Explainability for Transparent Conversational Information-Seeking

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

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

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

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

ABSTRACT

The increasing reliance on digital information necessitates advancements in conversational search systems, particularly in terms of information transparency. While prior research in conversational information-seeking has concentrated on improving retrieval techniques, the challenge remains in generating responses useful from a user perspective. This study explores different methods of explaining the responses, hypothesizing that transparency about the source of the information, system confidence, and limitations can enhance users' ability to objectively assess the response. By exploring transparency across explanation type, quality, and presentation mode, this research aims to bridge the gap between system-generated responses and responses verifiable by the user. We design a user study to answer questions concerning the impact of (1) the quality of explanations enhancing the response on its usefulness and (2) ways of presenting explanations to users. The analysis of the collected data reveals lower user ratings for noisy explanations, although these scores seem insensitive to the quality of the response. Inconclusive results on the explanations presentation format suggest that it may not be a critical factor in this setting.

CCS CONCEPTS

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

KEYWORDS

Conversational information-seeking; Explainable AI

ACM Reference Format:
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), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3626772.3657768

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657768

1 INTRODUCTION

The field of conversational information-seeking (CIS) focuses on systems designed for dialogue-based information retrieval, where the aim is to enable interactions that closely resemble human conversation [56]. In recent years, research in this space has primarily concentrated on improving various components of the response generation process, such as passage retrieval, reranking, query rewriting, and on making answers self-contained [38, 47, 56]. However, it remains a challenge to create a trustworthy conversational response from the retrieved information [44]. In transitioning from traditional search engine result pages to conversational information-seeking systems that limit responses to a few sentences, there is a significant concealment of underlying details such as the ranking of results and specifics about the sources. These details are essential for users to assess the scope, novelty, reliability, and topical relevance of the provided information [55]. Recently, retrieval-augmented generation (RAG) has been proposed [31], which is claimed to produce more factually correct and diverse content. RAG, however, does not solve issues around transparency, as it is not able to indicate low-confidence responses or identify potential flaws related to limitations of the retrieved results or of the response generation process itself. Since the user is provided only with a short textual response as the final outcome of the generation process, it becomes the responsibility of the conversational system to identify and communicate any potential limitations to its users, ensuring transparency and empowering users to evaluate response quality. While the importance of explainability is broadly recognized for AI [36] and has been extensively studied, for example, for decision support and recommender systems [37, 57], it has not received due attention for CIS systems.

In this study, we aim to fill this gap by investigating approaches to explaining conversational responses, as a means to increase the transparency of the system. Our focus is on *informational transparency*, disclosing information about the limitations or potential pitfalls in the response generation needed to enable appropriate understanding and assessment, in contrast to *functional* understanding of what the system can do, by exposing its capabilities and limitations or *mechanistic* understanding focused on how the system works [33]. In particular, the focus of this study lies on the sources used for generating the response, the system's confidence in the provided information, and potential limitations or pitfalls of the response. In contrast to prior research on reporting system

confidence [9, 42] or identifying particular limitations [21, 61], our emphasis is on effectively communicating this information to the user in a conversational setting; we thus assume the existence of components that estimate system confidence and perform the detection of limitations.

Specifically, based on the previous research in related domains, we choose to increase the transparency of a CIS system by explaining (1) the origin of presented information, i.e., source [4, 34, 53], (2) the system's confidence [9, 39], and (3) potential limitations of the generated response [45]; see Figure 1 for an illustration of an enhanced conversational response. Being transparent about these aspects of the response can enable users to make informed judgments about the presented information and increase their perceived usefulness of the response. We investigate the user's perception of the system response quality together with the type and quality of explanations. We ask the following two main research questions.

RQ1: How does the quality of responses and explanations affect user-perceived response usefulness? Individuals without specific training can only distinguish between human-generated and auto-generated texts at a level close to random chance [13]. Indeed, users easily overlook factually incorrect, unsupported, biased, or incomplete information. Therefore, we investigate the impact of the quality of the response and explanations provided by the system on users' assessment of the response. Specifically, imperfect responses in our study include factual errors or lack of viewpoint diversification, while noisy explanations introduce problems related to sources (information subjectivity or lack of support for the response), confidence (incorrect scores), or limitations (irrelevant information about pitfalls).

RQ2: What are effective ways to provide explanations to users? There are multiple approaches to providing users with explanations. One option is incorporate them as part of the natural language system utterance, ensuring that users are explicitly informed about the confidence and potential pitfalls of the response [42]. As an alternative, we explore utilizing various user interface elements to effectively convey the response's limitations [35, 48] or providing a granular scale of the system's confidence in generated response [48]. Building on findings from studies in recommender systems and automated decision making [37, 57], we seek to adapt and explore these concepts within the context of CIS systems.

To answer the above questions, we conduct a crowdsourcing-based user study with 160 participants asking about their perception of responses and explanations that vary in quality and presentation mode. The analysis of the collected data reveals that users can identify situations when their level of confidence in the response does not match the confidence reported by the system, but they cannot detect noise in the explanations of system limitations. In general, noisy explanations result in lower user-reported scores for response usefulness. In terms of presentation mode, we observe no significant advantage of any of the modes, which suggests that the explanation format is not critical for users.

Overall, our study seeks to establish a more trustworthy interaction in CIS dialogues by bridging the gap between system-generated responses and their usefulness to the users, by providing explanations. The main contributions of this work include (1) the first user study that provides explanations of source, confidence, and limitations in the CIS domain, (2) a manually curated dataset of responses and explanations, with noise incorporated in a controlled manner, and (3) generalizable results describing the impact of noise and the presentation mode of the explanations on response usefulness and user ratings of explanations of source, confidence, and limitations.

