Grounded and Transparent Response Generation for Conversational Information-Seeking Systems

Weronika Łajewska, Krisztian Balog

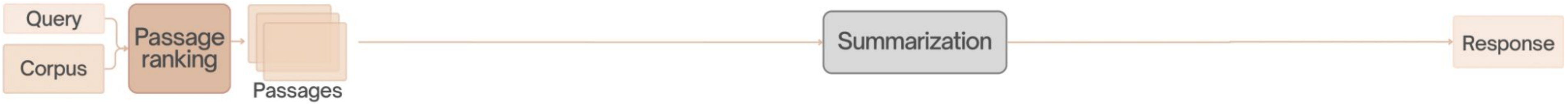
University of Stavanger, Norway

Our Motivation

- Conversational search is a less transparent setting that 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 in extremely difficult for users without knowledge about the topic

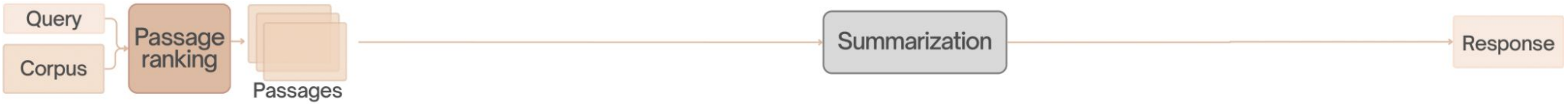
Overview of our Approach to Conversational Response Generation

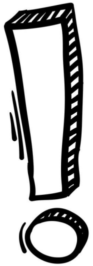
"A true teacher would never tell you what to do. But he would give you the knowledge with which you could decide what would be best for you to do."
— Christopher Pike, Sati



Overview of our Approach to Conversational Response Generation

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



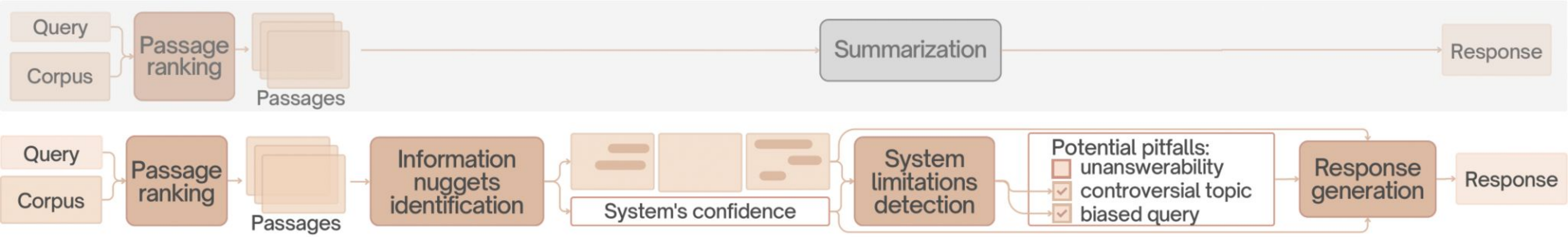


What can go wrong?

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

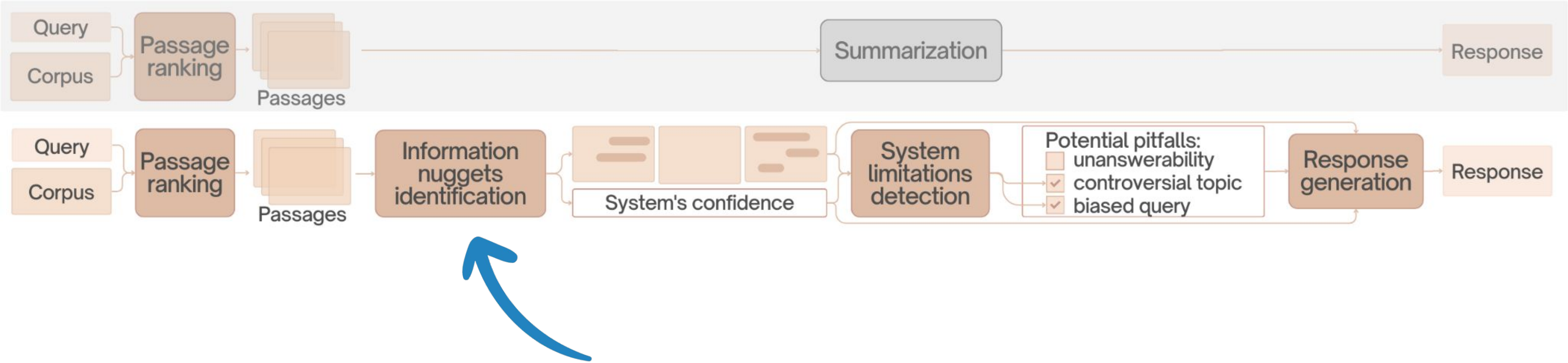
Overview of our Approach to Conversational Response Generation

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



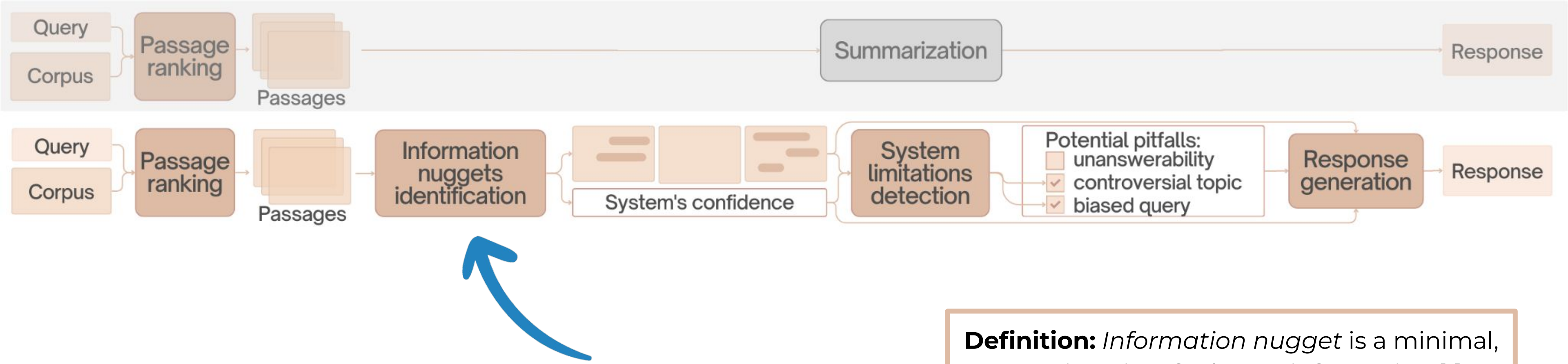
Overview of our Approach to Conversational Response Generation

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



Overview of our Approach to Conversational Response Generation

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



Definition: *Information nugget* is a minimal, atomic units of relevant information [1]

[1] Pavlu, Virgil et al. "IR system evaluation using nugget-based test collections." Web Search and Data Mining (2012).

