

Grounded and Transparent Response Generation for Conversational Information-Seeking Systems

Weronika Łajewska

Supervised by Krisztian Balog

University of Stavanger, Norway

Conversational Information-Seeking (CIS) Systems

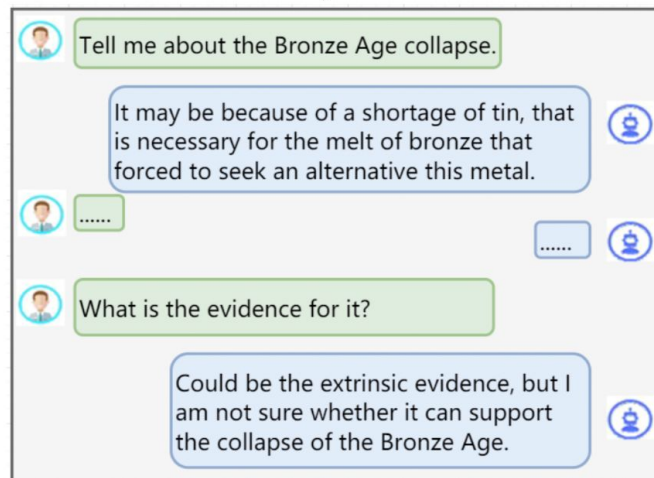
what is the evidence for bronze age collapse?



Search Engine Results Page
(SERP) based System



Conversational
Search System



Problems with CIS Systems

- 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



Grounded and Transparent Response Generation

for Conversational Information-Seeking Systems

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

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

Sources

▼ Sources

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

https://en.wikipedia.org/wiki/Blue_Lives_Matter#:~:text=History,-

[A%20golf%20cart&text=On%20December%2020%2C%202014%2C%20in,and%20retire](https://en.wikipedia.org/wiki/Blue_Lives_Matter#:~:text=History,-A%20golf%20cart&text=On%20December%2020%2C%202014%2C%20in,and%20retire)

Limitations + system confidence

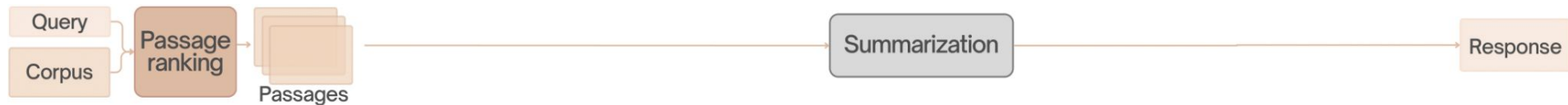
Assistants's confidence in the response:



Controversial topic, multiple viewpoints possible, only some discussed!

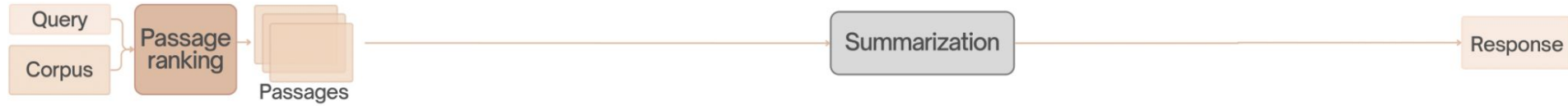
Overview of our Approach to Conversational Response Generation

*"A true teacher would never tell you what to do. But he would give you the knowledge with which you could decide what would be best for you to do."
— Christopher Pike, Sati*



Overview of our Approach to Conversational Response Generation

"A true teacher would never tell you what to do. But he would give you the knowledge with which you could decide what would be best for you to do."
— Christopher Pike, Sati

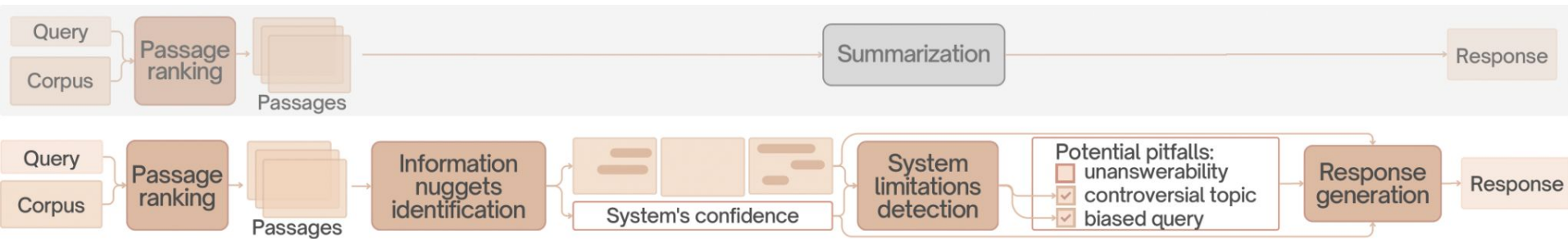


What can go wrong?