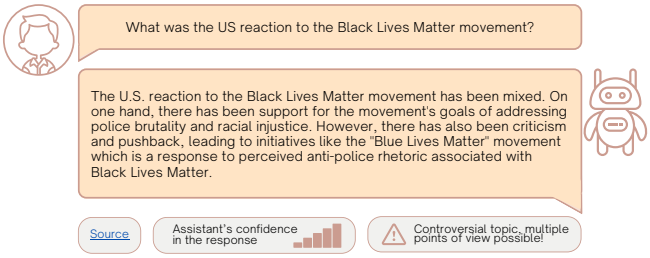

Figure 1: Information-seeking dialogue with a CIS system with explanations (sources, confidence, and limitations).

The paper is accompanied by an online repository, containing the manually generated CIS responses and explanations, user study results, and scripts for data analysis at https://bit.ly/TransparentCIS.

2 RELATED WORK

While traditional search engines provide ranked lists of documents, conversational response generation aims to aggregate top-ranked passages into coherent answers [44]. The task of generating summaries from retrieved results was first piloted in the 2022 edition of the TREC Conversational Assistance track (CAsT) [38]. Generative language models have been widely adopted for response generation [59]; however, the final step of aggregating supporting facts through abstractive summarization has its challenges, including factual errors [52] and hallucinations [8, 22, 51]. Despite advancements, CIS systems still face limitations such as unanswerability [12, 41, 43, 50], biases and lack of viewpoint diversification [2, 17, 19, 45], or queries with impossible conditions [21, 30]. Even though research has been done in related fields, such as text classification [25, 61], question answering [32, 41], or reading comprehension [21, 60] towards detecting these issues, communicating detected problems to users is still a largely unexplored area in CIS. To ensure the transparency of the system, the response should disclose system capabilities and potential limitations, thereby managing user expectations [3, 40].

Unlike previous studies that concentrated on detecting limitations, our work emphasizes the effective communication of potential flaws in the system's output to the user. Such limitations can be revealed using natural language utterances [42], using analogy [20], incorporating user interface elements [26, 35], or by providing a granular scale of the system's confidence [48]. Studies in the context of recommender systems have demonstrated that reporting the system's confidence in predictions can offer valuable information to users, aiding them in making informed decisions [48]. Moreover, research on human-AI interactions has highlighted the significance of system performance feedback in shaping human trust and reliance on AI systems [35, 42]. Both the level of confidence displayed by machine learning models and their observed accuracy influence people's belief in the model's predictions, and users' willingness to follow the model's suggestions, especially after observing the model's performance in practice [42].

Figure 2: Examples of responses with explanations for the query: *What was the US reaction to the Black Lives Ma er movement?*

Effective user-system interactions, aligning user expectations, and building trust in the systems that we attempt to achieve by communicating explanations about the response to the user are the main axes of explainable AI (XAI). According to human-AI interactions design guidelines, the system should communicate its capabilities, reliability, and the rationale behind its decisions [1]. XAI research in the context of decision-making emphasizes the role of explanations in improving user comprehension [10] and increasing human trust in the system [58]. Explanations can vary in terms of presentation format (textual vs. visual) [58], the level of interactivity [10], complexity [53], or reasoning styles [9]. Explanations can also be used to reveal the system's confidence [9], system accuracy indicators [27], data sources [53], answer attribution [4, 34] and the correctness of system suggestions [9]. While explanations can enhance users' adherence to the system's advice, they may also lead to incorrect mental models of the systems—overconfidence or overreliance—especially among users with low domain expertise [9]. Therefore, understanding the perceived usefulness of explanations is the first step in designing reliable and transparent CIS systems.

3 METHODOLOGY

We aim to investigate the user's perception of the (1) system response quality, (2) type and quality of explanations, and (3) presentation of explanations. We assume a *retrieval-augmented generation system* that, given a query, performs the following steps: (1) it retrieves passages and identifies the information nuggets in the top retrieved results containing key pieces of information answering a user query; (2) it synthesizes the identified snippets (i.e., information nuggets) into a concise and natural language response; (3) it returns the system's confidence in the provided response; and (4) based on the provided query, retrieved information nuggets, and returned confidence, it identifies the potential pitfalls and limitations that could have contributed to flaws in the response. We consider three types of explanations the system may provide: (1) the underlying *source*, to help users verify the response's factual correctness and broader context; (2) the system's *confidence* in the provided response, to give users insights about how certain the outcome of response generation is; and (3) potential *limitations or pitfalls* to warn the user about flaws in the response or the source.

The study's main goal is to investigate whether explanations provided by the system can make the user's response assessments easier or increase the information's usefulness. We provide crowd workers with different configurations of responses and explanations, varying in quality and presentation mode, and ask them to indicate their perception of different system response dimensions. Inspired by work in the area of explainable decision-making systems [9], we explore two different ways of presenting the explanations about the response limitations and system confidence: textual and visual presentation; see Figure 2.

3.1 Experimental Design

We have defined ten experimental conditions using different variants of the response and explanations.[1] Covering all combinations of factors (explanation components × quality × presentation mode) exhaustively would be unfeasible. Therefore, we select a subset of experimental conditions that best represent what we are trying to measure in our study. The selected conditions vary along three main dimensions: (1) response quality, (2) quality of the explanations (i.e., source, system confidence, limitations), and (3) presentation style (see Table 3). More details about experimental conditions and the different explanation variants can be found in Section 4.1.

