On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models?

Anonymous ACL submission

Abstract

Knowledge-grounded conversational models are known to suffer from producing factually invalid statements, a phenomenon commonly called hallucination. In this work, we investigate the underlying causes of this phenomenon: is hallucination due to the training data, or to the models? We conduct a comprehensive human study on both existing knowledge-grounded conversational benchmarks and several state-of-the-art models. Our study reveals that the standard benchmarks consist of >60% hallucinated responses, leading to models that not only hallucinate but even amplify hallucinations. Our findings raise important questions on the quality of existing datasets and models trained using them. All our annotations will be publicly released.

1 Introduction

Knowledge-grounded conversational models, powered by large pre-trained language models (Radford et al., 2019; Brown et al., 2020; Raffel et al., 2020), are well-known to generate factually incorrect statements, a phenomenon commonly called *hallucination* (Dziri et al., 2021b; Rashkin et al., 2021b). A large commonality in the majority of prior work seeks to address hallucination by ameliorating the model (Shuster et al., 2021; Mielke et al., 2020; Dziri et al., 2021a; Rashkin et al., 2021b), but no attempt has been made so far to audit the conversational benchmarks to the best of our knowledge.

On one hand, knowledge-grounded conversational benchmarks may contain hallucinations due to error-prone collection protocols, or due to a design framework that encourages informativeness over faithfulness. Existing dialogue systems are typically trained on corpora crowd-sourced through online platforms (Dinan et al., 2018; Gopalakrishnan et al., 2019; Moon et al., 2019). With loose incentive to come up with faithfully-grounded utterances on the provided knowledge, crowdworkers may ignore knowledge-snippets altogether, use

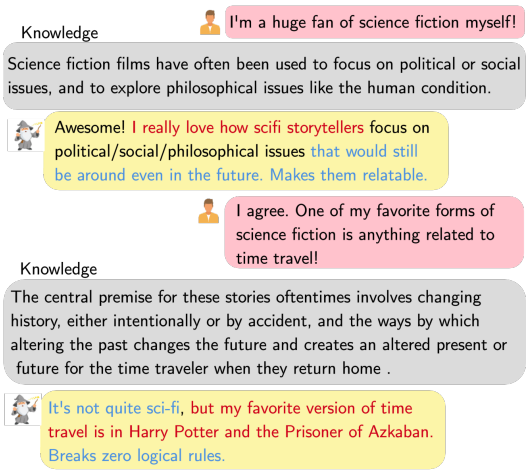


Figure 1: An example of a hallucinated conversation from the Wizard of Wikipedia dataset (Dinan et al., 2018). The wizard (yellow) is hallucinating information that cannot be inferred from the knowledge-snippet: hallucinated subjective content (red) and hallucinated objective content (blue).

their personal knowledge or sometimes assume a fictional persona, resulting in conversations that are rife with subjective content and unverified factual knowledge. Figure 1 shows a hallucinated conversation from the WoW dataset (Dinan et al., 2018),

On the other hand, neural conversational models are not necessarily designed to generate faithful outputs, but to mimic the distributional properties of the data. This kind of optimization will likely push the models to replicate and even amplify the hallucination behaviour at test time (Bender et al., 2021). The presence of even few hallucinated responses may skew the data distribution in a way that curbs the model's ability to generate faithful responses (Kang and Hashimoto, 2020).

In this work, drawing insights from the linguistic coding system for discourse phenomena (Stiles, 1992) and evaluation frameworks such as BEGIN (Dziri et al., 2021b) and AIS (Rashkin et al., 2021a), we annotate responses from the three widely-used knowledge-grounded conversational benchmarks: Wizard of Wikipedia (Dinan et al., 2018), CMU-DoG (Zhou et al., 2018) and Topi-

calChat (Gopalakrishnan et al., 2019). Our analysis reveals surprisingly that more than 60% of the responses are hallucinated in the three datasets, with major hallucination modes that manifest principally through the expression of subjective information (e.g., thoughts, beliefs, feelings, intentions, personal experiences) and the expression of unsupported objective factual information. Further, to understand if neural conversational models make this hallucination more severe, we annotate responses generated by several state-of-the-art models, including ones that are designed to alleviate hallucinations. We find that the generated responses consist of an even larger portion of hallucinations, in comparison with the training data. Our findings question the quality of current conversational datasets, their appropriateness to train knowledge-grounded conversational systems, and the robustness of existing models.