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

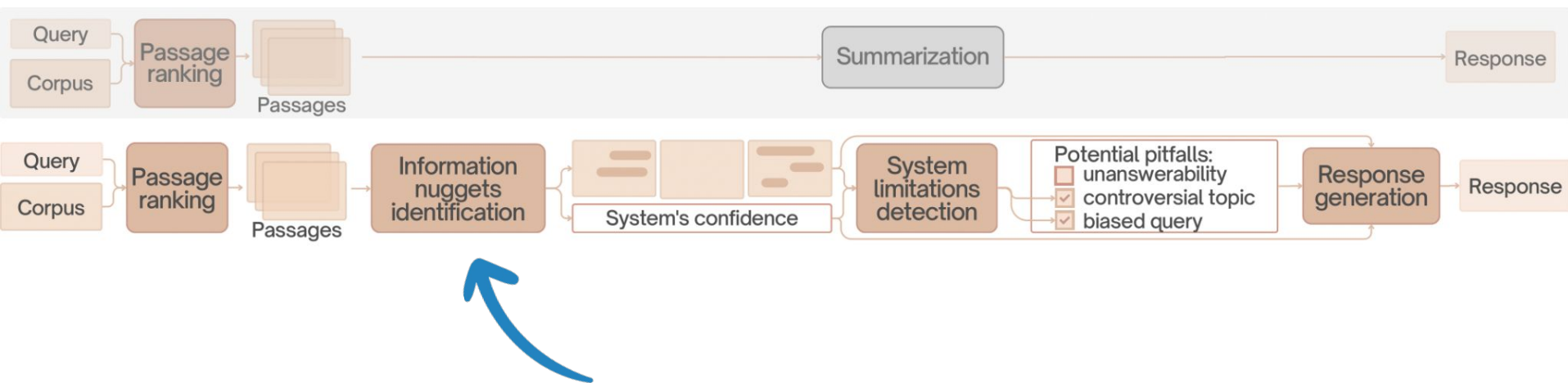
Overview of our Approach to Conversational Response Generation

"A true teacher would never tell you what to do. But he would give you the knowledge with which you could decide what would be best for you to do."
— Christopher Pike, Sati



Overview of our Approach to Conversational Response Generation

"A true teacher would never tell you what to do. But he would give you the knowledge with which you could decide what would be best for you to do."
— Christopher Pike, Sati



Information Nuggets Identification

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

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

Sources

▼ Sources

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

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

Limitations + system confidence

Assistant's confidence in the response:



Controversial topic, multiple viewpoints possible, only some discussed!

Dataset with Information Nuggets

- **Problem setting:** Conversational response generation
 - It extends beyond passage retrieval + summarization
- **Goal:** snippet-level annotations of relevant passages, to enable
 1. the training of response generation models that are able to ground answers in actual statements
 2. the automatic evaluation of the generated responses in terms of completeness
- **Main contributions:**
 1. Crowdsourcing task design and protocol to collect high-quality annotations
 2. A CAST-snippets dataset of 1.8k query-passage pairs annotated from the TREC 2020 and 2022 Conversational Assistance track

CAsT-snippets Sample

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

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

CAsT-snippets Dataset Summary

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

- Data quality
 - Inter-annotator agreement exceeds even that of expert annotators
- Comparison against other datasets
 - More snippets annotated per input text; also, snippets are longer

Dataset	Input text	Avg. snippets length (tokens)	# snippets per annotation
CAsT-snippets	Paragraph	39.6	2.3
SaaC [1]	Top 10 passages	23.8	1.5
QuaC [2]	Wikipedia article	14.6	1

[1] Pengjie Ren, Zhumin Chen, Zhaochun Ren, E. Kanoulas, Christof Monz, and M. de Rijke. 2021. Conversations with Search Engines: SERP-based Conversational Response Generation. ACM Transactions on Information Systems 39, 4 (2021), 1-29

[2] 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 Findings of the Association for Computational Linguistics: EMNLP '18, 2174-2184.

