The Dangers of trusting Stochastic Parrots: Faithfulness and Trust in Open-domain Conversational Question Answering

Sabrina Chiesurin* **Dimitris Dimakopoulos*** **Marco Antonio Sobrevilla Cabezudo**
Arash Eshghi **Ioannis Papaioannou** **Verena Rieser†** **Ioannis Konstas**
Alana AI
hello@alanaai.com

Abstract

Large language models are known to produce output which sounds fluent and convincing, but is also often wrong, e.g. "unfaithful" with respect to a rationale as retrieved from a knowledge base. In this paper, we show that task-based systems which exhibit certain advanced linguistic dialog behaviors, such as lexical alignment (repeating what the user said), are in fact preferred and trusted more, whereas other phenomena, such as pronouns and ellipsis are dis-preferred. We use open-domain question answering systems as our test-bed for task based dialog generation and compare several open- and closed-book models. Our results highlight the danger of systems that appear to be trustworthy by parroting user input while providing an unfaithful response.

1 Introduction

With the advent of large language models (LLM), Question Answering Systems have become open-domain and conversational, meaning that they are able to generate fluent and informative responses to questions about nearly any topic and over several turns (Adlakha et al., 2022). However, these systems are also known to produce factually incorrect statements, commonly referred to as *hallucinations* (Rashkin et al., 2021b; Dziri et al., 2022b). These two properties taken together require the system as well as the user to ensure that they mutually understand each other – a process also known as *conversational grounding* (Clark and Brennan, 1991).

Empirical studies of dialogue have shown that people use different kinds of context-dependent linguistic behavior to indicate grounding, including use of fragments, ellipsis and pronominal reference (Fernandez and Ginzburg, 2002; Eshghi and Healey, 2016). Other studies show that lexical alignment in a response, i.e. repeating and adopting the interlocutor's lexical items (Pickering and Garrod, 2004; Branigan et al., 2010), can play a similar role, see examples in Figure 1.

Question: When *will* the new *Dunkirk film be released on DVD?*

Rationale:
"*Dunkirk*" was *released* digitally on 12 December **2017**, and on 4K Ultra HD, Blu-ray, and *DVD* on ***18 December*** in the ***United Kingdom*** and ***19 December*** in the ***United States***

Linguistic Phenomenon	Answer	User Pref.	Faith.
Lexical Alignment	The new *Dunkirk film will be released on DVD* on **September 19, 2017**.	✓	✗
Pronominal	It will be on **18 December 2017**		✓
Fragment	**18 December 2017**		✓

Figure 1: Responses with different forms of conversational linguistic phenomena and token grounding: **Blue** indicates tokens from the question are repeated in the response (*lexically aligned*). **Bold** corresponds to content tokens in the response *grounded* in the knowledge source; **red** tokens are hallucinations, i.e., not *faithful* to the dialogue and rationale. The last two columns indicate user preference and faithfulness, respectively.

There is initial evidence in related fields that generating grounding phenomena will lead the user to trust the system more, such as conversational assistants for educational (Linnemann and Jucks, 2018) and medical applications (Bickmore et al., 2021) as well as in the field of HRI (Bossens and Evers, 2022). At the same time, we argue that systems that exhibit more grounding behavior are not necessarily more faithful to the dialogue and input rationale, which can lead to unjustified trust.

In order to explore these hypotheses, we first analyze conversational grounding phenomena via automatic annotation of linguistic properties for open-domain QA. We consider responses generated by different GPT-3 variants (Brown et al., 2020), and state-of-the-art Retrieve-and-Generate models on the TopiOCQA development set (Adlakha et al., 2022). We evaluate the performance of models via several automatic surface-level, and semantic-based metrics against multiple references and a chosen rationale from a gold Wikipedia passage.

*Equal Contribution.
†Now at Google DeepMind.

Findings of the Association for Computational Linguistics: ACL 2023, pages 947–959
July 9-14, 2023

Models	Length (μ)	Structure (%)			Align			Pron (%)
		Frag	Short	Long	P	R	F1	
DPR+FiD	9.1	64.6	33.4	2.0	6.1	8.7	6.2	23.1
DPR+GPT-3	24.4	13.9	56.7	29.4	14.8	37.0	19.5	10.4
GPT-3	20.1	12.5	55.8	31.7	18.3	38.0	22.9	12.1
Human	11.2	57.0	36.2	6.8	6.6	10.8	7.2	18.7

Table 1: Linguistic phenomena of responses for different models on the development set of TopiOCQA.