2 Hallucinations in Benchmarks

We conduct a human study on three English crowdsourced knowledge-grounded conversational benchmarks: Wizard of Wikipedia (WoW), CMU-DoG and TopicalChat. These datasets consist of dialogues between two speakers, where the goal is to communicate information about particular topics while speakers are presented with a knowledge snippet relevant to the current turn. More details about these datasets are provided in §A.

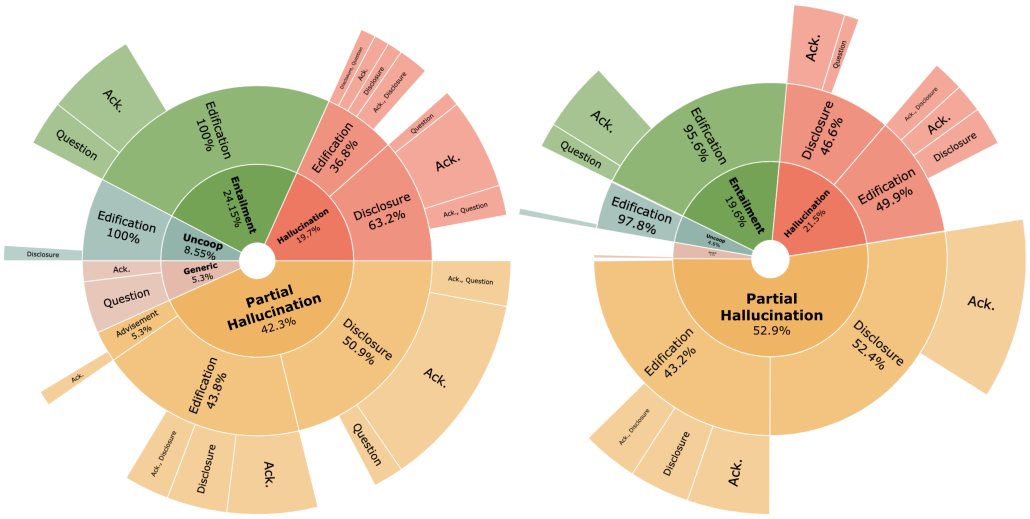
Response Classification Taxonomy Following the definitions of the BEGIN taxonomy (Dziri et al., 2021b) of response classification and the AIS framework (Rashkin et al., 2021a) of evaluating response attribution, we annotate each response based on whether it can be inferred exclusively from the knowledge-snippet as follows:[1] **Entailment**: a response is fully supported by the knowledge. **Hallucination**: a response's factual correctness cannot be fully verified from the knowledge-snippet (even if it is true in the real world). **Partial Hallucination**: part of the response is hallucinated while the rest is entailed. **Generic**: a response that is vague and does not convey any factual information. **Uncooperative**: an entailed response that does not follow the principles of conversational cooperation according to Gricean maxims (Grice, 1989). We provide more details of these classes and examples in §D. To understand the linguistic nature of hallucinations, we further annotate responses based on a linguistic coding system for discourse phenomena, dubbed Verbal Response Modes (VRM; Stiles 1992). Concretely, we label a turn with the following speech acts: **Disclosure**, **Edification**, **Advisement**, **Confirmation**, **Question** and **Acknowledgement (Ack.)**. Table 1 displays the definition for each VRM type. We opted for the VRM taxonomy as it offers a simple way of codifying responses into categories that are sufficient for our analysis whereas one can also opt for a more demanding annotation scheme (Bunt et al., 2020).

[1]We omit the label contradiction and off-topic from BEGIN as we consider it a subcategory of hallucination.

2.1 Human Evaluation Study

We follow a two-stage annotation protocol where we first ask two linguists to judge the attribution of 200 randomly sampled train responses with respect to the source knowledge. Details about experts can be found in §E. For inter-annotator agreement, we measure Fleiss' Kappa scores on both BEGIN and VRM. WoW achieved 0.89 on BEGIN and 0.78 on VRM, indicating substantial agreement. Annotations on CMU-DoG and TopicalChat achieved nearly similar agreement (See §F). The high agreement scores align with the findings in AIS on WoW (Rashkin et al., 2021a).

