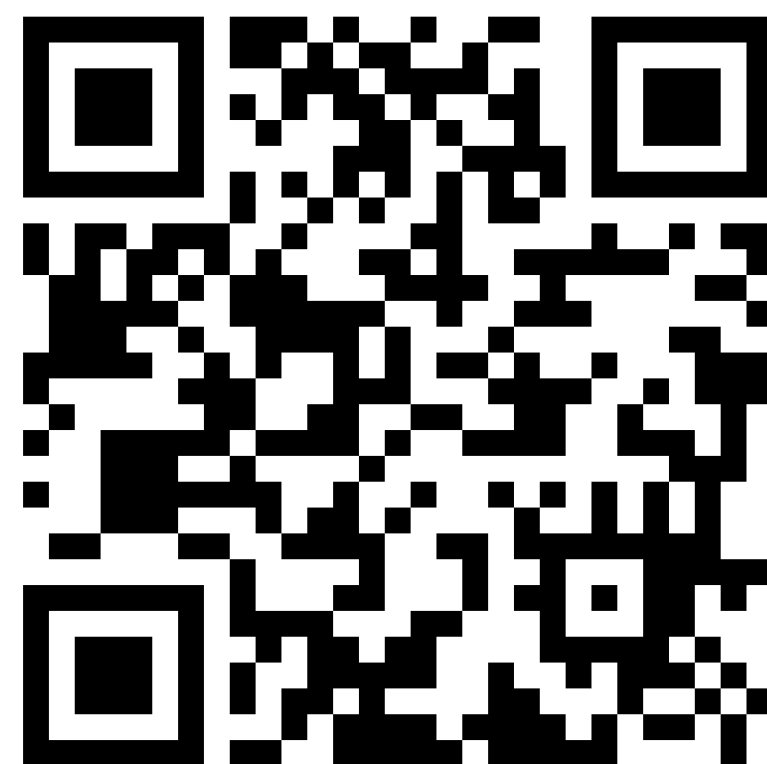


Explainability for Transparent Conversational Information-Seeking



Weronika Łajewska¹, Damiano Spina², Johanne Trippas², Krisztian Balog¹

¹Information Access & Artificial Intelligence, University of Stavanger, Norway; ²RMIT University, Melbourne, Australia

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!

HIT	
I)	HIT instructions
II)	Familiarity with Conversational Assistant
III)	System Description
IV)	<div>CIS interaction 1<div>a) Queryb) Topic familiarityc) Response (with explanations)d) Attentiveness Checke) Response assessment</div>CIS interaction 2<div>....</div>CIS interaction 10<div>....</div></div>
V)	Post-task questionnaire
VI)	Demographic information

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

RESULTS

	Usefulness	Other Response Dimensions										
		Rel.	Correct.	Compl.	Comprehen.	Concise.	Serend.	Coheren.	Fact.	Fairness	Read.	Sat.
	All responses											
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 responses with explanations											
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. Response dimensions include: Relevance, Correctness, Completeness, Comprehensiveness, Conciseness, Serendipity, Coherence, Factuality, Fairness, Readability, Satisfaction. Boldface indicate statistically significant effects ($p < 0.05$). Effect size: L=Large, M=Medium, S=Small.

LESSONS LEARNED

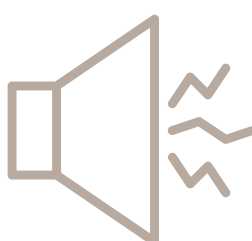
Users are not able to identify bias towards one specific point of view or factual errors without expert knowledge about the topic



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



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



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

