limitations



How to generate grounded and transparent responses?

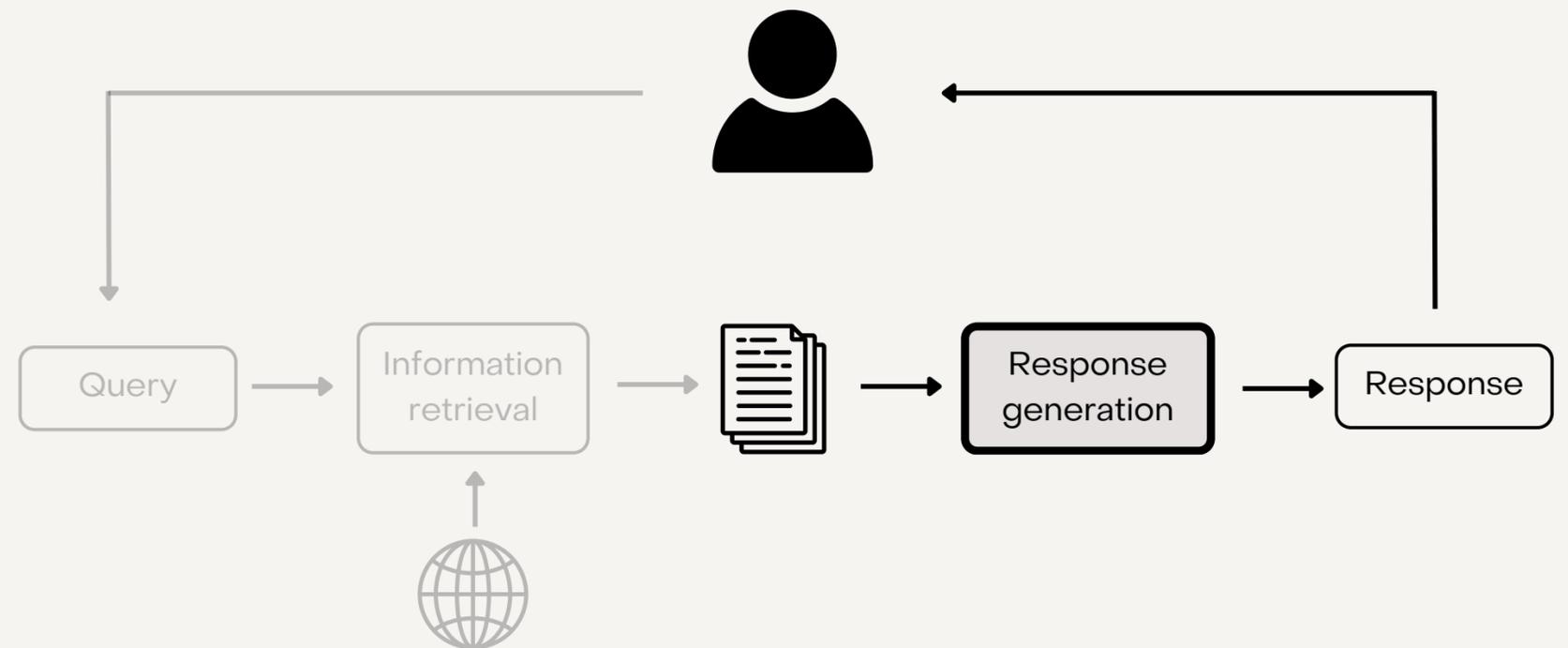


1. Detecting specific facts answering the question

2. Detecting response limitations

3. Synthesizing most important facts into a final response

4. Providing users with transparent response

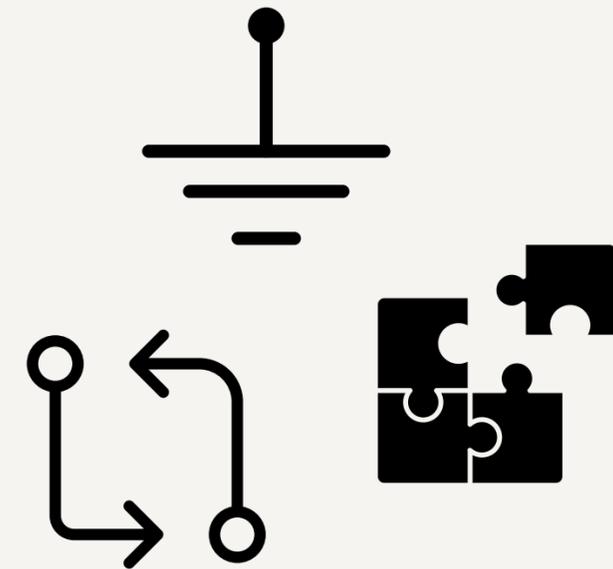




# Grounded Response Generation

We propose GINGER, Grounded Information Nugget-based Generation of Conversational Information-Seeking Responses, that:

- 1) ensures response grounding and source verification
- 2) controls response completeness
- 3) suggests relevant and answerable follow-up questions





# GINGER Response

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



**Passage 1:** On May 26, 2016, Louisiana was the first state to add police officers and firefighters to their state hate crime statute, when Governor John Bel Edwards signed an amendment from the legislature into law. This amendment was added, in part, as a response to the Black Lives Matter movement, which seeks to end police brutality against black people, with some advocates of the amendment using the slogan 'Blue Lives Matter'. Since the inception of Black Lives Matter, critics have found some of the movement's rhetoric anti-police, with the author of the amendment, Lance Harris, stating some 'were employing a deliberate campaign to terrorize our officers'. Despite the killing of a Texas sheriff in 2015 and the killings of two NYPD officers in the previous year, in response to the death of Eric Garner and the shooting of Michael Brown, there was little to no data suggesting hate crimes against law enforcement were a common problem when the bill was passed. A little less than two months after the amendment was passed, Baton Rouge was in the national spotlight after the Baton Rouge Police killing of Alton Sterling by two white police officers. This sparked protests in Baton Rouge, resulting in hundreds of arrests and increased racial tension nationally. In the week during those protests, five police officers were killed in Dallas, and the week after the protests, three more officers were killed in Baton Rouge. Both perpetrators were killed and the motives behind both shootings were responses to the recent police killings by police officers of black men.

**Passage 2:** 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 ...

**Passage 3:** White Lives Matter is an activist group created in response to Black Lives Matter. In August 2016, the Southern Poverty Law Center added 'White Lives Matter' to its list of hate groups. The group ...



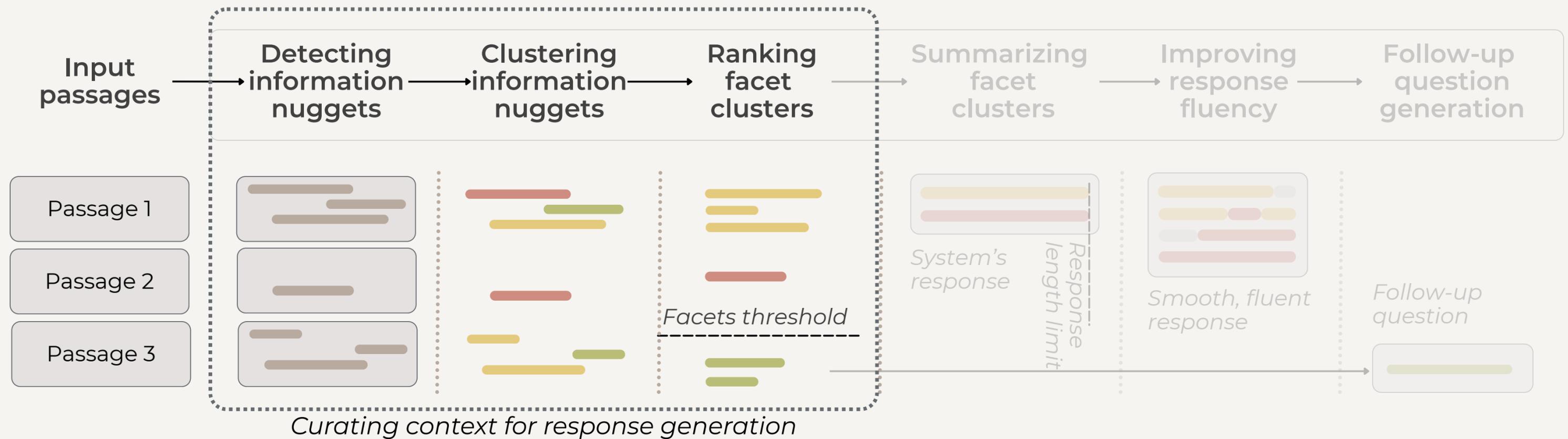
**Response:** The response to the Black Lives Matter movement in the US included the addition of an amendment and the emergence of Blue Lives Matter and All Lives Matter, movements supported by advocates of the police. The 'White Lives Matter' group was designated as a hate group by the Southern Poverty Law Center [P3], underscoring the diverse views on these movements among various races. During a wave of protests, five Dallas police officers were killed within a week, with another three officers killed in Baton Rouge the following week [P1]. Moreover, Black Lights Matter came under scrutiny for perceived anti-police statements [P1] and misrepresenting the objectives of Black Lives Matter.



**Follow-up question:** Do you want to learn more about how these arrests have influenced public perception and policy changes regarding racial issues in the US?