The ten experimental conditions resulted in ten different human intelligence tasks (HITs). In each HIT, crowd workers are asked to assess responses for ten queries. This is to ensure that the obtained results are to a large extent topic-independent. To avoid repeated judgments that would reduce the reliability of the study, we allow each crowd worker to complete only one HIT [49], i.e., we employ a between-subject design [24]. In each HIT, the order of query-response pairs is intentionally randomized. This is done to prevent any adverse effects on the given query-response pairs that might occur if they were consistently presented towards the end of the task, where worker fatigue could potentially influence the results.

3.2 Crowdsourcing Task Design

Figure 3 summarizes the design of the crowdsourcing tasks. Each HIT contains ten query-response pairs and is comprised of: I) HIT instructions providing task background; II) a questionnaire about the worker's familiarity with conversational assistants (see Table 2); III) a description of the system; IV) ten CIS interactions; V) a post-task questionnaire; and VI) a demographics questionnaire. Workers are not given specific examples of query-response pairs in the instructions to avoid bias. Part III contains a pre-use explanation of the system [11]. It aims at improving the following competencies of the users: (1) understanding the capabilities of the system, and

[1]We acknowledge that the variants for each transparency dimension are not exhaustive. Various UI elements can be used to present information, and different ways to introduce noise can be explored. However, since response-related explanations have not been explored in conversational search, we limit the first study in this area to solutions previously proposed for similar systems.

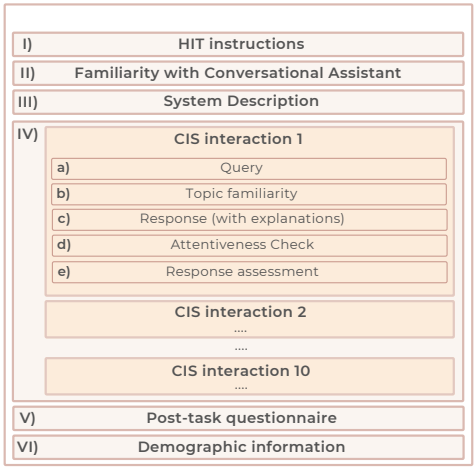

Figure 3: High-level design of the user study.

(2) understanding that the response is limited to 3 sentences only. We decompose part IV of the user study into ten subsections using independent CIS interactions to facilitate atomic microtask crowdsourcing [18]. Each CIS interaction contains (a) a query; (b) a topic familiarity questionnaire; (c) a system response possibly enhanced with explanations; (d) a corresponding attentiveness check; and (e) a CIS response assessment. CIS interactions are followed by a post-task questionnaire (Part V) investigating workers' experience of interacting with the assistant in general, not concerning specific responses. The questionnaire contains indirect questions about all three types of explanations enhancing the system response (see Table 2). The HIT finishes with a short demographics questionnaire (Part VI) asking workers' age, education level, and gender.[2]

a) ery and b) Topic Familiarity. The query is followed by a short questionnaire asking about interest, familiarity, and likelihood of posing a similar query [5] (see Table 2). In this user study, the worker's background knowledge and familiarity with the topic are dependent variables that we cannot control. Asking users to assess their familiarity with the topic enables us to condition the collected data on users' background knowledge [28].

c) Response. The system response synthesizes the information nuggets identified in the top retrieved results. The response can be enhanced with explanations that can be presented in different formats.

d) A entiveness Check. We present workers with an attentiveness check for each query-response pair, to detect poorly performing workers, cheat submissions, or bots [18]. Each attention check consists of three sentences related to the topic of the query, 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 [6]. This simple quality check enables us to filter out responses from workers who are not performing the task attentively or reading the responses carefully. Submissions that failed on more than 3 out of 10 attentiveness questions were rejected.

e) Response Assessment. In this part of the CIS interaction, workers are asked to evaluate different dimensions of the response variant presented for a given query. The question about each response dimension is answered on a four-point Likert scale. Explicitly asking users to report on its value is not helpful because they may have a different understanding of this concept [24]. Therefore, in our setup, user satisfaction is indirectly observable. To increase the ecological validity of our experiments, the questions do not use explicitly the names of the dimensions. Instead, we ask about each response dimension using an operational definition (see Table 1). This approach ensures a common understanding of the dimensions by all study participants. Both the response dimensions and the operational definitions are inspired by Cambazoglu et al. [7]'s work investigating answer utility for non-factoid question answering.

[2]Specific questions posed in each user study can be found in the online repository.

Table 1: Operational definitions used in the response assessment questionnaire for all response dimensions. They followed a statement: *The provided assistant's response ...* and were answered by crowd workers on a four-point Likert scale.

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

Table 2: Questions used for collecting data about the user experience of using conversational agents, their involvement in the topic, and their rating for explanations.

Variable	Question used in the user study
Conversational Agent Familiarity	How often do you use conversational assistants like Siri, Alexa, or Google Assistant?
Search with Agent Freq.	How often do you use conversational assistants to search for information?
Topic Familiarity	What is your level of familiarity with the topic of the question?
Interest in Topic	What is your level of interest in the question?
Similar Search Probability	What is the likelihood that you would search for this information?
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?

4 USER STUDY EXECUTION

We used the Amazon Mechanical Turk (MTurk) crowdsourcing platform to collect responses from online workers.[3] Data collection was run between 20 December 2023 and 9 January 2024, divided into two stages: a pilot (Section 4.2) and a main study (Section 4.3).

4.1 Data

A critical element of the study is selecting query-response pairs and explanations enhancing the responses that enable us to answer our research questions. We use ten queries selected from the TREC CAsT 2020 [16] and 2022 [38] datasets and two manually created responses for each query. Different variants of the responses (perfect

[3]Our institution does not require ethics approval for this kind of studies.

and imperfect) and explanations (accurate and noisy) are created manually by one of the authors of the paper. The noise in responses and explanations is introduced manually using framing, i.e., distorting the information presented to the users [27]. For each source, the specific information nuggets that contributed to the answer are highlighted, inspired by the CAsT-snippets dataset [29].