Given current limitations of automatic metrics, we annotate a subset of responses according to their plausibility, groundedness to the input source and faithfulness to the dialogue and input source *at the same time*. We also elicited a human preference task among the responses of each model. Finally, we conduct a series of human evaluation experiments where we provide responses to questions controlling for each of the linguistic phenomena under examination, and ask users to choose the one they perceive as more trustworthy. Our findings are summarised as follows:

- GPT-3 variants are generally more verbose and more lexically aligned to the question. In contrast, the human-authored responses in TopiOCQA are more elliptical and contain more pronominals. Unsurprisingly, the fine-tuned model emulates this behavior.
- GPT-3 variants are less faithful according to expert human annotations and the majority of automatic metrics.
- Surprisingly, users prefer open-book GPT-3 over the fine-tuned model although half of the time the preferred responses were unfaithful.
- Users trusted responses with high lexical alignment significantly more, whereas the effect was the opposite for elliptical responses, and answers containing pronominals.

2 Conversational Grounding Analysis

2.1 Dataset and Models

Dataset We use the development set of TopiOCQA comprising 205 information-seeking dialogues (2514 turns)[1].

Models We test a variety of models under two different settings. In the *closed-book* setting models have no access to domain-specific information other than what is stored in their own parameters; in the *open-book* setting models can leverage a set of relevant documents provided by the retriever.

For the open-book setting we used a fine-tuned Dense Passage Retriever (DPR; Karpukhin et al., 2020) as the retriever and experimented with two different readers: Fusion in Decoder (FiD; Izacard and Grave, 2021) fine-tuned on TopiOCQA, and GPT-3 (Brown et al., 2020)[2], where we concatenate passages returned from DPR with the dialogue context and use them as conversational prompt. For closed-book similar to Adlakha et al. (2022) we also use GPT-3, where the dialogue context is concatenated into a conversational prompt.

Notably, we could have also tuned GPT-3 either via prompt engineering or fine-tuning[3] so that it resembles the distribution of the target dataset. We decided against this for two reasons: firstly, the amount of engineering required would go beyond the focused scope of this work; second using vanilla GPT-3 variants is as close as possible to an ecologically valid scenario. For example, it is similar to how an end-user would be exposed to an LLM via a search engine, or a chat interface without any direct control of its prompt.

2.2 Dialogue Phenomena

We automatically annotate the following linguistic properties of responses:

Lexical Alignment is approximated based on unigram overlap between the response and corresponding question, i.e. the system repeating the same words as the user. This typically serves the purpose of implicitly confirming what was understood in task-based dialog. We compute the precision (P), recall (R) and F1. Figure 1 shows a response that lexically aligns to the question.

Syntactic Form We define three categories according to the syntactic structure, based on the constituency tree[4]:

- *short responses* comprise a single sentence

[1] A manual analysis of the dataset revealed that the linguistic phenomena under scrutiny are almost exclusively present.

[2] We used `davinci-003` in all our experiments.

[3] Fine-tuning GPT-3 would entail several rounds of hyperparameter tuning increasing the cost of the experiments.

[4] We used Stanza (Qi et al., 2020).

Models	F1 ↑	EM ↑	BLEU ↑	ROUGE ↑	BERT ↑	K-F1 ↑	K-F1++ ↑	Critic ↓	Q^2 F1 ↑	Q^2 NLI ↑
DPR+FiD	**55.3**	**33.0**	**44.74**	**56.3**	0.79	21.3	19.0	**55.9**	**32.8**	**35.9**
DPR+GPT-3	37.4	5.9	20.02	39.0	**0.81**	**28.4**	**22.6**	63.2	26.5	29.8
GPT-3	33.9	6.8	12.71	36.4	0.80	20.2	15.7	59.2	19.9	24.3
Human	70.1	40.2	58.63	70.8	0.83	33.0	29.3	20.7	59.9	63.6

Table 2: Model performance using automatic metrics on the development set of TopiOCQA.

with the tree's root being either a simple declarative clause (S), or a declarative sentence with subject-aux inversion (SINV); see the first two responses in Figure 1.

- *fragments* comprise an elliptic sentence, with its syntactic root not identified as either S or SINV; see last response in Figure 1.

- *long-form responses* are multi-sentence answers, which are rarely occurring. This is probably due to the conversational nature of TopiOCQA where complex questions are broken down into simpler ones across a dialogue.

Pronominals We identify the existence (or not) of a pronoun in a sentence in subject, or direct object position according to its dependency tree, e.g., "*It*" in the second response of Figure 1.

Table 1 summarizes the statistics of linguistic phenomena found in models and human responses. Note that GPT-3 variants produce more verbose, sentential and lexically aligned responses with the questions (see Recall column). In contrast, the fine-tuned model (DPR+FiD) generates shorter fragmented responses with more pronominals. This is expected as it follows the distribution of human responses, unlike the GPT-3 variants that have a very limited conditioning on the target distribution via the dialogue context getting encoded in the prompt.