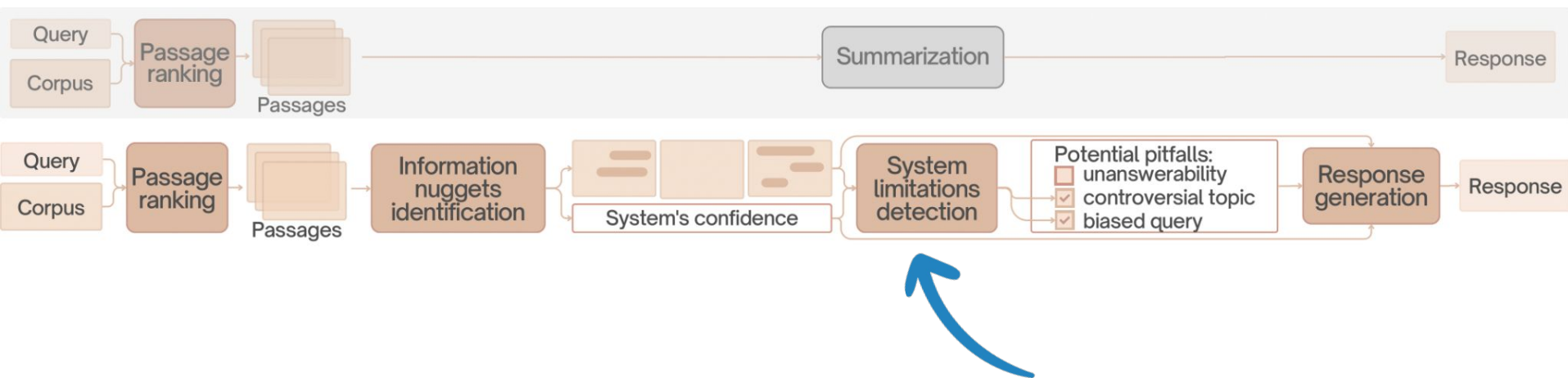
Challenges Identified

Challenges pointed out by the crowd workers that need to be addressed in conversational response generation:

- Only a partial answer is present
- Temporal considerations
 - Spans may need to be excluded given the time constraints in the query
 - Assessing temporal validity can be challenging based on the paragraph alone (without larger context)
- 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 span
 - Balancing between being concise and being self-contained
- Determining whether the evidence or additional information is needed or an entity alone is sufficient as an answer

Overview of our Approach to Conversational Response Generation

"A true teacher would never tell you what to do. But he would give you the knowledge with which you could decide what would be best for you to do."
— Christopher Pike, Sati



System Limitation Detection

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

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

Sources

▼ Sources

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

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

Limitations + system confidence

Assistants's confidence in the response:



Controversial topic, multiple viewpoints possible, only some discussed!



Response Generation Challenges in CIS

- **Problem setting:** Response generation in conversational information-seeking (CIS) scenario
- **Goal:** investigating the ability of users to recognize pitfalls in CIS responses
- **Research questions:**
 1. Can users effectively recognize the problem of query answerability and the problem of multiple viewpoints leading to response incompleteness in system responses?
 2. How do inaccurate, incomplete, and/or biased responses impact user experience?
- **Main contribution:**
 1. A novel methodology to study how users perceive query answerability and response incompleteness in CIS

Answerability Study

Query: 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?

Experimental Condition	Response Dimension		CIS System Response	Source
	Factual Corr.	Source		
Factually correct + valid source	✓	✓	<i>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...</i>	https://wanderingtrader.com/argentina/top-5-argentina-tourist-attractions/
Factually correct + no source	✓	✗	Same as above	–
Factually incorrect + invalid source	✗	✓ (invalid)	<i>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...</i>	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.)
Factually incorrect + no source	✗	✗	Same as above	–

Viewpoints Study

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

Experimental Condition	Response Dimension		CIS System Response
	Diversity	Balance	
Multiple viewpoints covered to the same extent	✓	✓	<i>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.</i>
Multiple viewpoints mentioned, one covered in detail	✓	✗	<i>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.</i>
Single viewpoint mentioned and covered	✗	✗	<i>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.</i>

Findings

- **Quantitative analysis** → users find it easier to identify problems with diversity and balance of viewpoints rather than factual errors and source validity in the responses
- **Analysis of a user experience** → self-reported overall satisfaction scores are not necessarily associated with the main response dimensions
- **Qualitative analysis of free-text comments** → credibility of the sources, as well as completeness, usefulness, and subjectivity of provided information impact the overall satisfaction of the users