Queries and Responses. The query selection process takes into account the potential challenges of the query and the familiarity of crowd workers with the topic. We select a subset of queries from the TREC CAsT datasets that are challenging in one of two aspects: (1) limited coverage of the topic in the corpus or lack of a full answer, resulting in factual errors; or (2) topic complexity or controversy resulting in an incomplete or biased response. By selecting these challenging queries we attempt to simulate scenarios where enhancing the system response with explanation can be beneficial for users. Additionally, queries selected in the first step are sorted according to the familiarity scores reported by crowd workers in a small crowdsourcing study that was set up to select the top ten queries that are deemed most well-known to users. This approach aims to ensure that users possess sufficient background to meaningfully assess responses and associated explanations. We consider two variants of the response for each query: perfect and imperfect. The perfect response, i.e., ground truth answer, is generated manually using the top retrieved results by one of the authors of the paper. The imperfect response is a manual modification of the ground truth answer to contain factual errors, be biased towards one point of view, or cover only one aspect of a complex problem. This way, we attempt to take into account significantly different versions of the responses in terms of their accuracy and quality.

Explanations. We provide explanations related to (1) source, (2) system confidence, and (3) limitations, which are instantiated in two variants: accurate and noisy.

(1) Sources. The "Source" component is an expandable element within the response, encompassing the complete text of the paragraph used for generating the response. It includes annotations of information nuggets [29], highlighting crucial pieces of information within the passage. Additionally, workers receive a link to the entire webpage from which the passage originates [34]. This allows them to access the full text of the document, aiding in the assessment of its relevance, which is particularly beneficial for long, non-navigational queries [23]. The URLs are anchored to the specific section of the webpage where the passage is located. Additionally, based on the URL, workers can assess the credibility or authority of the source. The noisy source pertains to the query's topic but lacks information that supports the provided response [34]. It corresponds to the initial passage from the Wikipedia page related to the general query topic, allowing for an assessment of users' diligence in verifying the provided explanations.

(2) System Confidence. Within conversational response generation, confidence can be assessed along several different dimensions:

- The confidence that the identified snippets contain the full, complete answer to the question, not only part of it.
- Given that the response is limited to only 3 sentences, the confidence that the top-k snippets used in the response provide a sufficient coverage of the retrieved information.
- The confidence that the response generated with LLM using the selected snippets is accurate; this accuracy is tied to the model's fluency in the topic, assessing how adept the model is in crafting content on a given subject, which can be influenced by the volume of data during LLM training or topic popularity.

System confidence is either communicated in textual form (*"the system confidence in the provided response is ...%"* appended at the end of the response) [9] or through an additional UI element. Given that users best understand confidence displays inspired by well-known displays in other areas [48], we decided to use a bar chart presentation that is often associated with cell phone connectivity. Adding noise to this component results in system confidence being reverted, i.e., although the provided response is correct, a low system confidence is reported. We consider the confidence of 1–2 out of 5 for imperfect responses and 4–5 for perfect responses. We skip confidence of 3 as it is ambiguous and we skip confidence of 0 as it represents the situation when the system should not show any response, but state that an answer could not been found.[4]

(3) Response Limitations. We have identified several key areas of potential challenges and problems that could impact the usefulness of provided responses. These issues, while not exhaustive, serve as a starting point for consideration in our user study. Among the challenges related to the topic, we recognize the potential issues related to controversy, leading to a lack of viewpoint diversification, and complexity, resulting in response incompleteness. Source-related challenges include the subjectivity of the source text used, possibly outdated source information, sources influenced by commercial interest, promoting specific products, or brands, and reliance on unverified or not reputable sources. Query-related issues encompass biases or ambiguities in queries, time-sensitive queries requiring current information, queries lacking sufficient context, privacy-sensitive queries involving private or confidential information, and speculative queries seeking insights into future events. Additionally, search and system issues may arise, such as rare topics insufficiently covered in the corpus, lack of credible sources supporting the response, or retrieved passages containing contradictory information.

The query-response pairs selected for this study contain factual errors, are incomplete, or rely on subjective sources. Additionally, challenges related to the topic, source, or query, not identified in the subset of query-passage pairs used in this study, are also considered to explore whether users can more easily identify these issues based on the presence and correctness of explanations provided by the system. Issues purely related to search or system failures, where the system is aware of its inability to find sources that answer the question, fall outside the scope of this study. In such cases, the system should inform the user about no answer found without trying to produce a response. Response limitations are communicated either in a textual form by appending running text at the end of the system response [14, 42] or using an additional UI element resembling a warning message (inspired by fact-checking warning labels [26]).

[4]Users' reactions to such extreme confidence scores is not a subject of this study, but could be explored in future work once it has been established that users find confidence explanations useful.

Table 3: Experimental conditions considered in the user study; components may be included without noise (+); with some inaccuracies (~); or not provided in the system's output (–). (T) indicates textual and (V) visual presentation mode.

	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	–	–

Adding noise to limitation explanation results in communicating irrelevant limitations, i.e., if the topic is controversial, the system informs the user about query ambiguity or possibly outdated source information. We aim for the noisy limitations to be easily distinguishable after reading the query and the response carefully. The goal of this study is to investigate the limitation explanations rather than the detection of specific limitations, therefore the noise in the limitations is aimed to be easy to spot.