The second round corresponds to a large-scale annotation of 4K randomly sampled train responses using non-expert annotators from AMT. This round is crucial to ensure that the obtained results from the experts are reliable enough to draw conclusions about the quality of the data. As human annotation is expensive, we perform the non-expert annotations only on the WoW benchmark while restricting ourselves to expert annotations on CMU-DoG and TopicalChat data. We choose WoW over the other two datasets as the source knowledge is more amenable to faster annotation (TopicalChat: 250 words > CMU-DoG: 200 words > WoW: 25 words). Details about our AMT task design and how we ensure data quality can be found in §G. In total, we selected 4 trusted workers to annotate the 4k responses. To compute the inter-annotator agreement between the workers, we assign three workers per dialogue response in a secondary task, and ask each of them to judge 500 responses. Reported Fleiss' Kappa agreements were 0.75 for BEGIN and 0.61 for VRM indicating high agreement (but lower than the experts which is expected).



(a) Expert annotations (200 responses) (b) Non-expert annotations (4000 responses)

Figure 2: BEGIN and VRM breakdown of responses from WoW. The inner circle shows the breakdown of BEGIN classes and the outer shows the VRM types in each BEGIN type: Hallucination (red), Entailment (green), Partial Hallucination (yellow), Generic (pink), and Uncooperative (blue).

(Q1) How much hallucination exists in the benchmarks? Figure 2 shows the breakdown of each BEGIN categoty in WoW and compares expert annotations versus AMT workers. Surprisingly, WoW is fraught with hallucinations. Expert annotations on 200 responses show that hallucinated responses are largely mixed with faithful content (42.3% v.s. 19.7% fully hallucinated responses), which amounts to 62% hallucinations in total. These results generalize even on larger data; we can see that the portion of hallucinated responses increased to 74.4% when evaluated on 4K samples. Our analysis shows similar trends on the CMU-DoG and TopicalChat benchmarks. TopicalChat contains 62% responses that are purely hallucinated against only 12% responses that are fully entailing the source knowledge and CMU-DoG has similar results. Detailed results are presented in §J and exemplars of hallucinated responses are depicted in §K. These findings raise the question on the quality of dialogue datasets.

(Q2) What are the hallucination strategies used in human-human data? Figure 2 and Figure 5 show the VRM breakdown for each BEGIN category in the three benchmarks. We make the following observations: The majority of hallucinations belong to *disclosure* (i.e., subjective information) in all benchmarks (63.2%, 56.2% and 61.5% in WoW, CMU_DoG and TopicalChat respectively). Although the strategy of sharing subjective information such as thoughts, opinions and feelings is natural in conversations, it often comes at a cost of ignoring the knowledge snippet in these datasets. Moreover, *edification* is also a common phenomenon in hallucinated responses, suggesting that humans not only discuss subjective information but also bring extra unsupported facts, either true or false. Other linguistic modes are also associated with hallucinations such as acknowledging unsupported claims or asking irrelevant questions. Conversely, entailment responses have high percentage of edification ($> 70\%$) with information inferred from the knowledge snippet.

VRM Type	Description
Disclosure	Reveal the speaker's subjective opinions, personal experience, thoughts and feelings.
Edification	Concerns information that is objective.
Advisement	Corresponds to guiding the behaviour of the addressee through: commands, requests, suggestions, advice, permission, prohibition.
Confirmation	Compares the speaker's experience with the other's by expressing shared ideas or by agreement, disagreement.
Question	Concerns requesting information or guidance.
Acknowledge	Expresses no content, it conveys only receipt of communication from the other's speaker.

Table 1: The definitions of the VRM types with examples.