Towards Filling the Gap in Conversational Search: From Passage Retrieval to Conversational Response Generation

Weronika Łajewska and Krisztian Balog

University of Stavanger, Norway

32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 2023

This Study

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

CAsT-snippets Sample

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

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

CAsT-snippets Sample

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

Passage: HOME - UN Climate Change Conference (COP26) at the SEC – Glasgow 2021 Uniting the world to tackle climate change. The UK will host the 26th UN Climate Change Conference of the Parties (COP26) in Glasgow on 1 – 12 November 2021. The COP26 summit will bring

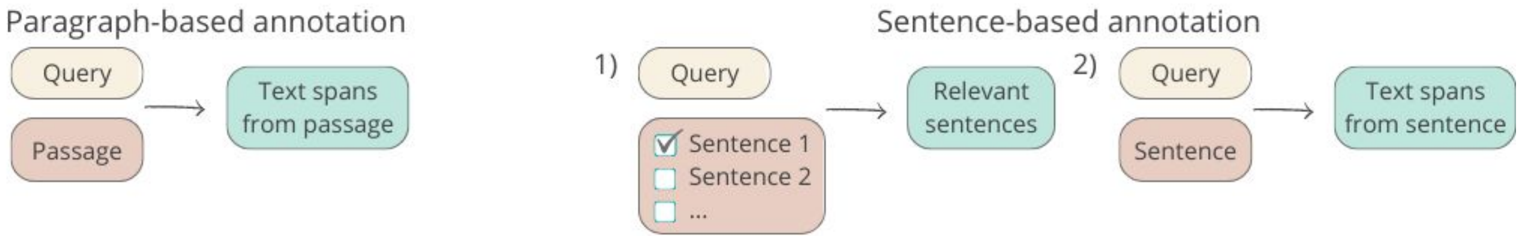
The seemingly straightforward task of highlighting relevant snippets turns out to be not that simple.

We need to accelerate the #RaceToZero Join wef, MPPindustry, topnigel & gmunozabogabir for a series of events demonstrating the need for systemic change to accelerate the global transition to net zero. Starting May 27th Learn more #ClimateBreakthroughs | #COP26 Twitter 1397069926800654339 COP26 COP26 · May 24, 2021 1396737733649846273 #TechForOurPlanet is a new challenge programme for #CleanTech startups to pilot and showcase their solutions at #COP26! Innovators can apply to six challenges focusing around core climate issues and government priorities.

Preliminary Study

A comparison of different task designs, platforms, and worker pools

- **Task designs**: paragraph-based vs. sentence-based annotation

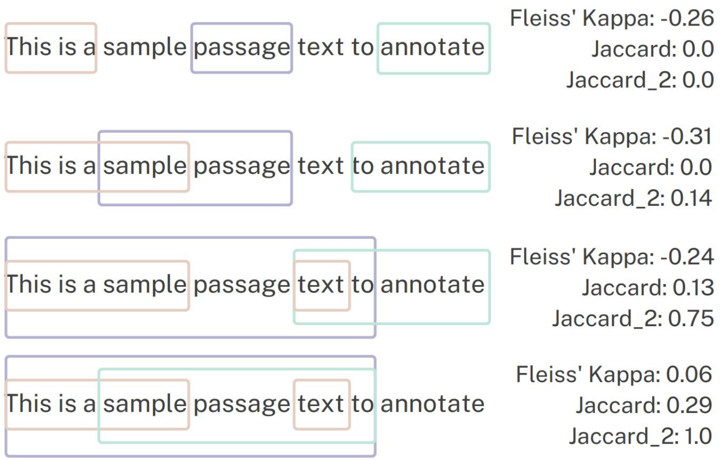


- **Platforms and workers:**
  - Amazon MTurk (regular vs. master workers)
  - Prolific
  - Expert annotators (PhD students)

Evaluation Measures

Traditional measures of inter-annotator agreement are insufficient

- Fleiss’ Kappa and Krippendorff’s Alpha are measures for categorical annotations that rely on a binary notion of agreement



Evaluation Measures

Traditional measures of inter-annotator agreement are insufficient

- Fleiss' Kappa and Krippendorff's Alpha are measures for categorical annotations that rely on a binary notion of agreement
- **Here:** we need to measure the degree to which snippets selected by different workers overlap
    - Inter-annotator agreement: Jaccard similarity (also a less strict variant, k-Jaccard)

$$J(t) = \frac{|\bigcap_{i=1}^{n} snippets(t, w_i)|}{|\bigcup_{i=1}^{n} snippets(t, w_i)|},$$

Evaluation Measures

Traditional measures of inter-annotator agreement are insufficient

- Fleiss' Kappa and Krippendorff's Alpha are measures for categorical annotations that rely on a binary notion of agreement
- **Here:** we need to measure the degree to which snippets selected by different workers overlap
  - Inter-annotator agreement: Jaccard similarity (also a less strict variant, k-Jaccard)
  - Similarity to expert annotators: "ROUGE-like" variant of precision and recall

$$p_t^{i,j} = \frac{|snippets(t, w_i) \cap snippets(t, e_j)|}{|snippets(t, w_i)|},$$

$$r_t^{i,j} = \frac{|snippets(t, w_i) \cap snippets(t, e_j)|}{|snippets(t, e_j)|}.$$

Results

Inter-annotator agreement

Task variant	Annotators	Jaccard	Jaccard_k		
			k = 4	k = 3	k = 2
Paragraph-based	MTurk regular (n=5)	0.02	0.08	0.21	0.48
	MTurk master (n=5)	0.18	0.35	0.53	0.73
	Prolific (n=5)	0.14	0.27	0.44	0.65
	Expert (m=3)	0.25	-	-	0.54
Sentence-based	MTurk regular (n=3)	0.35	-	-	0.71
	MTurk master (n=3)	0.47	-	-	0.76

