Explainability for Transparent Conversational Information-Seeking

Weronika Łajewska[1], Damiano Spina[2], Johanne Trippas[2], Krisztian Balog[1]

[1]University of Stavanger, Norway
[2]RMIT University, Melbourne, Australia

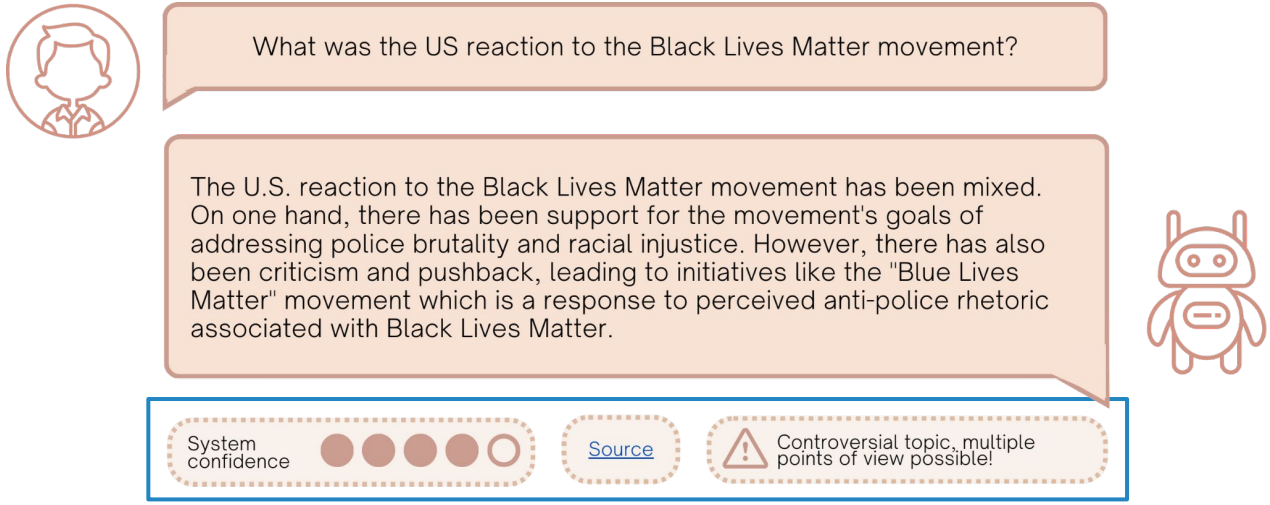
Where to find me: weronika.lajewska@uis.no, https://werlaj.github.io/

Our Motivation

"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

- Conversational search is a less transparent setting than a 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
- Recently proposed RAG systems are claimed to produce more factually correct content, but do not solve issues around transparency

This Study



- **Problem setting:** Conversational response generation
- **Goal:** Examining various ways of explaining the generated response
- **Main contributions:**
  - User study exploring explanations of source, confidence, and limitations in the Conversational Information Seeking (CIS) domain
  - Manually curated dataset of responses and explanations, with noise incorporated in a controlled manner
  - Analysis of the impact of noise and the presentation mode of the explanations on response usefulness

Explainable Response

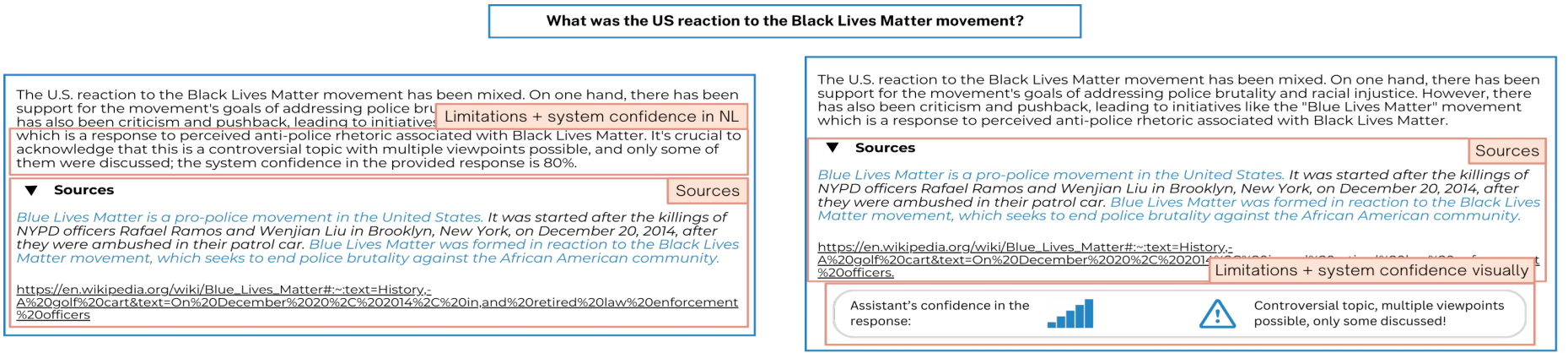
We attempt to increase the transparency of a CIS system by explaining:

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

Explanations presentation mode:

1) Textual explanation

2) Visual explanation

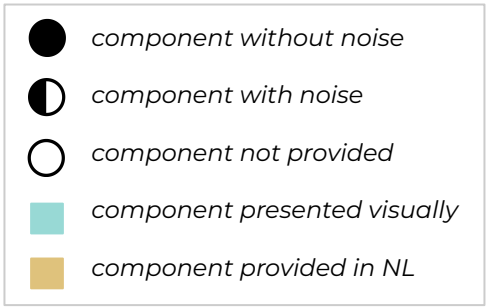


Experimental Conditions

The selected experimental conditions vary along three main dimensions:

1) **Response quality:** ground-truth response or imperfect response with biases or factual errors
2) **Quality of the explanations:** accurate or noisy explanation
3) **Presentation mode:** additional UI component or in natural language explanation

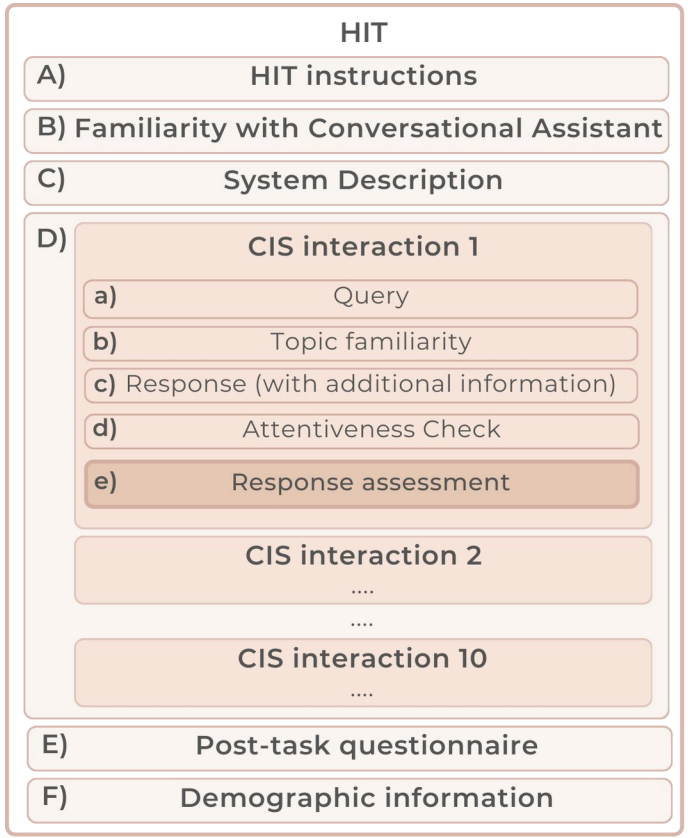
	EC1	EC2	EC3	EC4	EC5	EC6	EC7	EC8	EC9	EC10
Response	●	●	◐	◐	●	●	◐	◐	●	◐
Source	●	●	●	●	◐	◐	◐	◐	○	○
Confidence	●	●	●	●	◐	◐	◐	◐	○	○
Limitation	●	●	●	●	◐	◐	◐	◐	○	○



Noise in response → Incomplete, biased or factually incorrect response
Noise in source → Source related to the topic but not supporting the provided response
Noise in confidence score → Reverted confidence score
Noise in limitation information → Irrelevant limitation

Experimental Design

- In each human intelligence task, crowd workers are asked to assess responses for 10 queries
- Responses differ in their quality and may be enhanced with explanations
- Explanations differ in terms of quality and presentation mode
- Each HIT contains the same response variant for all ten queries, employing a between-subject design



Response Assessment Questionnaire

- Workers are asked to evaluate different dimensions of the response variant presented for a given query
- The question about each response dimension is answered by workers on a four-point Likert scale
- Questions use operational definitions of the response dimensions instead of explicit names of the dimensions

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

Variable	***Question used in the user study Conversational***
Source Explanation	*To what extent were the provided responses supported?*
Limitation Explanation	*To what extent did the assistant help you realize the potential limitations of the responses?*
Confidence Explanation	*To what extent are you aware of the assistant's confidence in the provided responses?*

Research Questions

1. Can users detect noise in the responses and explanations?
2. How does the quality of responses and explanations impact user experience?
3. What are effective ways to provide explanations to users?