Experimental Conditions (EC). The subset of experimental conditions selected for this user study is summarized in Table 3. The conditions vary along three main dimensions: (1) response quality, (2) explanation quality, and (3) presentation of explanation. EC1 and EC2 represent a perfect system response with accurate explanations. More specifically, the explanations cover the source supporting the response, as well as the system's confidence score; the limitation explanation is not included because the response has no inaccuracies in this case. EC3 and EC4 correspond to imperfect responses that may contain some factual errors or be biased towards one specific point of view, but are accompanied by accurate explanations related to source, confidence, and limitations. EC5 and EC6 represent the perfect response accompanied by noisy explanations. EC7 and EC8 correspond to imperfect responses that contain flaws and noisy explanations. The last group of conditions, EC9 and EC10, represents the response (either perfect or imperfect) without explanations.

4.2 Pilot Study

We ran a pilot study (MTurk; N=15; 3 HITs; US$3 per HIT, proportional to US minimum wage), where HITs corresponded to three experimental conditions selected from the 10 described in Table 3: EC3, EC4, and EC7. The selected conditions encompass border cases, featuring variations in both the presentation mode of explanations (EC3 vs. EC4) and the quality explanations (EC3 vs. EC7), and deliberately involve imperfect responses to simulate the most natural scenarios. In their overall feedback, crowd workers primarily expressed concerns about the length of the task and the payment which was accordingly increased in the large-scale data collection.

We performed a power analysis by employing one-way ANOVA with the experimental condition as an independent variable and user-reported response usefulness as a dependent variable [46]. The results indicate that 16 workers are required to observe a statistically significant effect of explanation quality on the perceived usefulness of system responses, whereas 56 workers are required for a statistically significant effect of the explanation presentation mode. Considering four additional pairs of experimental conditions with varying presentation modes, we expect that gathering data from 14 unique workers per HIT (56 from power analysis divided by 4 pairs of conditions) is adequate to observe a statistically significant effect of presentation mode across all ten experimental conditions. Based on this analysis, we decided to recruit 16 unique workers per HIT in our main study.

4.3 Main Study

Crowd workers with a greater than 97% approval rate, over 5,000 approved HITs, and located in the US were qualified to participate in the study. Workers were paid US$4 for successful HIT completion. Workers who failed 4 out of 10 attentiveness checks or more were rejected. Altogether we collected 273 submissions, out of which 113 were discarded due to failed attentiveness checks. Accepted tasks were submitted by 160 unique workers (16 per HIT), with the following user-reported demographics: 95 male, 60 female, 5 in "other" category (none in "prefer not to say"); age breakdown: 18–30 (39), 31–45 (76), 46–60 (41), 60+ (4); highest degree: Ph.D. or higher (3), master's (34), bachelor's (111), high school (12).

5 RESULTS

To answer our research questions, we first analyze the sensitivity of our experiment. Tables 4 and 5 show the results of the one- and two-way ANOVA tests for statistical significance on user-reported dimensions, using a significance level of $\alpha = 0.05$. Whenever applicable, the effect size of a given factor is classified based on the formula for the unbiased estimator and scales used by Culpepper et al. [15]. Given the large number of factors defining each experimental condition, we treat response quality, quality of explanations, and their presentation mode 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. The analysis performed to answer RQ1 (Section 5.2) and RQ2 (Section 5.3) is based only on the results with the statistically significant effects discussed in Section 5.1.

5.1 User's Perception of Response and Explanations

Response Quality. Table 4 shows that *response quality* has a statistically significant effect only on user-reported correctness of the response. Completeness, factuality, and fairness are not influenced by the quality of the response, even though some responses contained manually injected errors related to these dimensions (e.g., bias towards one specific point of view, factual errors, or covering only one aspect of the topic). This insensitivity of user-reported response dimensions to the quality of provided information may suggest that users are not able to identify some of the problems with the response without expert knowledge about the topic.

Explanations. Our experiments include two experimental conditions where explanations are not provided (i.e., EC9–EC10). In order to understand the impact of quality and presentation mode of explanations, we conducted an additional analysis on the data from HITs representing only EC1–EC8 (reported in the bottom part of Table 4) and we focused our analysis on these results. In terms of *explanation quality*, we observe that introducing noise in explanations has a statistically significant effect on almost all user-reported response dimensions, suggesting that noisy explanations have a

Table 4: 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.

	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)**

Table 5: Results of two-way ANOVA. Boldface indicates statistically significant effects ($p < 0.05$). Effect size: S=Small.

	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)**

strong impact on user experience in general. However, the quality of explanations does not impact user assessment of correctness and factuality, dimensions related to factual errors in the response. It means that users seem to assess the factual correctness of the response independently of the quality of the explanations provided by the system. In terms of *presentation mode*, we observe a statistically significant effect only for the presence/absence of explanations on the usefulness of the response—a statistically significant effect is observed in the top part but not in the bottom part of the table. Similarly, the user-reported conciseness, fairness, and relevance of the response are impacted only by the presence/absence of explanations. This implies insensitivity of the response dimensions to the way explanations are presented.

User Ratings for Source, Confidence, and Limitations. The impact of noise in the source is solely tied to the presence or absence of the source—no statistically significant effect is observed for EC1–EC8. However, the presentation mode of the source affects user ratings for the explanations independently of its presence or absence (statistical significance persists when excluding EC9 and EC10), even though the source is presented in the same format in both presentation modes. This may be due to the wording of the question about source explanation in the questionnaire—it does not explicitly mention sources, and therefore is open for other interpretation, especially when sources are not provided. In the case of noise in confidence explanation, it significantly affects user ratings. However, concerning presentation mode, we can only discern the effect of its presence or absence, not the specific mode of presentation. Regarding limitations, there is no statistically significant effect of noise in the corresponding explanation, but there is of presentation mode. User ratings for explanations related to limitations are influenced by the presentation mode, not the mere presence of this explanation. This implies that, in general, the impact of noise on explanations is only related to the confidence and the impact of the presentation mode only to the limitations. The effect of quality and presentation mode on other explanations—based on the user ratings—was not significant in this user study.