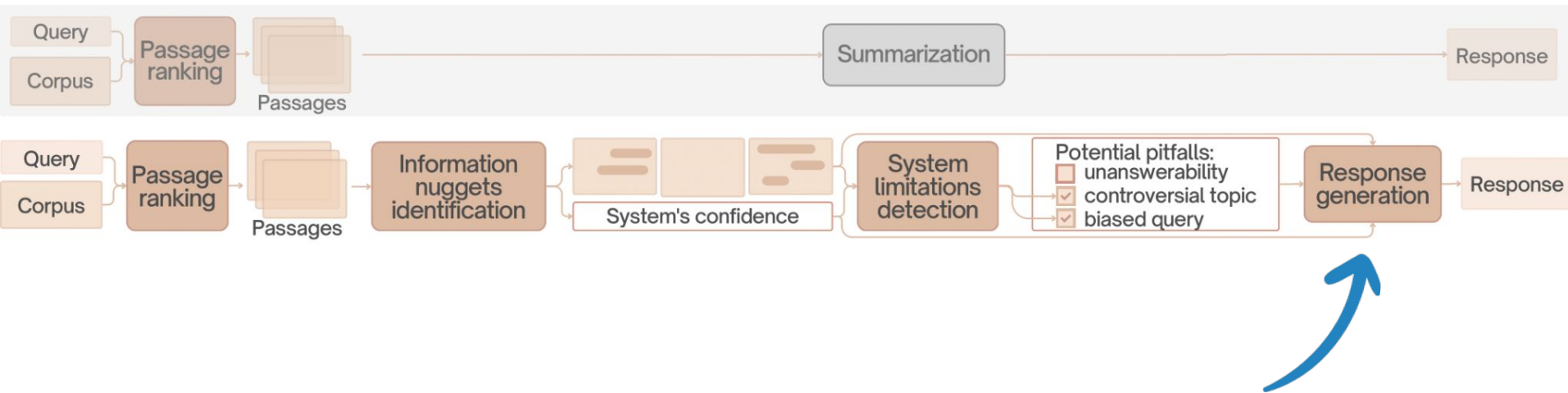
Dependent Variable	Independent Variable(s)	F	p-value	Unbalanced Estimator	Effect Size
<i>Answerability Study</i>					
Factual Correctness	Experimental Condition	1.330	0.264	0.003	-
Confidence in Answer Accuracy		0.721	0.540	-0.002	-
Overall Satisfaction		1.190	0.313	0.002	-
<i>Viewpoints Study</i>					
Diversity		31.774	0.000	0.186	L
Transparency	Experimental	21.751	0.000	0.133	M
Balance	Condition	17.514	0.000	0.109	M
Overall Satisfaction		17.687	0.000	0.110	M

Dependent Variable	Explanatory Variables	p-value
<i>Answerability Study</i>		
Overall Satisfaction	Familiarity	0.248
	Factual Correctness	0.069
	Confidence in Answer Accuracy	0.012
<i>Viewpoints Study</i>		
Overall Satisfaction	Familiarity	0.972
	Diversity	0.209
	Transparency	0.436
	Balance	0.003

⇒ simple source attribution is not enough to ensure effective interaction with the system

Overview of our Approach to Conversational Response Generation

*"A true teacher would never tell you what to do. But he would give you the knowledge with which you could decide what would be best for you to do."
— Christopher Pike, Sati*



Generation of Response Communicating Limitations

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

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

Sources

▼ Sources

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

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

Limitations + system confidence

Assistants's confidence in the response:



Controversial topic, multiple viewpoints possible, only some discussed!



Explainability for Transparent CIS



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

Research questions:

1. How does the quality of responses and explanations affect user-perceived response usefulness?
2. What are effective ways to provide explanations to users?

[Source](#)

Assistant's confidence in the response



Controversial topic, multiple points of view possible!



Conclusions about Explanations in CIS

- 1) We observe lower user ratings for noisy explanations.
- 2) We observe a significant effect of familiarity with the topic on response assessment.
- 3) Our study is not conclusive about the preferred way of presenting explanations to the user.
- 4) There is an effort-gain trade-off that needs to be taken into account when enhancing the response with additional information.

Thank you for your attention!

Questions?