Research Questions

1. **Can users detect noise in the responses and explanations?**
2. How does the quality of responses and explanations impact user experience?
3. What are effective ways to provide explanations to users?

Results

User's Perception of Response

	Usefulness	Other Dimensions										
		Rel.	Correct.	Compl.	Comprehen.	Conciseness	Serendipity	Coherence	Factuality	Fairness	Read.	Sat.
Response Quality	0.156 (S)	0.176 (S)	**0.003 (S)**	0.745 (–)	0.846 (–)	0.374 (S)	0.093 (S)	0.217 (S)	0.265 (S)	0.924 (–)	0.881 (–)	0.638 (S)

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

Results

User's Perception of Response

	Usefulness	Other Dimensions									
		Rel.	Correct.	Compl.	Comprehen.	Conciseness	Serendipity	Coherence	Factuality	Fairness	Read.	Sat.
Response Quality	0.156 (S)	0.176 (S)	**0.003 (S)**	0.745 (–)	0.846 (–)	0.374 (S)	0.093 (S)	0.217 (S)	0.265 (S)	0.924 (–)	0.881 (–)	0.638 (S)

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

- A statistically significant effect observed only on user-reported correctness of the response
- Insensitivity of user-reported response dimensions to the quality of provided information

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

Results

User's Perception of Explanations

Experiments including two conditions where explanations are not provided

	Usefulness	Other Dimensions										
		Rel.	Correct.	Compl.	Comprehen.	Conciseness	Serendipity	Coherence	Factuality	Fairness	Read.	Sat.
	All conditions (EC1–EC10)											
Explanation Quality	**0.0 (S)**	**0.0 (S)**	0.508 (S)	**0.003 (S)**	**0.0 (S)**	**0.001 (S)**	0.09 (S)	**0.002 (S)**	0.713 (–)	**0.0 (S)**	**0.032 (S)**	**0.0 (S)**
Presentation Mode	**0.019 (S)**	**0.0 (S)**	0.234 (S)	0.347 (S)	0.658 (–)	**0.001 (S)**	0.149 (S)	0.09 (S)	0.842 (–)	**0.001 (S)**	0.651 (–)	**0.0 (S)**
	Only conditions with explanations (EC1–EC8)											
Explanation Quality	**0.0 (S)**	**0.006 (S)**	0.256 (S)	**0.002 (S)**	**0.0 (S)**	0.122 (S)	0.319 (S)	**0.003 (S)**	0.504 (S)	**0.0 (S)**	**0.014 (S)**	**0.007 (S)**
Presentation Mode	0.872 (–)	0.686 (–)	0.096 (S)	0.895 (–)	0.38 (S)	0.399 (S)	0.86 (–)	0.377 (S)	0.739 (–)	0.78 (–)	0.771 (–)	0.071 (S)

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

Experiments only including the conditions where explanations are provided

Results

User's Perception of Explanations

	Usefulness	Other Dimensions									
		Rel.	Correct.	Compl.	Comprehen.	Conciseness	Serendipity	Coherence	Factuality	Fairness	Read.	Sat.
	All conditions (EC1–EC10)											
Explanation Quality	**0.0 (S)**	**0.0 (S)**	0.508 (S)	**0.003 (S)**	**0.0 (S)**	**0.001 (S)**	0.09 (S)	**0.002 (S)**	0.713 (–)	**0.0 (S)**	**0.032 (S)**	**0.0 (S)**
Presentation Mode	**0.019 (S)**	**0.0 (S)**	0.234 (S)	0.347 (S)	0.658 (–)	**0.001 (S)**	0.149 (S)	0.09 (S)	0.842 (–)	**0.001 (S)**	0.651 (–)	**0.0 (S)**
	Only conditions with explanations (EC1–EC8)											
Explanation Quality	**0.0 (S)**	**0.006 (S)**	0.256 (S)	**0.002 (S)**	**0.0 (S)**	0.122 (S)	0.319 (S)	**0.003 (S)**	0.504 (S)	**0.0 (S)**	**0.014 (S)**	**0.007 (S)**
Presentation Mode	0.872 (–)	0.686 (–)	0.096 (S)	0.895 (–)	0.38 (S)	0.399 (S)	0.86 (–)	0.377 (S)	0.739 (–)	0.78 (–)	0.771 (–)	0.071 (S)

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

Introducing noise in explanations has a statistically significant effect on almost all user-reported response dimensions → **Noisy explanations have a strong impact on user experience in general**