Similarity to expert annotations

Task variant	Annotators	F1
Paragraph-based	MTurk regular	0.36
	MTurk master	0.54
	Prolific	0.50
Sentence-based	MTurk regular	0.31
	MTurk master	0.41

Main findings

- Relative ordering: MTurk masters > Prolific > MTurk regular
- Paragraph-level > sentence-level (w.r.t. similarity with expert annotations)

⇒ use MTurk and paragraph-based design for the large-scale data collection

Data collection

Setup

Employ a small group of trained crowd workers, selected through a qualification task, and create an extended set of guidelines with help of the annotators

Qualification task

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

20 workers completed/15 passed

Initial guidelines

Discussion

Feedback on qualification task

Extended guidelines

Data collection

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

Individual feedback after each submitted batch

General comments/suggestions on a common Slack channel

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

Resulting Dataset: CAsT-snippets

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

- Data quality
  - Inter-annotator agreement exceeds even that of expert annotators
  - Similarity with expert annotations is on par with MTurk master workers
- Comparison against other datasets
  - More snippets annotated per input text; also, snippets are longer

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

[1] Pengjie Ren, Zhumin Chen, Zhaochun Ren, E. Kanoulas, Christof Monz, and M. de Rijke. 2021. Conversations with Search Engines: SERP-based Conversational Response Generation. ACM Transactions on Information Systems 39, 4 (2021), 1–29
[2] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In Findings of the Association for Computational Linguistics: EMNLP 20 (EMNLP '18). 2174–2184.

Challenges Identified

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

- Only a partial answer is present
- Temporal considerations
  - Spans may need to be excluded given the time constraints in the query
  - Assessing temporal validity can be challenging based on the paragraph alone (without larger context)
- Subjectivity of the passages originating from blogs or comments
- Indirect answers that require reasoning and background knowledge
- Determining the appropriate amount of context to include in each span
  - Balancing between being concise and being self-contained
- Determining whether the evidence or additional information is needed or an entity alone is sufficient as an answer

Summary

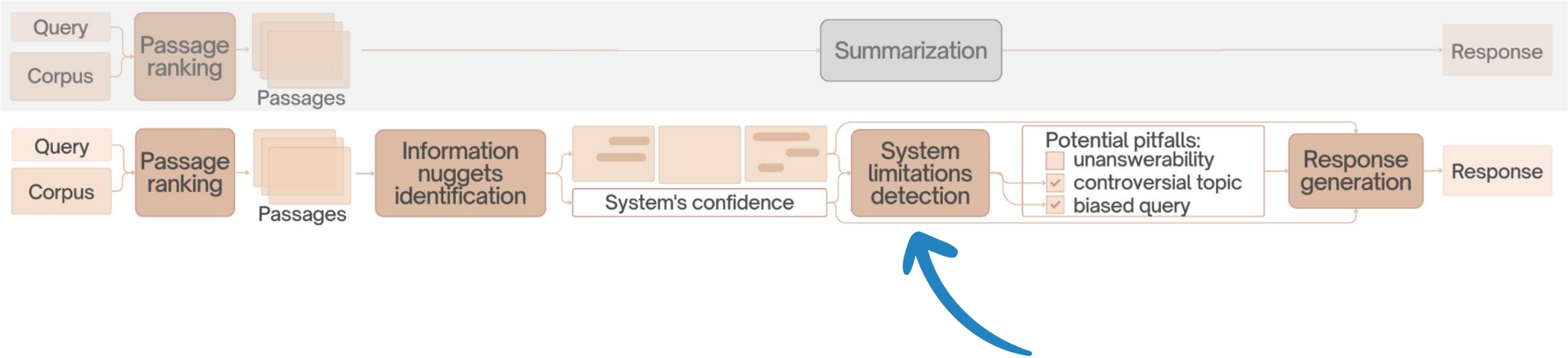
- Snippet-level annotations for conversational response generation (information-seeking queries)
- Several measures to ensure high data quality
  - Preliminary study to compare task variants and crowdsourcing platforms
  - Providing feedback and training to annotators throughout the data collection process
  - Incentive structure to engage crowd workers over a period of time and avoid worker fatigue
- Communication with workers also led to various insights regarding challenges in conversational response generation

---

Extended version on arXiv: https://arxiv.org/abs/2308.08911
Dataset: https://github.com/iai-group/CAsT-snippets

Overview of our Approach to Conversational Response Generation

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



Towards Reliable and Factual Response Generation: Detecting Unanswerable Questions in Information-Seeking Conversations

Weronika Łajewska and Krisztian Balog
University of Stavanger, Norway

46th European Conference on Information Retrieval (ECIR '24), March 2024

This Study

- **Problem setting:** Conversational information-seeking dialogue
- **Goal:** mechanism for detecting unanswerable questions for which the correct answer is not present in the corpus or could not be retrieved
- **Main contributions:**
  1. A dataset with answerability labels on three levels:
     i. sentences
     ii. paragraphs
     iii. rankings

	Answerable?	
	Yes	**No**
#question-sentence pairs (train+test)	6,395	19,043
#question-passage pairs (train+test)	1,778	1,932
#question-ranking pairs (test)	4,035	504