Query. We do not observe any statistically significant effects of the query on the user-reported response dimensions. This suggests that the results are topic-independent and generalizable. The proposed user study design mitigates the impact of the query on the results. Additionally, the interaction between response quality, quality of explanations, or presentation mode and the query does not have a statistically significant effect on user-reported scores for response satisfaction, usefulness, and explanations (see Table 5).

Topic Familiarity. Workers report that they are rather familiar with the query topics, which indicates that the process of query selection was successful. Following our hypothesis, users' background knowledge about the topic affects how they assess the response. It is visible on the effects reported for almost all response dimensions (see Table 4). Similar effects are observed for the user's interest in the topic and the likelihood of the user searching for a similar query. Additionally, we observe a statistically significant effect of all these three indicators of user involvement in the topic on the user ratings for explanations. It implies that these factors that we cannot control and are completely user-dependent directly impact the assessment of the responses we examined in this user study.

In terms of the results of two-way ANOVA (see Table 5), we observe a statistically significant effect of the interaction between the user's familiarity with the topic and the response quality on the user ratings for explanations related to limitation and the system's confidence. It confirms the intuitive relationship between the user's background knowledge and the quality of the response on their ability to correctly assess the explanations provided by the system and deem it useful or not. We do not observe a statistically significant effect of interaction between response quality and familiarity with the topic on the usefulness of the response or user satisfaction in general. The interaction between the noise in the explanations and the familiarity with the topic has a statistically

significant effect on the user ratings for all three explanations. It can follow from the fact that a user unfamiliar with the topic needs high-quality explanations from the system to be able to verify and use the provided response. The user ratings for explanations are influenced by the interaction between the way explanations are presented and the user's familiarity with the topic. This suggests that depending on the user's background knowledge, the preferred way of receiving explanations from the system may differ.

Familiarity with Conversational Agents. We observe a statistically significant effect of the user-reported frequency of interacting with conversational assistants (in general and for search specifically) on some of the user-reported response dimensions (see Table 4). Interestingly, we observe a medium-size effect of the frequency of using the conversational search on the user ratings for explanations related to source and limitations. Additionally, we observe that higher values for familiarity with conversational agents are associated with explanations without noise and visual presentation mode. It indicates that the user's familiarity with the system impacts their assessment of its additional components.

5.2 Effect of the Explanation Quality (RQ1)

Effect on the User-reported Response Dimensions. The top-left plot in Figure 4 shows that user-reported values for the usefulness of the responses are concentrated around higher values (3 and 4). However, noise in the explanations results in slightly lower usefulness scores. It indicates that the high-quality source, system confidence score, and information about the response limitations make the response more useful from the user's perspective. Minor differences in usefulness scores between perfect and imperfect responses (second and third set of bars in the plot) suggest that when explanations are not provided ("None" variant), users are less likely to object to the usefulness of imperfect responses. In general, the explanations are meant to increase the reliability and transparency of the system. However, they require additional time and effort from the user and the cost of "processing" explanations may be higher than the actual gain. This situation is visible in the second and third set of bars in the top-left plot in Figure 4 where the highest usefulness is reported for the responses that do not contain any explanations ("None" variant), independent of their quality. It suggests that the explanations either pollute the response or make the user more critical about it, resulting in reduced usefulness.

Effect on user ratings of explanations. Looking at the means of user ratings for source and confidence explanations (top-right plot in Figure 4), ratings are again skewed towards higher values, and scores for accurate explanations are slightly higher than for noisy explanations, especially for confidence. This suggests that users perceive noisy explanations as less useful in understanding system confidence and attributed sources—we do not observe statistically significant differences for the explanations related to limitations.

5.3 Effect of the Presentation Mode (RQ2)

Effect on the user-reported response dimensions. On average, we do not observe differences in the usefulness scores between textual and visual modes, but usefulness scores are significantly higher when no explanations are provided (bottom-left plot in Figure 4).

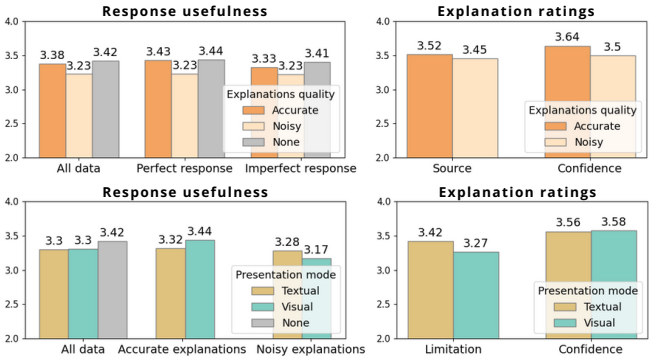

Figure 4: Mean scores for response usefulness and explanation ratings for different quality of the explanations (top) and presentation mode (bottom). All differences between the ratings within a given plot are statistically significant.

This is aligned with the one-way ANOVA results and suggests that the main issue is not the question of presentation mode but rather whether the explanations are necessary, hinting at the underlying trade-off between effort and gain. Nevertheless, we observe some differences in the user ratings for explanations when looking at responses accompanied by explanations with different quality. Namely, visual explanations result in higher usefulness scores for responses with accurate explanations, while in case of noisy explanations workers find the textual format more useful.

Effect on user ratings of explanations. Looking at the means of user ratings for explanations with respect to different presentation modes (bottom-right plot in Figure 4), the preferred presentation mode depends on the explained aspect of the response. (Note that user ratings for explanations related to the source are not informative in this case, as the source is always presented in the same way.) Namely, we observe slightly higher ratings for the textual presentation of limitations. In the case of confidence, the difference between presentation modes is very small with a slight preference towards visual presentation, which aligns with the results of one-way ANOVA. This suggests that further research is needed to better understand how to optimally integrate different aspects in the layout of transparent CIS responses.