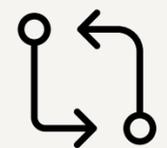
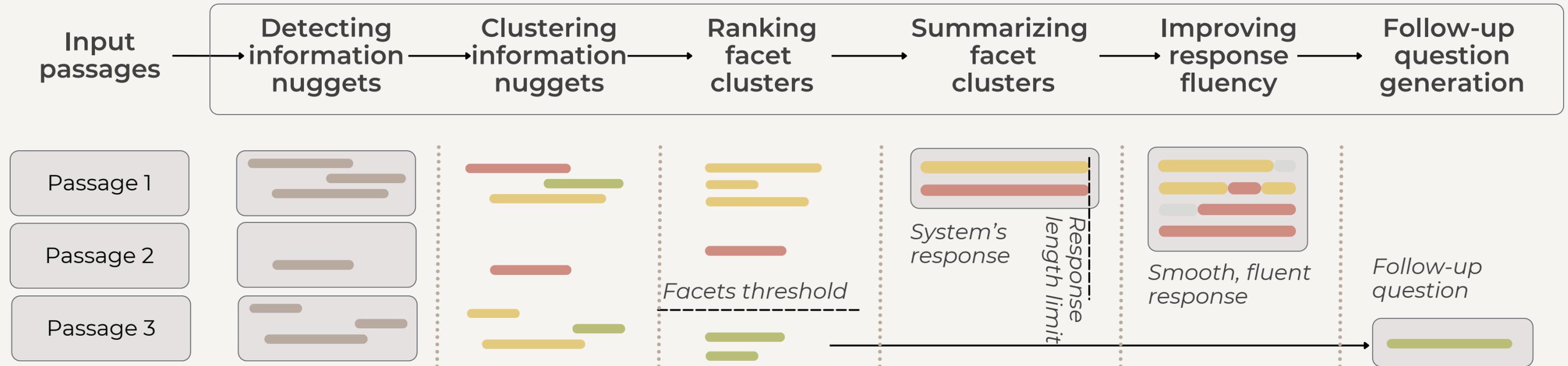
# GINGER Pipeline



Context curation employed in GINGER ensures that the most relevant information from retrieved passages is extracted from the sources, allowing summarization to operate on a shorter but more relevant context



# GINGER Pipeline



Conditioning follow-up question generation on information nuggets ensures relevance and answerability of the questions



By operating on information nuggets in all intermediate components of the pipeline we ensure the grounding of the final response in the source passages



# GINGER Evaluation

<i>Approach</i>	<i>V_strict</i>
Retrieval-augmented open-domain QA	0.442
Chain-of-Thought prompting with in-context learning examples	0.428
<b>GINGER (top 20 passages)</b>	<b>0.568 (+29%)</b>
GINGER (top 5 passages)	0.362 (-18%)
GINGER (top 20 passages, BM25)	0.554 (+25%)

## Key advantages of GINGER:

- maximizing information within response length limits
- providing source attribution for verifiability
- guiding users with relevant follow-up questions
- allowing control over response completeness

Ablation study shows that the main contributing factor to GINGER's performance is related to operating on information nuggets, as opposed to the effectiveness of individual components

# Addressing CIS Limitations



How to generate grounded and transparent responses?