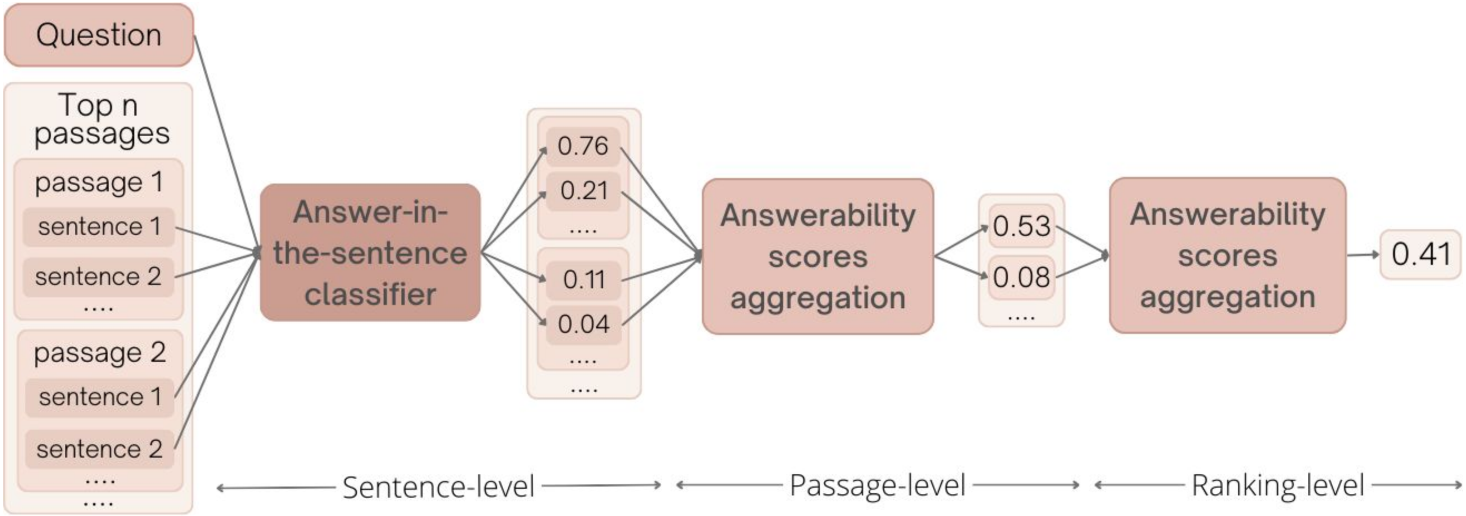
  2. A baseline approach for predicting answerability based on the top retrieved results.

CAsT-answerability Dataset

What's important for me to know about the safety of smart garage door openers?

Passage	Sentence	Answerability: Sentence	Answerability: Passage	Answerability: Ranking
MARCO_7107975	If you're looking to get a little more creative with your smart home gadgetry, try out something like Garageio.	0	1	1
	Echo can connect with this device to tell you if you've left your garage door open.	1		
	You can even say, Alexa, tell Garageio to close my garage door, and she will.	1		
MARCO_8270733	The Good The Chamberlain MyQ Garage is one of the most affordable smart garage-door openers, and also one of the easiest to install.	0	0	
	The Bad It works with a growing list of other smart home products, but notables like SmartThings and Revolv still don't have official support.	0		
	The Bottom Line Chamberlain's MyQ Garage should be the first on your list if you want to add some smarts to your garage door.	0		
	The MyQ isn't a garage door opener as it says in the headline, it's the equivalent of a remote for your existing garage door opener.	0		
	It works well and does exactly what you'd expect.	0		
MARCO_8270735	The LiftMaster MyQ Home and Property Control App empowers you to easily monitor and control your home or business from anywhere with your iPhone, or iPod touch.	0	1	
	Imagine receiving an alert if you left your garage or gate open.onitor and control your garage door, gate, commercial door and home lighting from anywhere with your smartphone.	1		
	*** Note: Requires LiftMaster MyQ hardware and a compatible garage door opener, gate operator or commercial door operator.	1		
	Learn more about compatible products and find a LiftMaster Dealer at LiftMaster.com.	0		

Overview of our Answerability Detection Approach



Results

- Data augmentation helps answerability detection only on sentence and answer levels
- *Max* aggregation on the passage level followed by *mean* aggregation on the ranking level gives the best results
- LLMs have a limited ability to detect answerability without additional guidance.

Classifier	Sentence	Passage		Ranking	
	Acc.	**Aggr.**	**Acc.**	**Aggr.**	**Acc.**
CAsT-answerability	0.752	Max	0.634	Max	0.790
				Mean	**0.891**
		Mean	0.589	Max	0.332
				Mean	0.829
CAsT-answerability augmented with SQuAD 2.0	**0.779***	Max	0.676*	Max	0.810*
				Mean	0.848*
		Mean	0.639*	Max	0.468*
				Mean	0.672*
ChatGPT passage-level (zero-shot)			**0.787***	T=0.33	0.839*
				T=0.66	0.623*
ChatGPT ranking-level (zero-shot)					0.669*
ChatGPT ranking-level (two-shot)					0.601*

Results

Does data augmentation help answerability detection?

- Data augmentation helps answerability detection only on sentence and answer levels
- *Max* aggregation on the passage level followed by *mean* aggregation on the ranking level gives the best results
- LLMs have a limited ability to detect answerability without additional guidance.

Classifier	Sentence	Passage		Ranking	
	Acc.	**Aggr.**	**Acc.**	**Aggr.**	**Acc.**
CAsT-answerability	0.752	Max	0.634	Max	0.790
				Mean	**0.891**
		Mean	0.589	Max	0.332
				Mean	0.829
CAsT-answerability augmented with SQuAD 2.0	**0.779***	Max	0.676*	Max	0.810*
				Mean	0.848*
		Mean	0.639*	Max	0.468*
				Mean	0.672*
ChatGPT passage-level (zero-shot)			**0.787***	T=0.33	0.839*
				T=0.66	0.623*
ChatGPT ranking-level (zero-shot)					0.669*
ChatGPT ranking-level (two-shot)					0.601*

Results

Which of the two aggregation methods performs better?

- Data augmentation helps answerability detection only on sentence and answer levels
- *Max* aggregation on the passage level followed by *mean* aggregation on the ranking level gives the best results
- LLMs have a limited ability to detect answerability without additional guidance.

Classifier	Sentence	Passage		Ranking	
	Acc.	Aggr.	Acc.	Aggr.	Acc.
CAsT-answerability	0.752	Max	0.634	Max	0.790
				Mean	$\boxed{\mathbf{0.891}}$
		Mean	0.589	Max	0.332
				Mean	0.829
CAsT-answerability augmented with SQuAD 2.0	**0.779***	Max	0.676*	Max	0.810*
				Mean	0.848*
		Mean	0.639*	Max	0.468*
				Mean	0.672*
ChatGPT passage-level (zero-shot)			**0.787***	T=0.33	0.839*
				T=0.66	0.623*
ChatGPT ranking-level (zero-shot)					0.669*
ChatGPT ranking-level (two-shot)					0.601*

Results

How competitive are these baselines in absolute terms?

- Data augmentation helps answerability detection only on sentence and answer levels
- *Max* aggregation on the passage level followed by *mean* aggregation on the ranking level gives the best results
- LLMs have a limited ability to detect answerability without additional guidance.