3 Hallucination Amplification in Models

Next, we investigate how much models amplify the hallucination phenomenon at inference time. We consider a range of representative models:

- **GPT2** (Radford et al., 2019; Wolf et al., 2019)

	Model	ROUGE↑	Hallucination Rate↓			Entailment Rate↑		
			Full	Partial	Overall	Entail.	Uncoop.	Overall
WoW	Gold	–	19.7	42.3	62.0	24.1	8.5	32.7
	GPT2	19.1	66.0	15.2	81.2	11.7	3.6	15.3
	DoHA	21.5	39.6	28.9	68.5	12.7	7.1	19.8
	CTRL	24.4	29.0	34.0	63.0	12.2	17.3	29.5
CMU DoG	Gold	–	61.4	5.1	66.5	16.2	4.1	20.3
	GPT2	13.7	75.0	6.0	81.0	5.0	5.5	10.5
	DoHA	15.4	62.5	10.0	72.5	8.1	5.9	14.0
	CTRL	19.3	62.3	6.0	68.3	9.1	12.4	19.8
Topical	Gold	–	46.8	17.1	63.9	22.9	0.5	23.4
	GPT2	13.2	71.5	8.5	80.5	6.0	5.5	11.5
	DoHA	16.3	52.5	26.5	79.0	9.0	5.0	14.0
	CTRL	18.3	47.5	29.5	77.0	8.5	10.0	18.5

Table 2: Amplification of models on the test data from WoW and CMU_DoG and TopicalChat. 'Entail.' and 'Uncoop.' mean entailment and uncooperative. Results are statistically significant with p-value< 0.01 according to McNemar's test.

is an autoregressive model which takes as input a concatenation of the knowledge and the history.

• **DoHA** (Prabhumoye et al., 2021) builds a BART-based conversational model (Lewis et al., 2020) for knowledge-grounding, with a two-view attention mechanism to handle separately the encoded document and the history during generation.

• **CTRL** (Rashkin et al., 2021b) augments the GPT2 model with control tokens (Keskar et al., 2019) that guide the generation towards less subjective and more entailed content.

We fine-tune each model on the benchmarks and use nucleus sampling (Holtzman et al., 2019) with $p = 0.6$ for decoding (more implementation details are in §B). As seen in Table 2, CTRL is the best model followed by DoHA based on the ROUGE score. Table 6 in §M shows a sample of generated responses. Similar to the analysis in §2, we task the same two linguists to analyze model-generated responses for 200 randomly-selected test samples from each benchmark. We seek to answer the following questions:

(Q3) Do state-of-the-art conversational models amplify hallucination? Table 2 shows the degree of amplification across different models trained on the three benchmarks. Numbers report the percentage of each class in the data. Contrasting this with human gold responses, the models not only hallucinate but also amplify the percentage of hallucinations. For example, GPT2 amplifies full hallucination by 19.2% in WoW, 14.5% in CMU-DoG and 16.6% in TopicalChat. Conversely, it reduces entailment by 17.4%, 9.8% and 11.9% respectively. This suggests that hallucination patterns are easier to learn than entailment. Among the three, CTRL hallucinates less followed by DoHA. Overall, these results demonstrate that hallucination is not only a reflection of training data issues, but also a consequence of the weaknesses of models.

(Q4) What are the hallucination strategies used by models? Surprisingly, different models use different strategies for hallucination. While DoHA and GPT2 predominantly rely on and amplify *disclosure*, CTRL relies on *edification*. This is because CTRL is trained explicitly to avoid pronouns (a crucial ingredient for disclosure) and to generate entailed responses. As a side-effect, it ends up amplifying uncooperative responses (by 100%, 202% in WoW and CMU-DoG as seen in Table 2). Full results of all models and datasets are in §L.

4 Related Work

Hallucination in neural language generation has recently received increasing attention from the NLP community, including machine translation (Raunak et al., 2021; Wang and Sennrich, 2020) and summarization (Durmus et al., 2020; Kang and Hashimoto, 2020). Hallucinations in knowledge-grounded neural dialogue generation is a nascent research problem (Roller et al., 2021; Mielke et al., 2020; Shuster et al., 2021; Dziri et al., 2021a; Rashkin et al., 2021b). Closest to our work are Dziri et al. (2021b) and Santhanam et al. (2021) who introduce testbeds for quantifying groundedness in dialogue systems including hallucinations, whereas we conduct a much finer-grained manual analysis on multiple benchmarks and models.