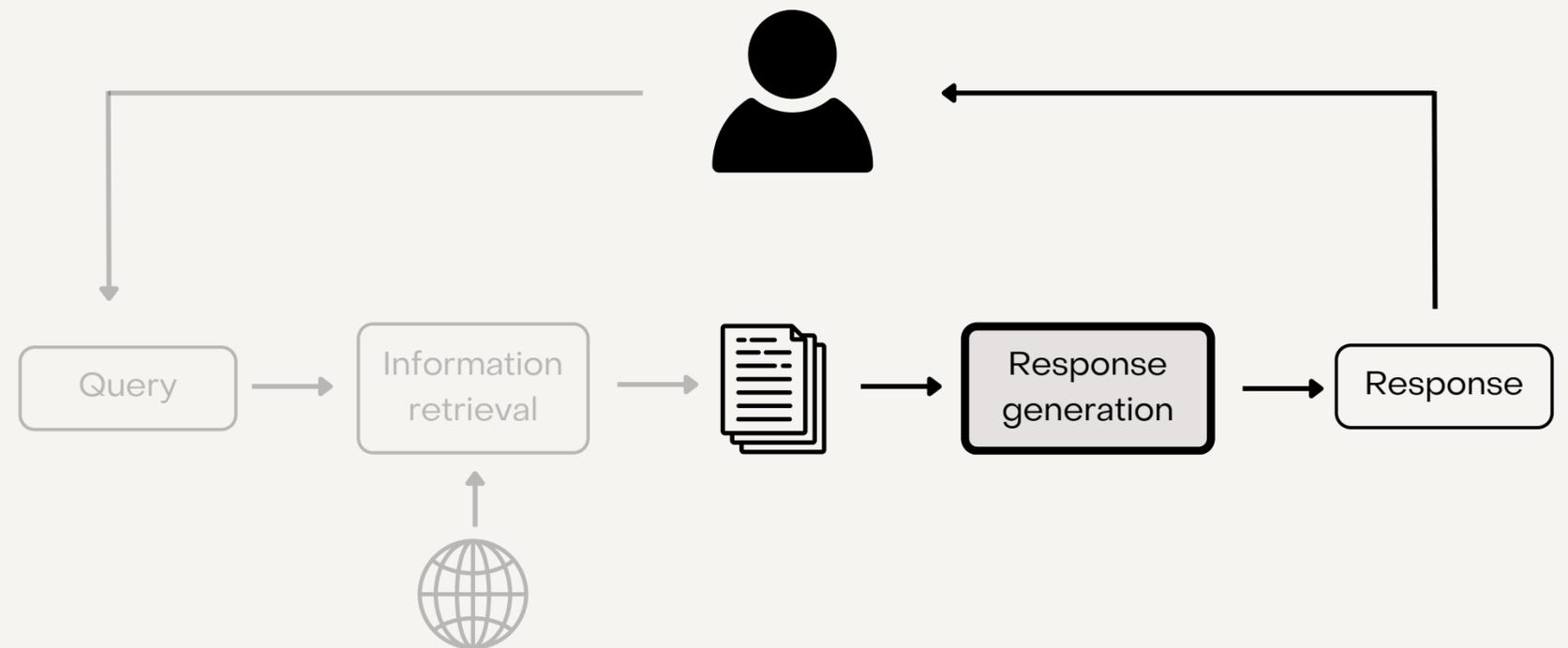


1. Detecting specific facts answering the question

2. Detecting response limitations

3. Synthesizing most important facts into a final response

4. Providing users with transparent response



# Transparent Response Generation



We run a user study exploring explanations of source, confidence, and limitations in CIS dialogues:

- 1) we use manually curated dataset of responses and explanations, with noise incorporated in a controlled manner
- 2) we analyse the impact of noise and the presentation mode of the explanations on response usefulness

**USER:** What was the US reaction to the Black Lives Matter movement?



**LLM:** 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.



[Sources](#)

System confidence



Controversial topic, multiple points of view possible!



# Variants of Explainable Responses

## Explanations Quality

Accurate or noisy explanation including:

- 1) The origin of presented information
- 2) The system's confidence
- 3) Potential limitations of the generated response

## Response Quality

- 1) Ground-truth response
- 2) Imperfect response with biases or factual errors

## Explanations Presentation

- 1) Textual explanations
- 2) Visual explanations

**Each crowdworker is asked to assess responses for 10 queries that differ in their quality and may be enhanced with explanations (differing in terms of quality and presentation mode)**

# Evaluation of Responses

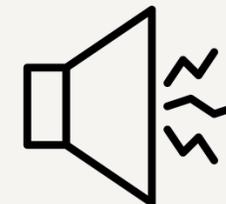
## Response Dimensions

Usefulness  
Relevance  
Correctness  
Completeness  
Comprehensiveness  
Conciseness  
Serendipity  
Coherence  
Factuality  
Fairness  
Readability  
Satisfaction

Users are not able to identify bias towards one specific point of view or factual errors without expert knowledge about the topic



Low-quality explanations have a strong impact on user experience and decrease the user-perceived usefulness of the response



# Evaluation of Explanations

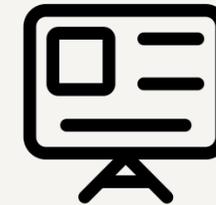
## Explanations Quality

To what extent were the provided responses supported?

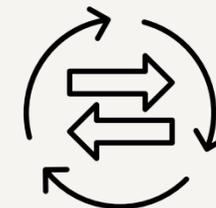
To what extent did the assistant help you realize the potential limitations of the responses?

To what extent are you aware of the assistant's confidence in the provided responses?

The presentation mode of explanations is not a critical factor in this setting



Not providing explanations is more useful than providing noisy ones (user gain and effort trade-off)



# Conclusions

---

# Conclusions



We propose a response generation pipeline that:



**1.** Detects specific facts answering the question

**2.** Detects response limitations

**3.** Synthesizes most important facts into a final response

**4.** Provides users with transparent response

## Limitations

01  
10

Nugget identification and answerability detection as binary tasks



Restricted Scope of Answerability



Limitations of User Studies



Single turn in a conversation vs. the conversation as a whole

● Thank you!

---

Weronika Łajewska  
weronikalajewska@gmail.com