Classifier	**Sentence**	**Passage**		**Ranking**	
	Acc.	**Aggr.**	**Acc.**	**Aggr.**	**Acc.**
CAsT-answerability	0.752	Max	0.634	Max	0.790
				Mean	**0.891**
		Mean	0.589	Max	0.332
				Mean	0.829
CAsT-answerability augmented with SQuAD 2.0	**0.779***	Max	0.676*	Max	0.810*
				Mean	0.848*
		Mean	0.639*	Max	0.468*
				Mean	0.672*
ChatGPT passage-level (zero-shot)			**0.787***	T=0.33	0.839*
				T=0.66	0.623*
ChatGPT ranking-level (zero-shot)					0.669*
ChatGPT ranking-level (two-shot)					0.601*

Summary

- Test collection with answerability labels for information-seeking queries

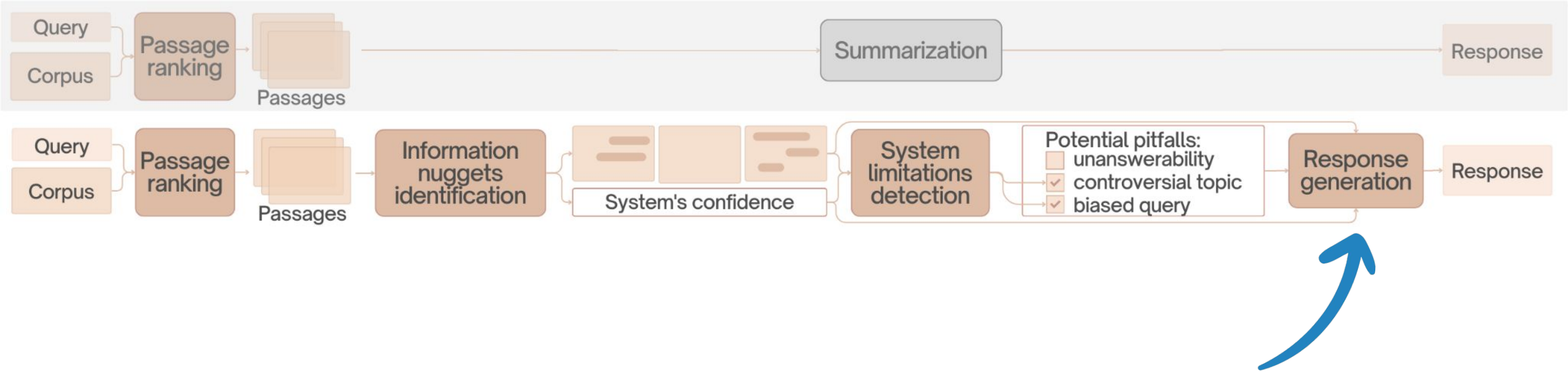
- Baseline approach based on:
  - sentence-level answerability classification
  - multi-step results aggregation

- Limitation → answerability is treated as a binary concept

Preprint on arXiv: https://arxiv.org/pdf/2401.11452
Resources: https://github.com/iai-group/answerability-prediction

Overview of our Approach to Conversational Response Generation

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



Explainability for Transparent Conversational Information-Seeking

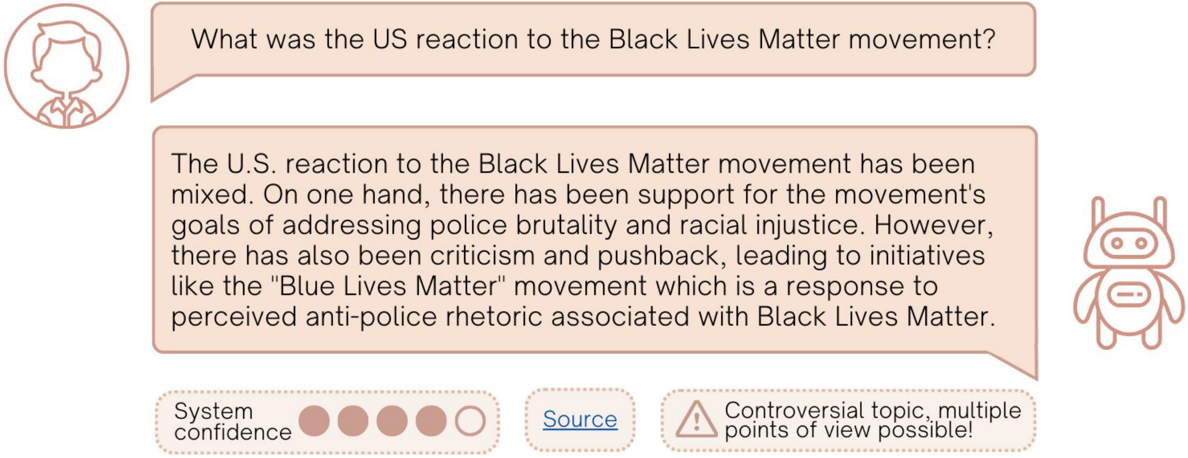
Weronika Łajewska[1], Krisztian Balog[1], Damiano Spina[2], Johanne Trippas[2]