3 Study of Faithfulness

Faithfulness Definition We extend the definition by Adlakha et al. (2022) to consider faithfulness both wrt the *dialogue* and rationale:

Given a dialogue history $\mathcal{H} = (u_1, ..., u_{n-1})$ *and knowledge* $\mathcal{K} = (k_1, ..., k_j)$ *at turn* n*, we say that* utterance u_n *is faithful with respect to* $\mathcal{K}$ *and* $\mathcal{H}$ *iff* $\exists \Gamma_n$ *such that* $\Gamma_n \models u_n \wedge E(\mathcal{H}, u_n) \neq \emptyset$*, where* $\models$ *denotes semantic consequence,* Γ_n *is a non-empty subset of* $\mathcal{K}$ *and E is the explicature of* u_n *in context* $\mathcal{H}$ as defined in (Rashkin et al., 2021a).

3.1 Automatic Evaluation

We first employ a wide range of automatic metrics to assess model performance grouped according to their similarity to a gold (human) reference (*reference-based*), or their faithfulness to the provided knowledge $\mathcal{K}$ (*reference-less*).

Reference-based metrics Following Adlakha et al. (2022) and Dziri et al. (2022a), we report F1 score, Exact Match (EM), BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). These measure the overlap-based similarity between the generated response and the gold answer[5].

Reference-less token-level metrics Similar to Dziri et al. (2022a) and Shuster et al. (2021), we report BERTScore (BERT) (Zhang et al., 2019), and Knowledge-F1 (K-F1). Notably, the latter calculates the unigram overlap between the response and a knowledge snippet $\mathcal{K}$, providing a verbatim measure of grounding to the input source.

We propose K-F1++, a variant of K-F1, that captures only the novel information in the generated response and discounts any lexical alignment to the question: it calculates the unigram overlap between the response and $\mathcal{K}$, after *subtracting* any tokens appearing in the question from the response.

Reference-less entailment metrics We report *Critic* (Dziri et al., 2022a), a dialogue-trained classifier determining if a response follows from a given snippet $\mathcal{K}$, and Q^2 (Honovich et al., 2021), which measures faithfulness via question answering.

3.2 Human evaluation studies

Similar to Glaese et al. (2022), Bai et al. (2022) and Thoppilan et al. (2022), we conducted a human evaluation to assess the faithfulness of given responses, followed by a human evaluation study to collect human preferences when presented with two possible responses to an existing conversation.

Faithfulness Judgment task Annotators are required to judge the plausibility of a response given the dialogue, the relevance of the gold passage to answer the question, and the faithfulness of the re-

[5]Note that results for Human don't go up to 100% as each output is compared with 3 additional human annotations.

sponse given the dialogue and the gold passage. In more detail, we consider the response to be grounded when it (or a paraphrase of it) is found in the document. We consider a response to be faithful if, in addition to being grounded, it answers the question and follows from the dialogue. For example, given i) a conversation about European countries, ii) a document about European capitals, iii) a query "*What is the capital of Spain?*", and iv) the response "*Castellano*", if "*Castellano*" is in the document, the response is grounded. However, it is not faithful with respect to the dialogue as it does not correctly answer the question. Two annotators[6] completed the annotation for each model on 500 instances from TopiOCQA.

Preference task Annotators are provided with a question, the previous dialogue and the gold passage that contains the answer, and are required to select their preferred response given two options. These are between a baseline model (DPR+FiD) and a model variant; they can also select both or none. We take a sample of 250 faithful and unfaithful instances from the previous task.

3.3 Results

Table 2 summarizes the automatic metrics. Baseline DPR+FiD outperforms the GPT-3 variants in all *reference-based* metrics. This is somewhat expected since the former is fine-tuned on the TopiOCQA dataset, whereas GPT-3 –despite being a much larger model– is evaluated in a zero-shot fashion. Surprisingly, DPR+GPT-3 outperforms the baseline in most *reference-less* metrics.

Interestingly, the absolute difference between K-F1 and K-F1++ with respect to the baseline (2.3%) is significantly smaller than that of the GPT-3 variants (5.8%, and 4.5%, respectively). This is probably due to the latter being more lexically aligned to the user question than the baseline (see Table 1), hence there are more overlapping tokens removed when computing K-F1++. Nevertheless, the GPT-3 variants maintain superior knowledge-grounding scores even based on the stricter K-F1++.

Table 3 paints a different story to the reference-less metrics: although all responses are regarded mostly plausible continuations to the dialogue, the GPT-3 variants (with the closed-book scoring worst) produce outputs that are less grounded and more unfaithful compared to DPR+FiD. We observed often the inclusion of extra information that could *potentially* be true but still not faithful to the input source. We leave fact checking of such extrinsic hallucinations to future work.

[6]The annotators comprise a hired annotator and one of the co-authors. Quality was ensured via multiple rounds of pilot annotations, until all disagreements were resolved.

Models	Plaus.	Ground.	Faith.
DPR+FiD	97.2	**62.4**	**57.8**
DPR+GPT-3	**100.0**	46.2	39.6
GPT-3	91.6	22.6	22.0
Human	99.8	98.6	93.0

Table 3: Faithfulness Judgement Task carried out by 2 expert annotators on a sample of 500 instances.