Results

User's Perception of Explanations

	Usefulness	Other Dimensions									
		Rel.	Correct.	Compl.	Comprehen.	Conciseness	Serendipity	Coherence	Factuality	Fairness	Read.	Sat.
	All conditions (EC1–EC10)											
Explanation Quality	**0.0 (S)**	**0.0 (S)**	0.508 (S)	**0.003 (S)**	**0.0 (S)**	**0.001 (S)**	0.09 (S)	**0.002 (S)**	0.713 (–)	**0.0 (S)**	**0.032 (S)**	**0.0 (S)**
Presentation Mode	**0.019 (S)**	**0.0 (S)**	0.234 (S)	0.347 (S)	0.658 (–)	**0.001 (S)**	0.149 (S)	0.09 (S)	0.842 (–)	**0.001 (S)**	0.651 (–)	**0.0 (S)**
	Only conditions with explanations (EC1–EC8)											
Explanation Quality	**0.0 (S)**	**0.006 (S)**	0.256 (S)	**0.002 (S)**	**0.0 (S)**	0.122 (S)	0.319 (S)	**0.003 (S)**	0.504 (S)	**0.0 (S)**	**0.014 (S)**	**0.007 (S)**
Presentation Mode	0.872 (–)	0.686 (–)	0.096 (S)	0.895 (–)	0.38 (S)	0.399 (S)	0.86 (–)	0.377 (S)	0.739 (–)	0.78 (–)	0.771 (–)	0.071 (S)

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

Introducing noise in explanations has a statistically significant effect on almost all user-reported response dimensions → **Noisy explanations have a strong impact on user experience in general**

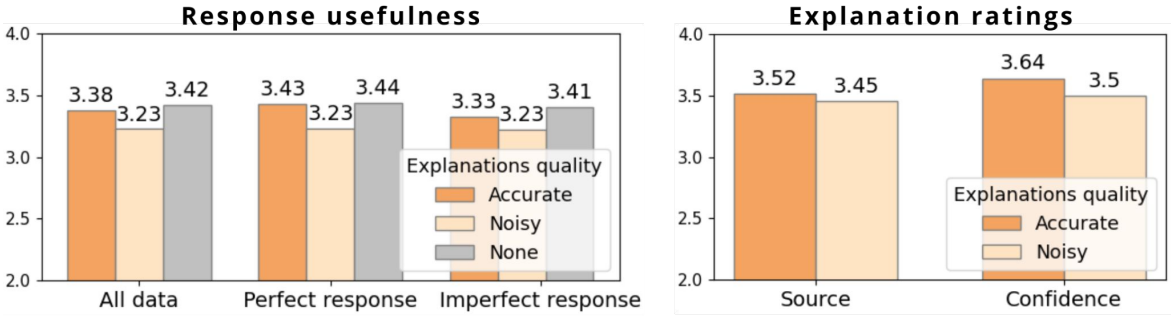
Response dimensions are insensitive to the way explanations are presented

Research Questions

1. Can users detect noise in the responses and explanations?
2. **How does the quality of responses and explanations impact user experience?**
3. What are effective ways to provide explanations to users?

Results

Effect of the Explanation Quality



Mean scores for response usefulness and explanation ratings for different quality of the explanations. All differences between the ratings within a given plot are statistically significant.

Results

Effect of the Explanation Quality

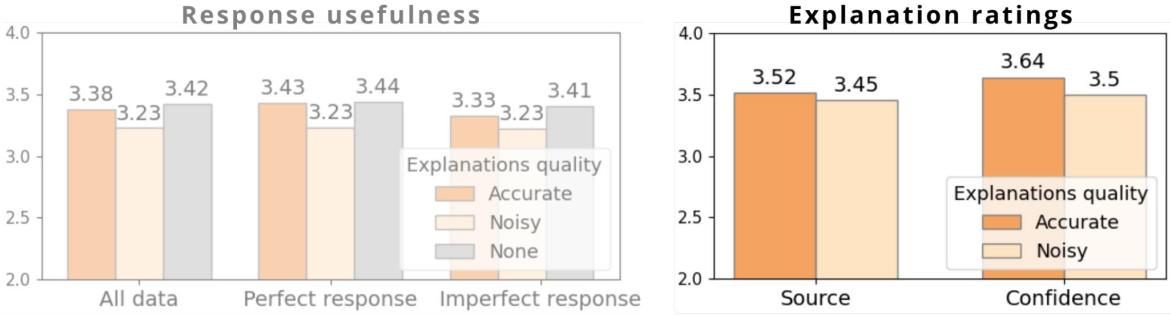
Mean scores for response usefulness and explanation ratings for different quality of the explanations. All differences between the ratings within a given plot are statistically significant.