[1]University of Stavanger, Norway

[2]RMIT University, Melbourne, Australia

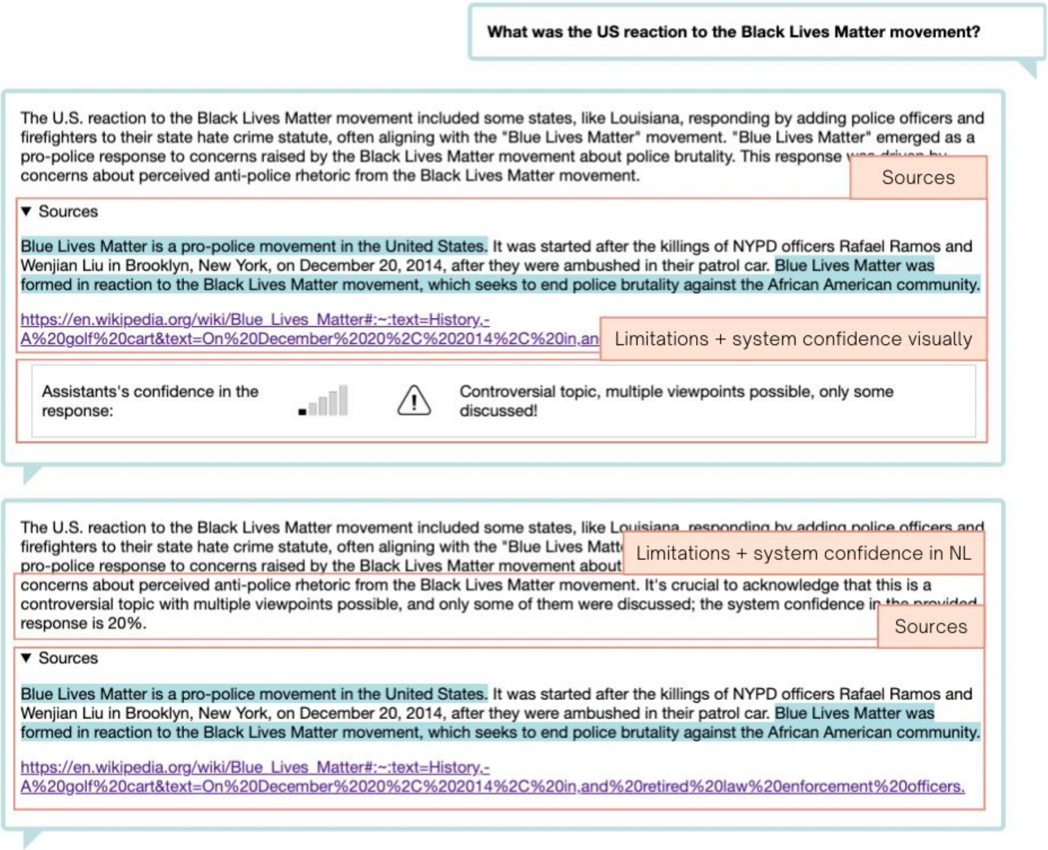
47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 2024

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 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:

- Visual
- Textual

Experimental Conditions

The selected conditions vary along three main dimensions:

1) response quality
2) quality of the explanations
3) presentation mode

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

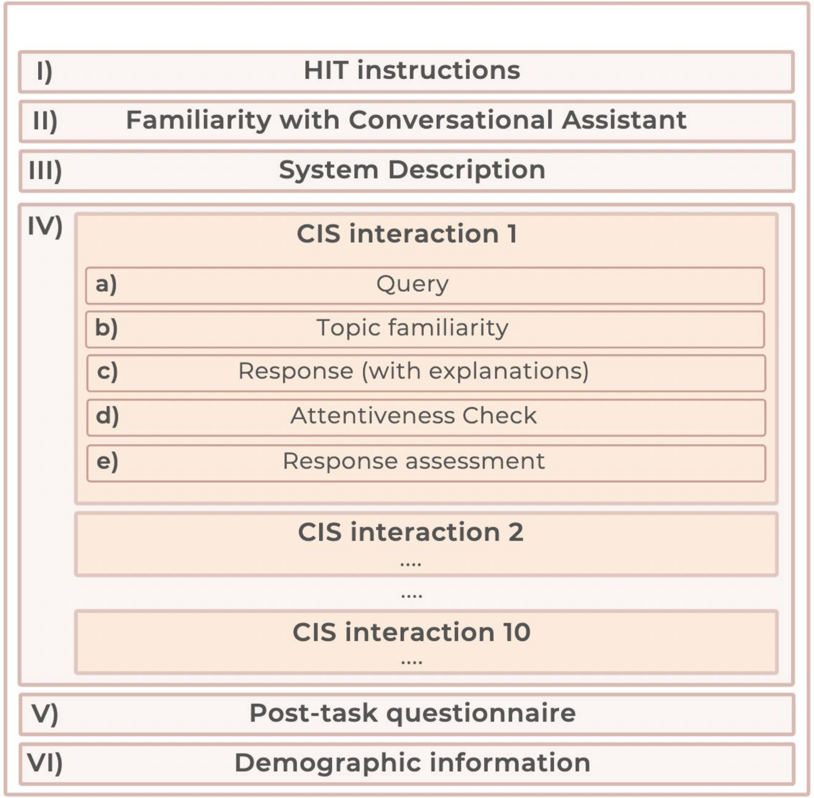
	EC1	EC2	EC3	EC4	EC5	EC6	EC7	EC8	EC9	EC10
Response	+, T	+, T	~, T	~, T	+, T	+, T	~, T	~, T	+, T	~, T
Source	+, T	+, T	+, T	+, T	~, T	~, T	~, T	~, T	–	–
Confidence	+, V	+, T	+, V	+, T	~, V	~, T	~, V	~, T	–	–
Limitations	+, V	+, T	+, V	+, T	~, V	~, T	~, V	~, T	–	–

- **+** *component without noise*
- **~** *component with inaccuracies*
- **–** *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

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**
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?

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)

- 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 some of the problems with the response without expert knowledge about the topic**

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
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)**

- 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											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)**

- 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

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)**	$\boxed{\textbf{0.0 (S)}}$	**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)**

- 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

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)**

- 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

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)**

- 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

Results

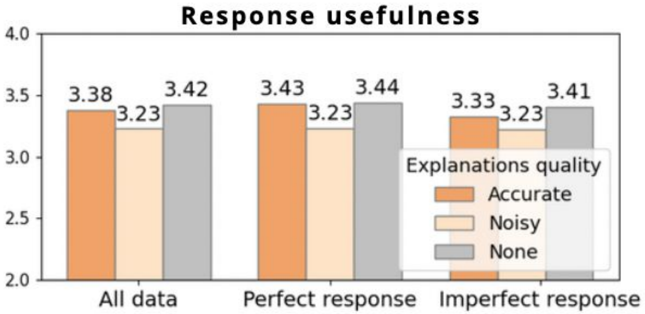
User's Perception of Explanations

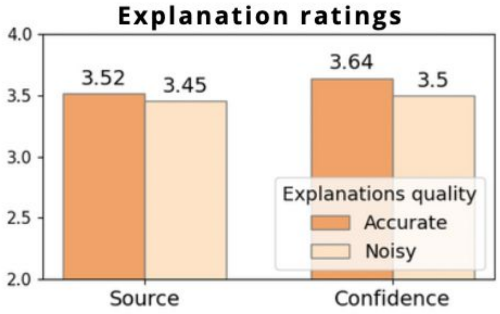
	Usefulness	Other Dimensions											Explanation		
		Rel.	Correct.	Compl.	Comprehen.	Conciseness	Serendipity	Coherence	Factuality	Fairness	Read.	Sat.	Source	Conf.	Limitation
All conditions (EC1–EC10)															
Explanation Quality	**0.0 (S)**	**0.0 (S)**	0.508 (S)	**0.003 (S)**	**0.0 (S)**	**0.001 (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)**