Model	Preferences		
	All (#)	Faith. (#)	Unfaith. (#)
DPR+FiD	33% (417)	85% (354)	15% (63)
None	12% (153)	-	-
DPR+GPT-3	**70% (883)**†	52% (459)	48% (424)
DPR+FiD	43% (539)	84% (451)	16%(88)
None	13% (173)	-	-
GPT-3	45% (559)	33%(186)	66% (373)
DPR+FiD	46% (578)	95% (547)	5% (31)
None	9% (109)	-	-
Human	**74% (931)**†	94% (879)	6% (52)

Table 4: Pair-wise Preference task results on a sample of 250 examples with 5 annotations. Baseline (DPR+FiD) is compared with GPT-3 variants, and human responses. Users can select both models or none. Total number of annotations per model is in parentheses. Last two columns denote a breakdown of selected responses that were faithful, or unfaithful. † indicates stat. sig. against the baseline using χ^2 goodness of fit ($p < .05$).

The most striking result according to the Preference task (Table 4) is that annotators preferred unfaithful responses over faithful ones, or rejected both options, even though they had access to the gold passage. DPR+GPT-3 overall was preferred 70% of times, with almost half preferences being towards unfaithful responses (48%). Similarly, GPT-3 was preferred 45% of the time with 66% of preferences being unfaithful. Again this supports our hypothesis that high lexical alignment has a great influence on users' choices, often bypassing the need to judge the accuracy of the response. Appendix A contains additional results on computing majority agreement per item among the 5 annotators for the Preference Task and a qualitative analysis of provided feedback.

4 Study of Trust

So far we have established that lexically aligned responses coming from GPT-3 variants are not necessarily faithful. The surface form seems to negatively affect users' preferences, obviating their need

Linguistic phenomena	Trust
High Lexical Alignment	**58%**†
None	10%
Low Lexical Alignment	32%
Pronouns	31%
None	19%
No Pronouns	**49%**†
Short answer	**66%**†
None	7%
Fragment	26%

Table 5: Human Evaluation experiment on Trust for various linguistic phenomena. High/Low lexical alignment threshold is set to 0.5, based on recall. † denotes pair-wise stat. sig. using χ^2 goodness of fit ($p < .05$).

to check the supporting source, and creating a risk of placing trust to an imperfect system. With this experiment, we investigate a more general trend between linguistic phenomena and user trust.

Human Evaluation Experiment Annotators are presented with the dialogue only, and are asked to choose the response they trusted more from two possible responses, or none. Going beyond just lexical alignment, we selected 15 pairs of responses[7], for every linguistic phenomenon in Section 2.2. We modified responses to ensure each specific phenomenon was the only difference between them. We collected 20 preferences for each response pair.

Results Table 5 shows that annotators trusted responses with high lexical alignment significantly more than those with low lexical alignment. Interestingly, they trusted significantly more short answers than fragments, and preferred responses that did not present pronouns. This is in contrast to literature (Eshghi and Healey, 2016), which primarily focused on human-to-human interactions; this could be down to people talking to a system (vs. a human), seeking stronger forms of evidence such as lexical alignment. Notably, the combination of the preferred presence and absence of phenomena aligns well with their calculated occurrences in the GPT-3 variants' responses (Table 1).

5 Conclusions

We investigated the performance of different models on the task of OCQA, measuring faithfulness and lexical phenomena. Automatic metrics highlighted how GPT-3 variants are less faithful than DPR+FiD, as confirmed by annotators in the faithfulness judgment task. We conducted a study on conversational grounding phenomena and a preference task, whose significant results demonstrated an effect of surface form in human preferences towards the more conversational GPT-3, even when unfaithful. Another experiment confirmed trust as being effected by high lexical alignment.

[7]Note that we select only faithful responses, explicitly informing participants.

Limitations

This work is constrained by the number of grounding phenomena analyzed, which is limited by the dataset domain and their straightforward automatic computation. We only focused on lexical alignment, the use of ellipsis (fragments) and pronouns, disregarding other phenomena such as repairs (e.g. asking for confirmation or clarification) (Purver et al., 2003), among others.

With respect to the linguistic phenomena, we simplified the calculation of the lexical alignment by regarding only the last two turns of a conversation (the user question and the system response). In this manner, we omitted the dynamic convergence over several turns (Mills and Healey, 2008). It should be noted though that this was decided based on manual observation of examples, the majority of which exhibited lexical alignment in the last two turns only. This could be a limitation of the OCQA domain, and/or a bias of the TopiOCQA dataset.

Another limitation is that the form of crowdsourcing experiments we performed are mostly diagnostic of certain conditions on a given dataset, and does not reflect more organic real-use cases. An ideal setup would be to collect whole dialogues in the form of an extrinsic evaluation, which would be more costly to perform.