High-quality source, system confidence score, and information about the response limitations make the response more useful from the user's perspective

The explanations either pollute the response or make the user more critical about it, in both cases resulting in reduced usefulness → **Not providing explanations is more useful than providing noisy ones**

Results

Effect of the Explanation Quality



Mean scores for response usefulness and explanation ratings for different quality of the explanations. All differences between the ratings within a given plot are statistically significant.

High-quality source, system confidence score, and information about the response limitations make the response more useful from the user's perspective

The explanations either pollute the response or make the user more critical about it, in both cases resulting in reduced usefulness → **Not providing explanations is more useful than providing noisy ones**

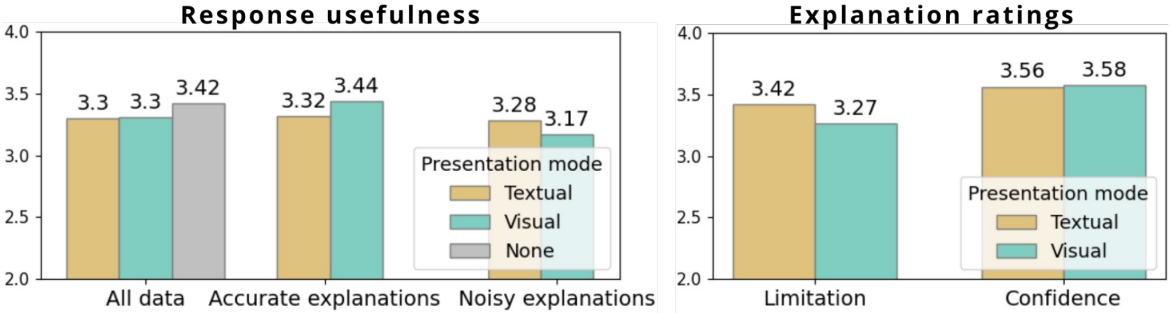
Users perceive noisy explanations as less useful in understanding system confidence and attributed sources

Research Questions

1. Can users detect noise in the responses and explanations?
2. How does the quality of responses and explanations impact user experience?
3. **What are effective ways to provide explanations to users?**

Results

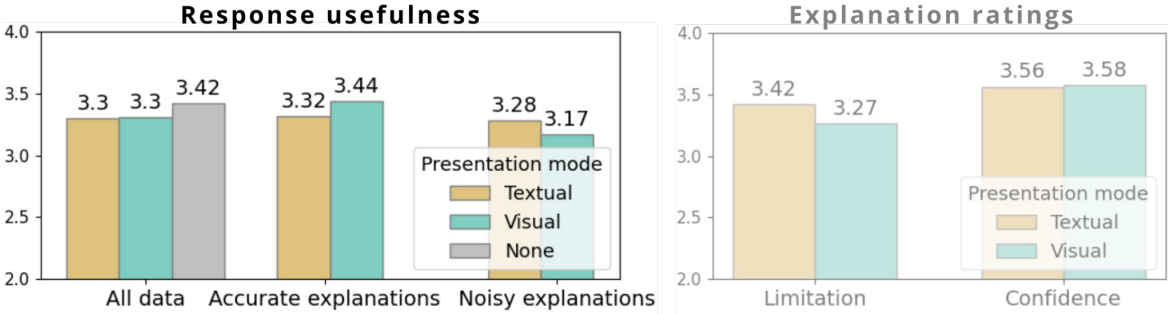
Effect of the Presentation Mode



Mean scores for response usefulness and explanation ratings for presentation modes. All differences between the ratings within a given plot are statistically significant.

Results

Effect of the Presentation Mode

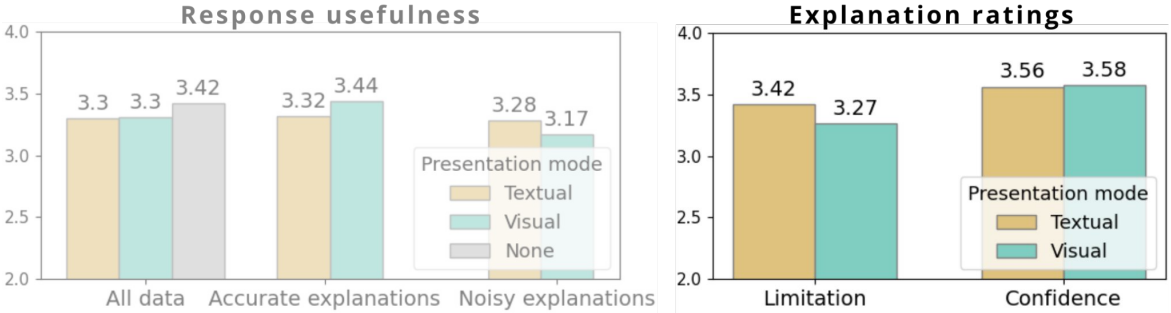


Mean scores for response usefulness and explanation ratings for presentation modes. All differences between the ratings within a given plot are statistically significant.