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

Results

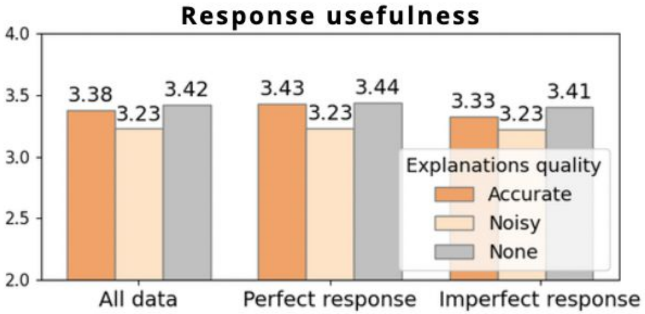
Effect of the Explanation Quality

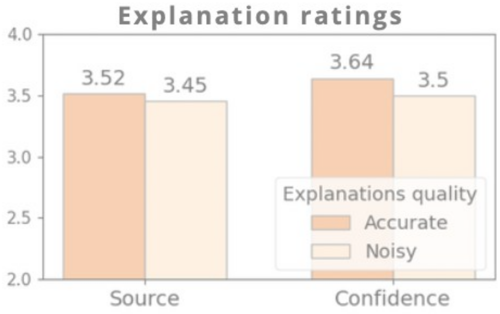




Results

Effect of the Explanation Quality





- 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

Results
Effect of the Explanation Quality

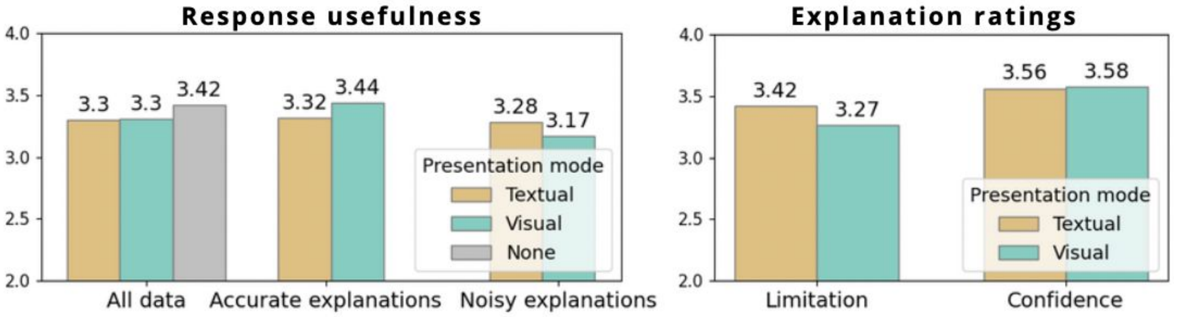
- High-quality source, system confidence score, and information about the response limitations make the response more useful from the user's perspective

- The explanations either pollute the response or make the user more critical about it, in both cases resulting in reduced usefulness

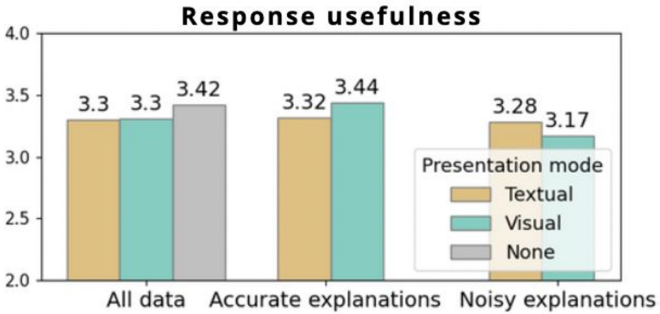
- Users perceive noisy explanations as less useful in understanding system confidence and attributed sources

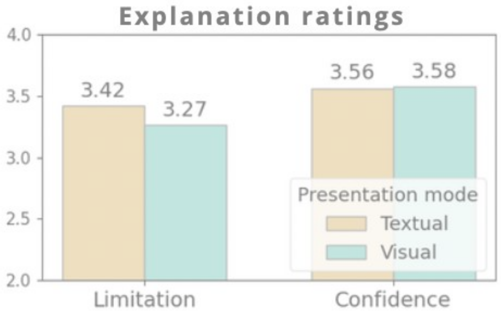
Results

Effect of the Presentation Mode



Results
Effect of the Presentation Mode

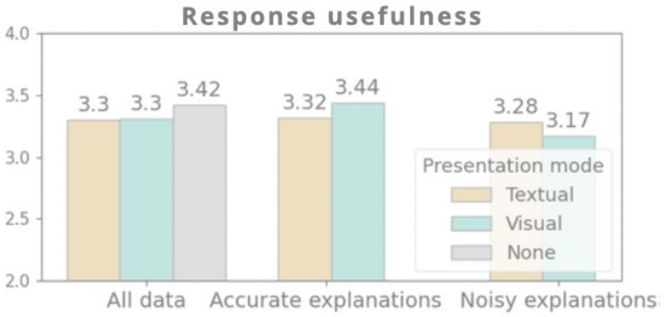


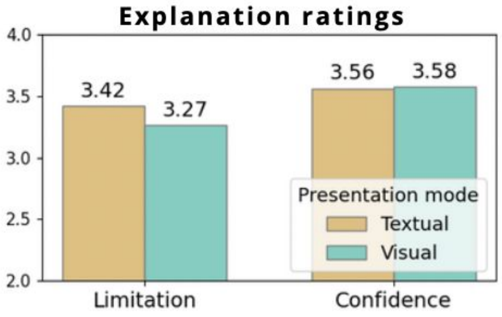


- The critical decision lies not in the method of presenting information but rather in determining whether the explanations are necessary → **trade-off between effort and gain**

Results

Effect of the Presentation Mode





- 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

Results
Qualitative Analysis

Number of comments stating that …

- … explanations enhance the understanding of the constraints of the system and the response → *18/160*
- … responses restricted to three sentences and a single source are insufficient in certain situations → *4/160*
- … interpreting explanations related to limitations and confidence scores is challenging → *3/160*
- … there is a mismatch between the source and the response → *0/160*

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 shared also for noisy explanations)

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:
  - high-quality explanations increase 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 in not a critical factor in this setting
  - user gain and effort trade-off (on explanations is more useful than providing noisy ones)
- Future work:
  - investigating 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

Preprint and resources will be available soon!