5 Conclusion

Our investigations demonstrate empirically that hallucination is a prevalent issue in both dialog benchmarks and models. Our analysis on three widely used benchmarks reveals that they are rife with hallucinations (more than half of the data), and the most common strategies people use are *disclosure* and *edification*. Moreover, we show that conversational models trained on these benchmarks not only hallucinate but also amplify hallucinations, even the models that were designed to alleviate this issue. This calls for a clean high-quality data release and careful design of trustworthy conversational systems. Before then, we strongly advocate practitioners to look at samples of any dataset—in order to uncover actionable insights—prior to their use or public release.

Impact Statement & Ethics

Risks in annotation Before starting the annotation, this study was thoroughly reviewed and approved internally by a review board. The benchmarks we audit were collected through AMT and thus may contain some non-normative or even profane and racist examples. Annotators were also asked to judge the outputs of several state-of-the-art conversational systems which may be in turn toxic and insensitive. We recognize the emotional burden that this presents to annotators (Roberts, 2016). Therefore, we added the following content warning in italic text in the header of each task: *This HIT may contain text that disturbs some workers. If at any point you do not feel comfortable, please feel free to skip the HIT.*

Risks in deployment Our analytical study reveals that a large portion of standard knowledge-grounded dialogue benchmarks is hallucinated, leading us to reflect on the potential harm of low-quality data releases for conversational models. In recent years, the conversational AI market has seen a proliferation of a variety of applications—which are powered by large pre-trained LMs—that span across a broad range of domains, such as customer support, education, e-commerce, health, entertainment, etc (Vakulenko et al., 2021). Ensuring that these systems are trustworthy is key to deploy systems safely at a large scale in real-world application, especially in high-stake domains (Sambasivan et al., 2021). However, even if we come up with a model that is robust enough against hallucination, it will be ultimately bounded by the data quality. We argue that fixing the models or the data to enforce faithfulness is a highly non-trivial task without an in-depth understanding of the various sources of hallucination. Our work thus represents the first effort to gain such an understanding and to inform the community about the unreliability of the existing benchmarks and models. As result, we believe it is important to raise these insights to the broader community.

References

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

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.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. 2020. The ISO standard for dialogue act annotation, second edition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 549–558, Marseille, France. European Language Resources Association.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021a. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2021b. Evaluating groundedness in dialogue systems: The begin benchmark. *arXiv preprint arXiv:2105.00071*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Paul Grice. 1989. *Studies in the Way of Words*. Harvard University Press.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Daniel Kang and Tatsunori B. Hashimoto. 2020. Improved natural language generation via loss truncation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 718–731, Online. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Sabrina J Mielke, Arthur Szlam, Y-Lan Boureau, and Emily Dinan. 2020. Linguistic calibration through metacognition: aligning dialogue agent responses with expected correctness. *arXiv preprint arXiv:2012.14983*.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

Shrimai Prabhumoye, Kazuma Hashimoto, Yingbo Zhou, Alan W Black, and Ruslan Salakhutdinov. 2021. Focused attention improves document-grounded generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4274–4287, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, 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.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Sarah T Roberts. 2016. Commercial content moderation: Digital laborers' dirty work, book chapter published in the intersectional internet: Race, sex, class and culture online, edited by s. u. noble b. tynes, and published by peter lang publishing.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. 2021. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. *arXiv preprint arXiv:2110.05456*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

William B Stiles. 1992. *Describing talk: A taxonomy of verbal response modes*. Sage Publications.

Svitlana Vakulenko, Evangelos Kanoulas, and Maarten de Rijke. 2021. A large-scale analysis of mixed initiative in information-seeking dialogues for conversational search. *arXiv preprint arXiv:2104.07096*.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

A Datasets

We conduct our analysis on the following datasets:

Wizard of Wikipedia: consists of dialogues between a "wizard" and an "apprentice", where the goal of the wizard is to communicate information about a particular topic. The Apprentice, in turn, is expected to seek information about the topic. At each turn, the wizard is presented with a knowledge snippet (typically a sentence) from Wikipedia while the apprentice is not; and the Wizard is allowed to form an utterance that does not use the evidence. We omit data points in which the wizard did not explicitly select a passage as evidence for the response. The dataset consists of 82722 grounded-responses in train, 8800 valid and 8690 test.