The critical decision lies not in the method of presenting information but rather in determining whether the explanations are necessary → **Trade-off between effort and gain**

Results

Effect of the Presentation Mode



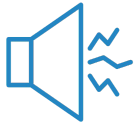
Mean scores for response usefulness and explanation ratings for presentation modes. All differences between the ratings within a given plot are statistically significant.

The critical decision lies not in the method of presenting information but rather in determining whether the explanations are necessary → **Trade-off between effort and gain**

The preferred presentation mode depends on explanations quality and the explained aspect of the response

Summary

- Manually curated dataset of responses and explanations, with noise incorporated in a controlled manner
- Analysis of the effect of noise and different presentation modes of the explanations on users' assessments of responses and explanations:



Low-quality explanations decrease the user-perceived usefulness of the response

Users are not able to detect factual errors or biases in the provided information

The format of explanations is not a critical factor in this setting

User gain and effort trade-off (no explanations are better than noisy ones)

Future work:

- Investigating the impact of response specificity and interactivity on user experience over time
- Analyzing users' assessment when provided with a broader context or previous interactions

---

Paper: https://doi.org/10.1145/3626772.3657768

Resources: https://github.com/iai-group/sigir2024-transparentCIS

Explainability for Transparent Conversational Information-Seeking

Resources: https://github.com/iai-group/sigir2024-transparentCIS

Where to find me: weronika.lajewska@uis.no, https://werlaj.github.io/

Results

User's Perception of Explanations

	Usefulness	Other Dimensions											Explanation		
		Rel.	Correct.	Compl.	Comprehen.	Conciseness	Serendipity	Coherence	Factuality	Fairness	Read.	Sat.	Source	Conf.	Limitation
	All conditions (EC1–EC10)														
Explanation Quality	**0.0 (S)**	**0.0 (S)**	0.508 (S)	**0.003 (S)**	**0.0 (S)**	**0.001 (S)**	0.09 (S)	**0.002 (S)**	0.713 (–)	**0.0 (S)**	**0.032 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	0.173 (S)
Presentation Mode	**0.019 (S)**	**0.0 (S)**	0.234 (S)	0.347 (S)	0.658 (–)	**0.001 (S)**	0.149 (S)	0.09 (S)	0.842 (–)	**0.001 (S)**	0.651 (–)	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**
	Only conditions with explanations (EC1–EC8)														
Explanation Quality	**0.0 (S)**	**0.006 (S)**	0.256 (S)	**0.002 (S)**	**0.0 (S)**	0.122 (S)	0.319 (S)	**0.003 (S)**	0.504 (S)	**0.0 (S)**	**0.014 (S)**	**0.007 (S)**	0.097 (S)	**0.0 (S)**	0.088 (S)
Presentation Mode	0.872 (–)	0.686 (–)	0.096 (S)	0.895 (–)	0.38 (S)	0.399 (S)	0.86 (–)	0.377 (S)	0.739 (–)	0.78 (–)	0.771 (–)	0.071 (S)	**0.0 (S)**	0.653 (–)	**0.0 (S)**

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

Introducing noise in explanations has a statistically significant effect on almost all user-reported response dimensions → **Noisy explanations have a strong impact on user experience in general**

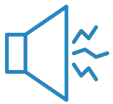
- The impact of noise on explanations is only related to the confidence

Results

User's Perception of Explanations

	Usefulness	Other Dimensions										Explanation		
		Rel.	Correct.	Compl.	Comprehen.	Conciseness	Serendipity	Coherence	Factuality	Fairness	Read.	Sat.	Source	Conf.	Limitation
	All conditions (EC1–EC10)														
Explanation Quality	**0.0 (S)**	**0.0 (S)**	0.508 (S)	**0.003 (S)**	**0.0 (S)**	**0.001 (S)**	0.09 (S)	**0.002 (S)**	0.713 (–)	**0.0 (S)**	**0.032 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	0.173 (S)
Presentation Mode	**0.019 (S)**	**0.0 (S)**	0.234 (S)	0.347 (S)	0.658 (–)	**0.001 (S)**	0.149 (S)	0.09 (S)	0.842 (–)	**0.001 (S)**	0.651 (–)	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**
	Only conditions with explanations (EC1–EC8)														
Explanation Quality	**0.0 (S)**	**0.006 (S)**	0.256 (S)	**0.002 (S)**	**0.0 (S)**	0.122 (S)	0.319 (S)	**0.003 (S)**	0.504 (S)	**0.0 (S)**	**0.014 (S)**	**0.007 (S)**	0.097 (S)	**0.0 (S)**	0.088 (S)
Presentation Mode	0.872 (–)	0.686 (–)	0.096 (S)	0.895 (–)	0.38 (S)	0.399 (S)	0.86 (–)	0.377 (S)	0.739 (–)	0.78 (–)	0.771 (–)	0.071 (S)	**0.0 (S)**	0.653 (–)	**0.0 (S)**