Ethics Statement

Dual Use Our results highlight a possible misuse scenario, where verbally fluent but factually incorrect text generated by models, such as GPT-3, is more convincing to users than text by models which are more faithful to the input rationale. This blind trust could be exploited to convince users of e.g. fake news, for example by generating more lexically aligned text.

Human data The methodology of this paper heavily relies on human data collection using crowd-sourcing. Workers were allowed to complete a maximum of 40 HiTs across annotations. They were payed 0.29$ per HiT for the preference task, while 0.20$ per HiT for the study on trust.

Annotators come from Australia, Canada, New Zeland, United Kingdom and United States. A total of 38 annotators were involved in the study of trust, and 115 were involved in the Preference task. Data collected using AMT are fully anonymized per the providers specifications.

Use of TopiOCQA We obtained the dataset through the public domain and do not intend to release part, or whole of it separately without the prior consent of its authors. We assume the authors have taken precautions against offensive content.

Acknowledgements

We would like to particularly thank Oliver Lemon for the discussions on the linguistic phenomena in conversation and trust. We also appreciate the valuable feedback we received by the rest of the technical team at Alana AI at various stages of the paper. Finally, we would like to thank the anonymous reviewers and annotators for the human evaluation.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Timothy W Bickmore, Stefán Ólafsson, and Teresa K O'Leary. 2021. Mitigating patient and consumer safety risks when using conversational assistants for medical information: Exploratory mixed methods experiment. *J Med Internet Res*, 23(11):e30704.

David M. Bossens and Christine Evers. 2022. Trust in language grounding: a new ai challenge for human-robot teams.

Holly P Branigan, Martin J Pickering, Jamie Pearson, and Janet F McLean. 2010. Linguistic alignment between people and computers. *Journal of pragmatics*, 42(9):2355–2368.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022a. Faithdial: A faithful benchmark for information-seeking dialogue.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022b. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Arash Eshghi and Patrick G. T. Healey. 2016. Collective contexts in conversation: Grounding by proxy. *Cognitive Science*, 40(2):299–324.

Raquel Fernandez and Jonathan Ginzburg. 2002. Non-sentential utterances in dialogue: A: Corpus-based study. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, pages 15–26, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769–6781, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Gesa Alena Linnemann and Regina Jucks. 2018. 'Can I Trust the Spoken Dialogue System Because It Uses the Same Words as I Do?'—Influence of Lexically Aligned Spoken Dialogue Systems on Trustworthiness and User Satisfaction. *Interacting with Computers*, 30(3):173–186.

Gregory Mills and Pat Healey. 2008. Semantic negotiation in dialogue: the mechanisms of alignment. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 46–53, Columbus, Ohio. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003. On the means for clarification in dialogue. In *Current and new directions in discourse and dialogue*, pages 235–255. Springer.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2021a. Measuring attribution in natural language generation models. *arXiv preprint arXiv:2112.12870*.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021b. Increasing faithfulness in knowledge-grounded dialogue with controllable features. 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)*, pages 704–718, Online. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Baseline vs	Agreement
DPR+GPT-3	86.4%
GPT-3	77.6%
Human	90%

Table 6: Majority Agreement per item (5 annotations) for the Preference Task between the Baseline (DPR+FiD) and models. Each row denotes majority reached at the corresponding % of the times.

Phenomenon	Agreement
Lexical Alignment	80%
Pronouns	53%
Fragment	86%

Table 7: Majority Agreement per item (20 annotations) for the Study of Trust across the different linguistic phenomena examined in this work. Each row denotes majority reached at the corresponding % of the times.

A Additional Human Evaluation Results

Majority Agreement Results

Following Glaese et al. (2022) we computed the majority agreement for each item, i.e., 5 and 20 annotations per item for the preference and trust studies, respectively. Tables 6 and 7 summarize the results. Similar to Glaese et al. (2022) there are cases when agreement is quite low, which is an interesting avenue for future work.

Qualitative Analysis of Feedback

Next, we conducted a simple qualitative analysis regarding how often annotators looked at the grounded document during the Preference Task. 286 out of 2170 feedback responses explicitly refer to the document to justify the preference expressed. Interestingly, There are in total 558 responses where GPT-3 variants were preferred over the baseline, of which only 27 (4%) refer to the document. In contrast, there are 359 of which 76 refer to the document (21%) when the baseline is preferred. Overall, feedback suggests that GPT-3 responses were mostly preferred due to other factors, such as the amount and variety of information, and conversational style.

B Human Evaluation Instructions and Interfaces

B.1 Faithfulness Judgment Task

Figures 2 and 3 illustrate the user interface implemented for the plausibility and faithfulness subtasks, respectively.