5.4 Qualitative Analysis

We manually investigate the feedback given by crowd workers regarding their ratings for the source, confidence, and limitations explanations, seeking insights and suggestions to enhance their content and presentation.[5] Many workers (18/160) pointed out that explanations related to limitations and confidence significantly enhanced their understanding of the constraints of both the system and the responses. The mention of encouragement towards information verification and critical thinking was consistent across various qualities (comments from EC1–EC8 HITs), and positive comments were also shared for noisy explanations (EC5–EC8). It suggests that workers may face challenges in identifying inaccuracies in the explanations. For instance, even though the provided sources did not align with the information in the response, none of the users

[5] Comments provided by workers can be found in the online repository.

mentioned these mismatches in their comments. Nevertheless, several crowd workers (4/160) emphasized the potential insufficiency of responses restricted to three sentences and a single source in certain situations. A few (3/160) crowd workers expressed uncertainty in interpreting explanations related to limitations and confidence scores, underscoring the need for additional explanations or tutorials describing the system interface before usage. For instance, some workers attempted to interpret the meaning of the confidence score on their own describing it as a "transparency measure to indicate the system's level of certainty regarding the accuracy or relevance of the information shared" or a "model's estimate of the accuracy and reliability of its responses." In terms of the presentation mode, one worker suggested that representing confidence score using percentages would be more precise and helpful than a "wifi connection symbol." This suggests that users might prefer a different display element, e.g., a fuel gauge [48], and perhaps also a finer confidence scale (which would require a more precise estimation of confidence). In HITs with no explanations (EC9 and EC10), workers highlighted their lack of awareness regarding response limitations and confidence. Some workers attempted to gauge system confidence by searching for implicit confidence signals like "I think" or "I believe" in the responses [39]. Overall, workers consistently emphasized that explanations enhance their understanding and encourage information verification and critical thinking. However, the comments reflect that workers are unlikely to identify flaws in the provided explanations.

6 DISCUSSION

Our results show that high-quality explanations related to the source, system confidence, and response limitations increase the user-perceived usefulness of the response and user ratings for explanations. Additionally, noise in the explanations of the response provided by the system has a significant impact on user experience in general (almost all response dimensions are affected). These results align with previous research in AI-assisted decision-making claiming that confidence scores can help calibrate people's trust in the system model, but they are not sufficient to increase the success rate of interactions [58]. In our study, we observe a significant effect of familiarity with the topic on response assessment, indicating the need for the user's background knowledge to complement the system's errors [58]. In terms of user's sensitivity to inaccuracies in the responses and explanations, we show that users are not able to detect factual errors or biases in the provided information. The qualitative analysis shows that workers do not point out these inaccuracies explicitly. Similarly, they cannot identify flaws in the explanations related to response limitations. This aligns with previous research, demonstrating that explanations might cause users to follow the system's advice more often, even when it is wrong [54]. Our study is not conclusive about the preferred way of presenting explanations to the user. We find that limitations tend to receive higher user ratings when presented in a textual form, whereas, for confidence, we observe the opposite trend, which complies with the findings reported in the field of recommender systems [48]. Additionally, limited mentions of the presentation mode in the free-text feedback obtained from crowd workers may imply that the format of explanations is not a crucial factor in this setting.

Insights from this study about communicating explanations to facilitate users' assessment of the provided information need to be put in a broader context of system explainability and the associated effort/gain trade-off [10]. While these explanations complement the system response with components that enable users to assess responses more objectively, they demand more time and effort than merely reading the provided response. Optimizing user gain is a complex task influenced by various factors. Firstly, the relationship between the user's gain and the effort associated with the amount of additional information is not linear; while more explanations generally increase gain, there is a tolerance threshold. Exceeding that threshold may overwhelm users, causing a drop in gain. Secondly, the overall quality of the system's response and explanations significantly impacts gain. This is evidenced by our findings: users struggle to detect flaws in provided responses when explanations contain noise or errors, and providing no explanations is more useful than providing noisy ones. Thirdly, the relevance of explanations depends on the topic's complexity and user familiarity, with more complex topics benefiting from adjusted and detailed information. Additionally, the optimal effort-gain trade-off is likely to be user-dependent, requiring personalized adjustments in the amount, level of detail, and presentation of the information, which is evidenced by various preferences for the confidence display we observed in the feedback. To our knowledge, investigating the adaptation of responses based on user preferences, previous interactions with CIS systems, and topic complexity has yet to be explored.

7 CONCLUSION

Response transparency has not received significant attention in a CIS setting. Our user study addresses this gap by examining various ways of explaining the source of the information provided by the system, the system's confidence in the response, and its limitations. We explore the effect of noise and different presentation modes of these explanations on users' assessments of responses and explanations. Results reveal lower user-reported scores when explanations contain noise, although these scores seem insensitive to the quality of the response. In terms of presentation mode, we do not observe significant differences between visual and textual explanations—suggesting that the format of explanations may not be a critical factor in this setting—but users presented with no explanations found the responses more useful. To our knowledge, this study is the first to examine response transparency in CIS, highlighting the need for further research to enhance transparency in CIS responses. In particular, some of the limitations of our experimental design can be addressed by studying the impact of response specificity and interactivity on user experience over time, analyzing user's assessment when provided with a broader context or previous interactions, and evaluating in more detail the relation between user's background knowledge and the usefulness of explanations.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. 1–13.