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



Introducing noise in explanations has a statistically significant effect on almost all user-reported response dimensions → **Noisy explanations have a strong impact on user experience in general**

- The impact of noise on explanations is only related to the confidence

- **Response dimensions are insensitive to the way explanations are presented**

- The impact of the presentation mode is only related to the limitations

Results
Qualitative Analysis

Comments stating that ...

- ... explanations enhance the understanding of the constraints of the system and the response → ***11%***
- ... responses restricted to three sentences and a single source are insufficient in certain situations → ***3%***
- ... interpreting explanations related to limitations and confidence scores is challenging → ***2%***
- ... there is a mismatch between the source and the response → ***<1%***

General conclusions:

- Overall, workers consistently emphasized that explanations enhance their understanding and encourage information verification and critical thinking
- Workers are unlikely to identify flaws in the provided explanations (positive comments also for noisy explanations)

Results
Pilot Study

- Ran on MTurk with 15 crowd workers and 3 HITs corresponding to EC3, EC4, and EC7 (US$3 per HIT)
- Feedback: crowd workers expressed concerns about the length of the task and the payment which was accordingly increased in the large-scale data collection
- Results of power analysis:
    - 16 workers are required to observe a statistically significant effect of explanation quality on the perceived usefulness of system responses
    - 56 workers are required for a statistically significant effect of the explanation presentation mode

      → **We recruited 16 unique workers per HIT in our main study.**

Results

Experiments sensitivity

- One- and two-way ANOVA to test statistical significance for the user-reported dimensions
- Response quality, quality of explanations, and their presentation mode are treated as three separate independent variables to simplify the interpretation of the results
- Each user-reported response dimension score and user rating for explanation is treated as a dependent variable

Results

Effect of Query, Topic Familiarity, Familiarity with Conv. Agents

	Usefulness	Other Dimensions										Explanation					
		Rel.	Correct.	Compl.	Comprehen.	Conciseness	Serendipity	Coherence	Factuality	Fairness	Read.	Sat.	Source	Conf.	Limitation		
Query	0.341 (S)	0.911 (–)	0.939 (–)	0.84 (–)	0.733 (–)	0.449 (S)	0.66 (–)	0.543 (–)	0.724 (–)	0.098 (S)	0.125 (S)	0.254 (S)	1.0 (–)	1.0 (–)	1.0 (–)		
Topic Familiarity	**0.017 (S)**	**0.0 (S)**	0.285 (S)	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (M)**	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	**0.002 (S)**	**0.0 (S)**	**0.0 (M)**	**0.0 (S)**	**0.0 (S)**		
Interest In Topic	**0.0 (S)**	**0.007 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	0.053 (S)	**0.0 (M)**	0.115 (S)	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (M)**	**0.0 (S)**	**0.0 (S)**		
Similar Search Prob.	**0.0 (S)**	**0.0 (S)**	**0.001 (S)**	**0.0 (M)**	**0.0 (S)**	**0.0 (S)**	**0.0 (M)**	**0.002 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (M)**	**0.0 (S)**	**0.0 (S)**		
Conv. Agent Familiarity	0.079 (S)	**0.0 (S)**	0.077 (S)	**0.001 (S)**	**0.0 (S)**	0.093 (S)	**0.0 (S)**	**0.003 (S)**	**0.0 (S)**	0.079 (S)	**0.005 (S)**	**0.004 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**		
Search with Agent Freq.	**0.0 (S)**	**0.002 (S)**	0.351 (S)	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (M)**	**0.0 (S)**	0.533 (–)	0.426 (S)	**0.0 (S)**	**0.0 (S)**	**0.0 (M)**	**0.0 (S)**	**0.0 (M)**		

- No statistically significant effect of the query on the user-reported response dimensions
- A significant effect of familiarity with the topic on response assessment indicates the need for the user's background knowledge to complement the system's errors