Task Instructions:

In this task you will evaluate the quality of a system-generated response to a user query. The system is trying to help the user learn about a particular topic by answering their questions. We want to rate the system response quality based on how well it represents the sources provided. You will need to answer four questions. The first question is about plausibility. Only if the answer is plausible, you will be asked to answer other questions. Some ratings will result in other categories being skipped. The task interface will guide you through the flow.

Note: The system-generated responses may appear very fluent and well-formed, but contain slight inaccuracies that are not easy to discern at first glance. Pay close attention to the text. Read it carefully as you would when proofreading.

B.2 Preference Task

Figure 4 depicts the interface for the preference task in the context of the dialogue and gold passage.

Task Instructions:

In this task, you will continue a conversation between a system and a user by selecting your preferred answer. For each question you will see two different answers, and we want you to carefully decide which one is better. Read the Conversation carefully and find a reason to select one answer over the other. If this is not possible due to high or low quality of both answers, you can check "All completions are high quality" or "All completions are low quality" depending on the situation. A document to back up the claims made in the answers is provided.

Optional: in the feedback box, please justify your choice of best answer. Be specific about which parts of the question and answers mattered in your choice, especially when comparing two satisfactory answers.

B.3 Study of Trust

Figure 5 shows a screenshot of the trust task given the dialogue only.

Task Instructions:

In this task, you will continue a conversation between a system and a user by selecting most trustworthy response. For each question you will see two different answers, and we want you to carefully decide which one is most trustworthy. If you cannot decide between the two, you can check "I can't decide". Note that all answers provided are correct.

Optional: if you can't decide, please write why.

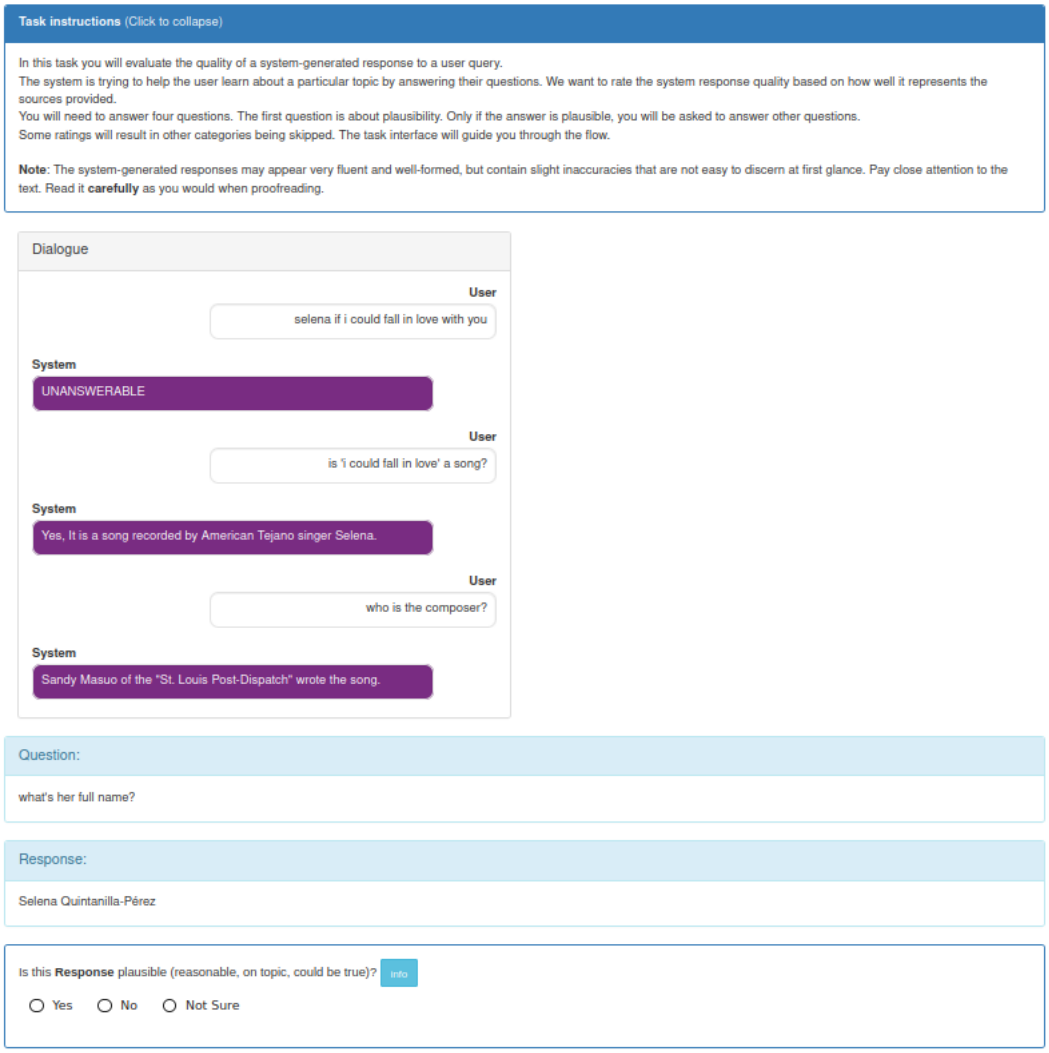

Figure 2: Interface used to collect faithfulness. The annotator is asked to answer the question about plausibility of the response first, without looking at the document. The annotation stops at this point if the response is not plausible.