[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, M. 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] Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2023. Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models. arXiv:2212.08037 [cs.CL]

[5] Valeria Bolotova, Vladislav Blinov, Yukun Zheng, W. Bruce Croft, Falk Scholer, and Mark Sanderson. 2020. Do People and Neural Nets Pay Attention to the Same Words: Studying Eye-tracking Data for Non-factoid QA Evaluation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20)*. 85–94.

[6] 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.

[7] 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.

[8] Ziqiang Cao, Wenjie Li, Sujian Li, Furu Wei, and Yanran Li. 2016. AttSum: Joint Learning of Focusing and Summarization with Neural Attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics (COLING '16)*. 547–556.

[9] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. 2023. Supporting High-Uncertainty Decisions through AI and Logic-Style Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. 251–263.

[10] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. 1–12.

[11] Chun-Wei Chiang and Ming Yin. 2022. Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models. In *Proceeding of the 27th International Conference on Intelligent User Interfaces (IUI '22)*. 148–161.

[12] 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 2018 (EMNLP '18)*. 2174–2184.

[13] 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.

[14] Felipe Costa, Sixun Ouyang, Peter Dolog, and Aonghus Lawlor. 2018. Automatic Generation of Natural Language Explanations. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion (IUI '18)*. 1–2.

[15] 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.

[16] 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)*.

[17] Tim Draws. 2021. Understanding How Algorithmic and Cognitive Biases in Web Search Affect User Attitudes on Debated Topics. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. 2709.

[18] 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.

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

[20] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–29.

[21] Kevin Huang, Yun Tang, Jing Huang, Xiaodong He, and Bowen Zhou. 2019. Relation Module for Non-Answerable Predictions on Reading Comprehension. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL '19)*. 747–756.

[22] 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.

[23] Gabriella Kazai, Bhaskar Mitra, Anlei Dong, Nick Craswell, and Linjun Yang. 2022. Less is Less: When are Snippets Insufficient for Human vs Machine Relevance Estimation?. In *Advances in Information Retrieval: 44th European Conference on IR Research (ECIR '22)*. 153–162.

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

[25] Youngwoo Kim and James Allan. 2019. Unsupervised Explainable Controversy Detection from Online News. In *Advances in Information Retrieval: 41th European Conference on IR Research (ECIR '19)*. 836–843.

[26] Timo K. Koch, Lena Frischlich, and Eva Lermer. 2023. Effects of fact-checking warning labels and social endorsement cues on climate change fake news credibility and engagement on social media. *Journal of Applied Social Psychology* 53, 6 (2023), 495–507.

[27] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. 1–14.

[28] 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.

[29] 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.

[30] Gyeongbok Lee, Seung-won Hwang, and Hyunsouk Cho. 2020. SQuAD2-CR: Semi-supervised Annotation for Cause and Rationales for Unanswerability in SQuAD 2.0. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC '20)*. 5425–5432.

[31] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*. 9459–9474.

[32] Jinzhi Liao, Xiang Zhao, Jianming Zheng, Xinyi Li, Fei Cai, and Jiuyang Tang. 2022. PTAU: Prompt Tuning for Attributing Unanswerable Questions. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. 1219–1229.

[33] Q. Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. arXiv:2306.01941 [cs.HC]

[34] 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.

[35] 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)*. 1–16.

[36] Don Monroe. 2018. AI, explain yourself. *Commun. ACM* 61, 11 (2018), 11–13.

[37] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Model. User-adapt. Interact.* 27, 3-5 (2017), 393–444.

[38] 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)*.

[39] Marissa Radensky, Julie Anne Séguin, Jang Soo Lim, Kristen Olson, and Robert Geiger. 2023. "I Think You Might Like This": Exploring Effects of Confidence Signal Patterns on Trust in and Reliance on Conversational Recommender Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. 792–804.

[40] 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.

[41] 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 (ACL '18)*. 784–789.

[42] 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 Systems (CHI '22)*. 1–14.

[43] 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.

[44] 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).

[45] 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)*.

[46] Tetsuya Sakai. 2018. *Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power*. The Information Retrieval Series, Vol. 40.

[47] Ivan Sekulić, Krisztian Balog, and Fabio Crestani. 2024. Towards Self-Contained Answers: Entity-Based Answer Rewriting in Conversational Search. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '24)*. 209–218.

[48] Guy Shani, Lior Rokach, Bracha Shapira, Sarit Hadash, and Moran Tangi. 2013. Investigating confidence displays for top- N recommendations. *Journal of the American Society for Information Science and Technology* 64, 12 (2013), 2548–2563.

[49] 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.

[50] Elior Sulem, Jamaal Hay, and Dan Roth. 2022. Yes, No or IDK: The Challenge of Unanswerable Yes/No Questions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT '22)*. 1075–1085.

[51] 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.

[52] 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)*.

[53] Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M. Carroll. 2021. Exploring and Promoting Diagnostic Transparency and Explainability in Online Symptom Checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. 1–17.

[54] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* 291 (2021), 103404.

[55] Yunjie (Calvin) Xu and Zhiwei Chen. 2006. Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology* 57, 7 (2006), 961–973.

[56] 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.

[57] Yongfeng Zhang and Xu Chen. 2020. Explainable Recommendation: A Survey and New Perspectives. *Found. Trends Inf. Retr.* 14, 1 (2020), 1–101.

[58] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT '20)*. 295–305.

[59] 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.

[60] Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2021. Retrospective Reader for Machine Reading Comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI '21, Vol. 35)*. 14506–14514.

[61] Lei Zhong, Juan Cao, Qiang Sheng, Junbo Guo, and Ziang Wang. 2020. Integrating Semantic and Structural Information with Graph Convolutional Network for Controversy Detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL '20)*. 515–526.