Thank you for your attention!

Questions?

Preliminary Study

Dataset: TREC CAsT'20 and '22 (top 5 passages according to relevance score for each query)

Input: query + passage/sentence

Output: snippet-level annotations in passage

Task Variant	Annotator	Time	# workers	Acceptance rate	Cost
Paragraph	MTurk regular	182s	5	50%	$0.36
	MTurk master	63s	5	90%	$0.38
	Prolific	154s	5	79%	$0.51
	Expert	96s	3	-	-
Sentence	MTurk regular	977s	3	72%	$0.43
	MTurk master	305s	3	87%	$0.56

Results (Large-scale Data Collection)

Inter-annotator agreement

Task variant	Annotator	Jaccard	Jaccard_2
Paragraph -based	MTurk regular (n=5)	0.02	0.48
	MTurk master (n=5)	0.18	0.73
	Prolific (n=5)	0.14	0.65
	Expert (m=3)	0.25	0.54
	Large-scale (topics 1,2) (m=3)	**0.38**	**0.62**
	Large-scale (all data) (m=3)	**0.33**	**0.61**
Sentence -based	MTurk regular (n=3)	0.35	0.71
	MTurk master (n=3)	0.47	0.76

Comparison to expert annotations

Task variant	Annotator	F1
Paragraph -based	MTurk regular	0.36
	MTurk master	0.54
	Prolific	0.50
	Large-scale (topics 1,2) (m=3)	**0.54**
Sentence -based	MTurk regular	0.31
	MTurk master	0.41

Amazon MTurk - Paragraph-based Design

Your task is to identify all the text spans that contain key pieces of the answer to a given question.

Text spans should contain a **single piece of information**, be **as short as possible** while **self-contained**, and **can not overlap**.

Highlight the text spans in this passage that should be included in the answer to the question **Cool. Can you tell me how to make a moisturizer at home?**

You'll receive a crumbly, waxy substance. Here's how to turn it into your own homemade moisturizer -- a lovely luxury for yourself, and a wonderful gift too. This is my personal recipe, which I've used almost exclusively as a moisturizer -- face, hands, elbows, everything -- for over a year. Sadly, it has not yet reversed the aging process -- but my skin is noticeably healthier. That's good enough for me. Ingredients 8 ounces (1 cup) of raw shea butter* 3 ounces of extra virgin olive oil, jojoba oil or another non-comedogenic nut oil 1 teaspoon of vitamin E oil Essential fragrance oils (I like almond and orange) *If you're a curly girl like me, make a hair cream by halving the amount of shea and adding 4 ounces of coconut oil. Method Place the shea butter in a small metal bowl. Put the bowl into a pot of water and heat it slowly, stirring occasionally. When the shea butter is soft enough to stir but not melted (it will be lumpy), add the olive and E oils. Whip the mixture to high heaven with an egg beater. To speed it up, try whipping on high speed for five minutes, then putting the bowl in the fridge for five minutes.

Amazon MTurk - Sentence-based Design

Instructions:

Choose **all** sentences that contain information that should be included in the answer to the question.

Task:

Question: **How much would making my own deodorant cost?**

- [ ] Before You Start, You'll Need Coconut oil (or 1/2 as much of a liquid oil if you are allergic to coconut oil) shea butter , cocoa butter or mango butter (or a mix of all three) beeswax (pastilles)
- [ ] Optional: Vitamin E oil baking soda (Omit this if you have sensitive skin and just use extra arrowroot) organic arrowroot powder or non-gmo cornstarch 2-3 capsules of high quality probiotics that don't need to be refrigerated (I love Bio Kult brand)- optional
- [ ] Optional: Essential oils of choice – I used about 20 drops of lavender essential oil Deodorant Bar Ingredients ½ cup coconut oil ½ cup shea butter , cocoa butter or mango butter (or a mix of all three equal to 1 part) ½ cup + 1 tsp beeswax 1 teaspoon Vitamin E oil – optional 3 tablespoons baking soda
- [ ] (Omit this if you have sensitive skin and just use extra arrowroot or cornstarch) 1/2 cup organic arrowroot powder 2-3 capsules of high quality probiotics that don't need to be refrigerated (optional)
- [ ] Optional: Essential oils of choice – I used about 20 drops of lavender essential oil and also like citrus and frankincense Deodorant Bar
- [ ] Instructions Combine coconut oil, shea (or other) butter, and beeswax in a double boiler, or a glass bowl over a smaller saucepan with 1 inch of water in it.

Submit

Your task is to identify all the text spans that contain key pieces of the answer to a given question.

Text spans should contain a **single piece of information**, be **as short as possible** while **self-contained**, and **can not overlap**.

Highlight the text spans in this sentence that should be included in the answer to the question **I remember Glasgow hosting COP26 last year, but unfortunately I was out of the loop. What was the conference about?**

If countries cannot agree on sufficient pledges, in another 5 years, the emissions reduction necessary will leap to a near-impossible 15.5% every year.

Prolific

Paragraph-based Design

Snippet annotation task 1

Identify all the text spans that contain key pieces of the answer to a given question

Instructions

Your task is to identify all the text spans that contain key pieces of the answer to a given question.

Text spans should contain a **single piece of information**, be **as short as possible** while **self-contained**, and **can not overlap**.

In **each** passage highlight the chosen text spans using green text highlight:

Do not edit the text of the passages!

Task

Question 1:

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

Passage 1:

The initial pledges of 2015 are insufficient to meet the target, and governments are expected to review and increase these pledges as a key objective this year, 2021. The updated Paris Agreement commitments will be reviewed at the climate change conference known as COP 26 in Glasgow, UK in November 2021. This conference will be the most important ...

Questionnaires

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?

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

User's Perception of Explanations

	Explanation		
	Source	**Conf.**	**Limitation**
	All conditions (EC1–EC10)		
Explanation Quality	**0.0 (S)**	**0.0 (S)**	0.173 (S)
Presentation Mode	**0.0 (S)**	**0.0 (S)**	**0.0 (S)**
	Only conditions with explanations (EC1–EC8)		
Explanation Quality	0.097 (S)	**0.0 (S)**	0.088 (S)
Presentation Mode	**0.0 (S)**	0.653 (–)	**0.0 (S)**

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

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	Source	Explanation 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)**