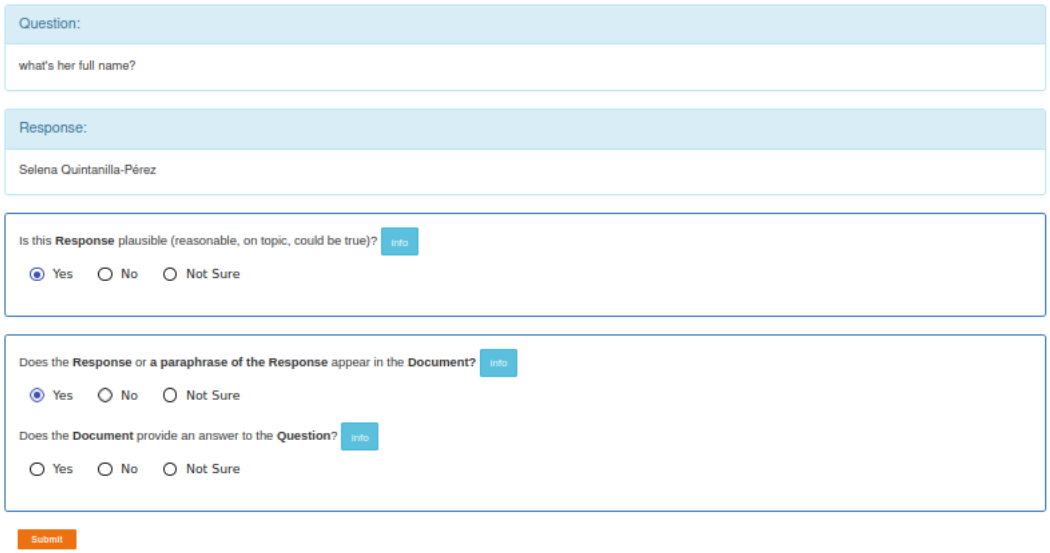

Figure 3: Interface used to collect faithfulness. The annotator has now access to the document and can annotate.