Results

One-way ANOVA

	Usefulness	Other Dimensions											Explanation		
		Rel.	Correct.	Compl.	Comprehen.	Conciseness	Serendipity	Coherence	Factuality	Fairness	Read.	Sat.	Source	Conf.	Limitation
	All conditions (EC1–EC10)														
Response Quality	0.156 (S)	0.176 (S)	**0.003 (S)**	0.745 (–)	0.846 (–)	0.374 (S)	0.093 (S)	0.217 (S)	0.265 (S)	0.924 (–)	0.881 (–)	0.638 (S)	0.697 (–)	0.456 (S)	0.445 (S)
Explanation Quality	**0.0 (S)**	**0.0 (S)**	0.508 (S)	**0.003 (S)**	**0.0 (S)**	**0.001 (S)**	0.09 (S)	**0.002 (S)**	0.713 (–)	**0.0 (S)**	**0.032 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	0.173 (S)
Presentation Mode	**0.019 (S)**	**0.0 (S)**	0.234 (S)	0.347 (S)	0.658 (–)	**0.001 (S)**	0.149 (S)	0.09 (S)	0.842 (–)	**0.001 (S)**	0.651 (–)	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**
Query	0.341 (S)	0.911 (–)	0.939 (–)	0.84 (–)	0.733 (–)	0.449 (S)	0.66 (–)	0.543 (–)	0.724 (–)	0.098 (S)	0.125 (S)	0.254 (S)	1.0 (–)	1.0 (–)	1.0 (–)
Topic Familiarity	**0.017 (S)**	**0.0 (S)**	0.285 (S)	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (M)**	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	**0.002 (S)**	**0.0 (S)**	**0.0 (M)**	**0.0 (S)**	**0.0 (S)**
Interest In Topic	**0.0 (S)**	**0.007 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	0.053 (S)	**0.0 (M)**	0.115 (S)	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (M)**	**0.0 (S)**	**0.0 (S)**
Similar Search Prob.	**0.0 (S)**	**0.0 (S)**	**0.001 (S)**	**0.0 (M)**	**0.0 (S)**	**0.0 (S)**	**0.0 (M)**	**0.002 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (M)**	**0.0 (S)**	**0.0 (S)**
Conv. Agent Familiarity	0.079 (S)	**0.0 (S)**	0.077 (S)	**0.001 (S)**	**0.0 (S)**	0.093 (S)	**0.0 (S)**	**0.003 (S)**	**0.0 (S)**	0.079 (S)	**0.005 (S)**	**0.004 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**
Search with Agent Freq.	**0.0 (S)**	**0.002 (S)**	0.351 (S)	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**	**0.0 (M)**	**0.0 (S)**	0.533 (–)	0.426 (S)	**0.0 (S)**	**0.0 (S)**	**0.0 (M)**	**0.0 (S)**	**0.0 (M)**
	Only conditions with explanations (EC1–EC8)														
Explanation Quality	**0.0 (S)**	**0.006 (S)**	0.256 (S)	**0.002 (S)**	**0.0 (S)**	0.122 (S)	0.319 (S)	**0.003 (S)**	0.504 (S)	**0.0 (S)**	**0.014 (S)**	**0.007 (S)**	0.097 (S)	**0.0 (S)**	0.088 (S)
Presentation Mode	0.872 (–)	0.686 (–)	0.096 (S)	0.895 (–)	0.38 (S)	0.399 (S)	0.86 (–)	0.377 (S)	0.739 (–)	0.78 (–)	0.771 (–)	0.071 (S)	**0.0 (S)**	0.653 (–)	**0.0 (S)**

Results

Two-way ANOVA

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

Experimental Conditions

The selected conditions vary along three main dimensions:

1) **Response quality:** ground-truth response or imperfect response with biases or factual errors
2) **Quality of the explanations:** accurate or noisy explanation
3) **Presentation mode:** additional UI component or in natural language explanation

We have defined ten experimental conditions using different variants of the response and explanations:

	EC1	EC2	EC3	EC4	EC5	EC6	EC7	EC8	EC9	EC10
Response	+, T	+, T	~, T	~, T	+, T	+, T	~, T	~, T	+, T	~, T
Source	+, T	+, T	+, T	+, T	~, T	~, T	~, T	~, T	–	–
Confidence	+, V	+, T	+, V	+, T	~, V	~, T	~, V	~, T	–	–
Limitations	+, V	+, T	+, V	+, T	~, V	~, T	~, V	~, T	–	–

+ *component without noise*
~ *component with noise*
– *component not provided*
V *component presented visually*
T *component provided in text*

We use ten queries selected from the TREC CAsT 2020 and 2022 datasets and two manually created responses for each query