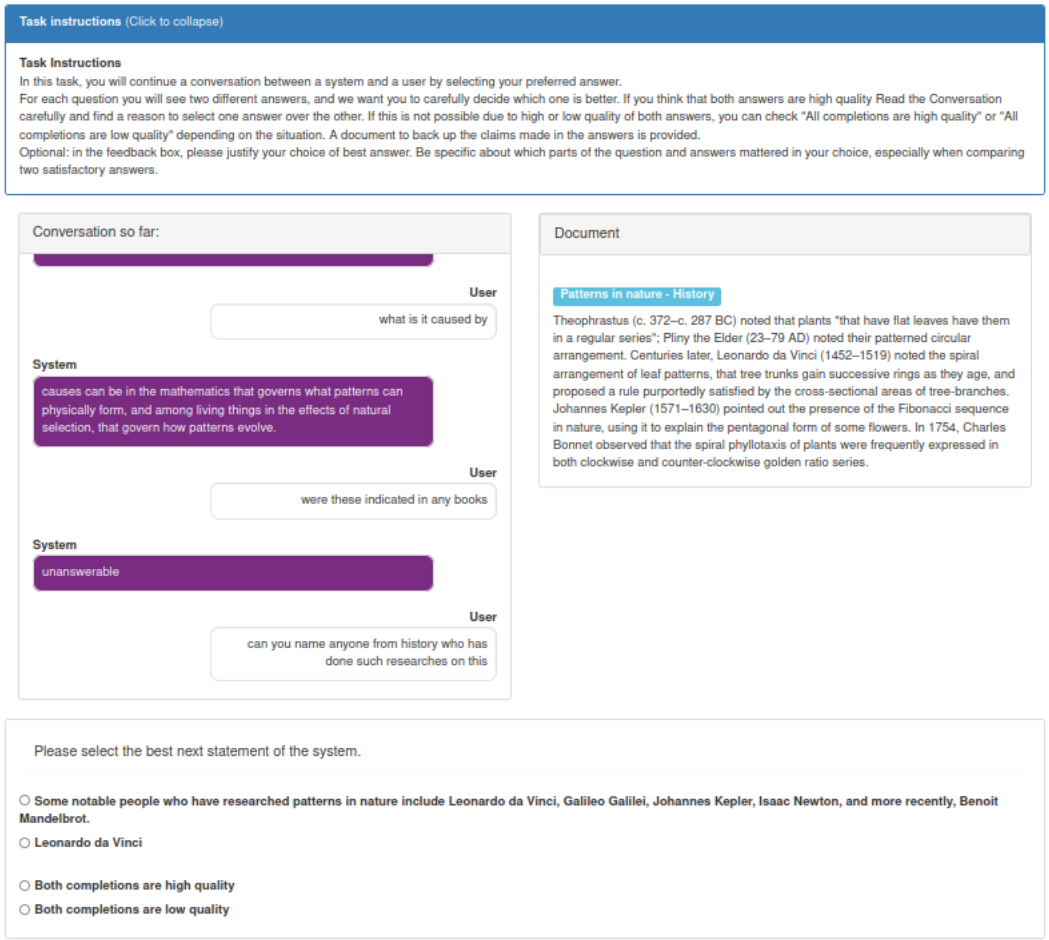

Figure 4: Interface used to collect the human evaluation for preferences

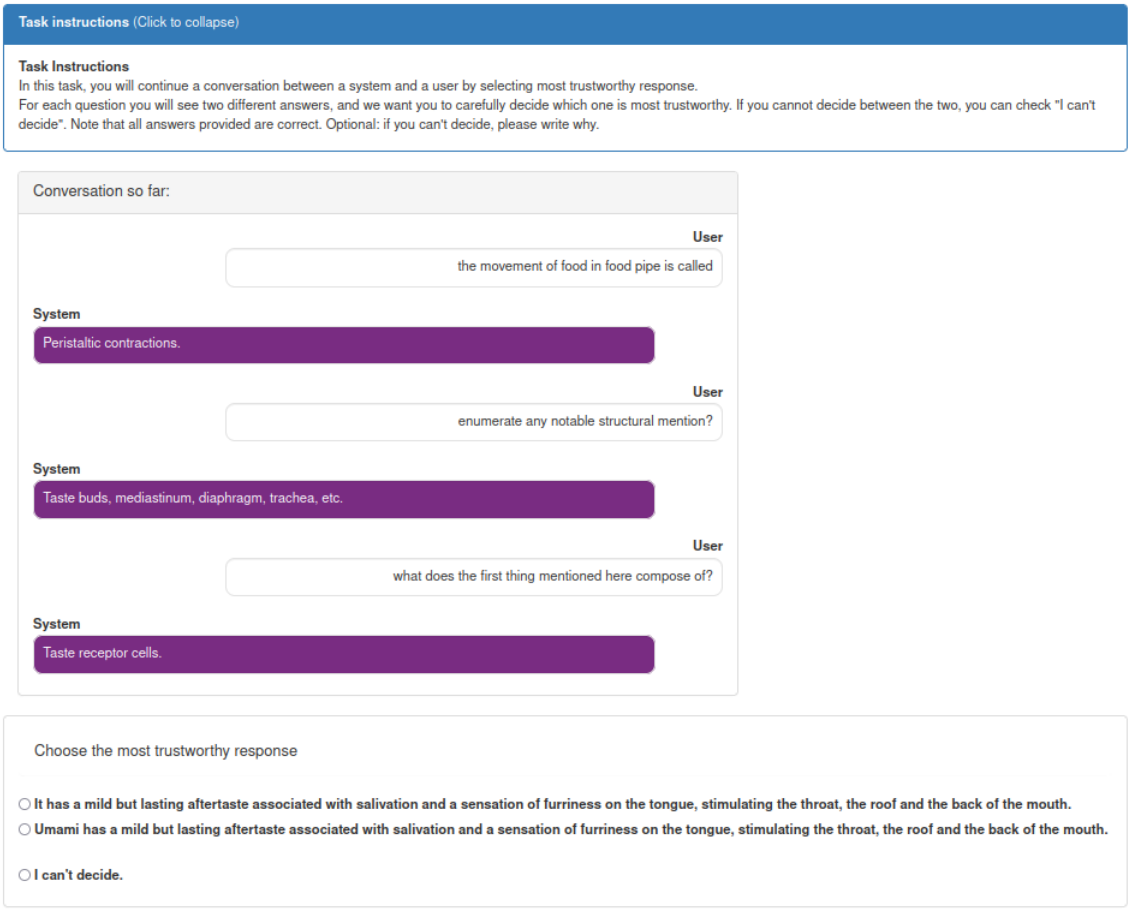

Figure 5: Interface used to collect the human evaluation for the study of trust

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
6

☑ A2. Did you discuss any potential risks of your work?
7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

2

☑ B1. Did you cite the creators of artifacts you used?
2

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The dataset we used is in the public domain

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
6

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
6

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
7

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
2

C ☑ Did you run computational experiments?

3

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We only run evaluation experiments that have a low running overhead/footprint.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
3

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
3,4

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
6

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Approved by the internal committee of our company

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
6