CMU-DoG: consists of conversations focusing only on the movie domain. Each response is grounded on a document (typically a section from Wikipedia). Workers are asked to either persuade the other speakers to watch the movie based on the knowledge or to discuss the content of the document with them. In total, there are 78136 grounded responses in train, 13800 in valid and 13796 in test.

TopicalChat: consists of dialogues conversing around a variety of topics. Workers are provided with relevant facts from Reddit, Wikipedia and news articles. The collection protocol consists of two scenarios: symmetric and asymmetric. In the symmetric scenario, workers will have access to the same source knowledge and in the asymmetric scenario, they will have access to different sources. In total, the dataset has 292215 grounded responses in train, 23601 in valid and 23623 in test.

B Implementation Details

GPT2: We implement this model using the Pytorch Huggingface Transformers library (Wolf et al., 2020) and the Pytorch-lightning library[2]. During training, we use the Adam optimizer (Kingma and Ba, 2015) with Dropout (Srivastava et al., 2014) on a batch size of 32 with a learning rate of 6.25×10^{-5} that is linearly decayed. The maximum dialogue history length is set to 3 utterances. The model early-stops at epoch {6, 10, 10} respectively for WoW, CMU-DoG and TopicalChat. The average runtime is {1.5, 3, 3} hours for WoW, CMU-DoG and TopicalChat respectively.

DoHA: We use the code and the pre-trained model on CMU-DoG that are publicly available by the authors at their Github's account [3]. For WoW and TopicalChat, we follow closely the authors' training procedure described in (Prabhumoye et al., 2021) and we trained two models on both datasets. The average runtime of these models is {5, 10} hours for WoW and TopicalChat respectively.

CTRL: The code is not publicly available for this model. We were able to reproduce the results ourselves by following training details in the paper and having multiple discussions with the authors. Similar to GPT2, we implement this model using the Pytorch Huggingface Transformers library and the Pytorch-lightning library.

For each dataset, we save the best model based on the validation set. Training for all models is done on an Nvidia V100 GPU 32GB and for inference, we use nucleus sampling with p=0.6. To compute the ROUGE score, we used 'rouge_scorer' from https://github.com/google-research/google-research/tree/master/rouge.

C Definition of VRM

Table 3 contains VRM definitions with examples.

D Definitions of the BEGIN Taxonomy

Here are more detailed definitions of each of the BEGIN classes. Examples can be found in Table 5:

- **Entailment**: The response can be verified as true based solely on the knowledge-snippet. Any factual information which it contains can be found in or derived from the knowledge provided.
- **Hallucination**: The response cannot be verified as true based solely on the knowledge-snippet. It is comprised of information which cannot be found in or derived from the knowledge provided. More specifically, responses about personal opinions, experiences, feelings, internal assessments of reality that cannot be attributed to the information present in the source document, are considered hallucination.

[2] https://github.com/PyTorchLightning/pytorch-lightning

[3] https://bit.ly/3bBup2M

VRM Type	Description	Example
Disclosure	Reveal the speaker's subjective opinions, personal experience, thoughts, feelings, wishes, and intentions.	*"I think science fiction is an amazing genre. Future science, technology they're all interesting."*
Edification	Concerns information that is, in principle, objective.	*"Recycling includes items like metal and plastic."*
Advisement	Corresponds to guiding the behaviour of the addressee through: commands, requests, suggestions, advice, permission, prohibition.	*"You should be patient and persistent to succeed."*
Confirmation	Compares the speaker's experience with the other's by expressing shared ideas/memories/beliefs, or by agreement/disagreement	*"I agree that love encompasses a variety of different emotional and mental states."*
Question	Concerns requesting information or guidance.	*"What is your favorite song?"*
Acknowledge	Expresses no content, it conveys only receipt of communication from the other's speaker.	*"Mmm. OK,..."*, *"Yeah, ..."*, *"Hello, ..."*

Table 3: The definitions of the VRM types with examples.

	BEGIN	VRM
CMU_DoG	0.85	0.78
TopicalChat	0.83	0.72

Table 4: Fleiss Kappa Scores on 200 train Human-Human responses taken from the CMU-DoG and TopicalChat benchmarks.

- **Partial Hallucination**: Part of the response can be verified as true using the knowledge-snippet (i.e. is entailed), while another part of the response cannot (i.e. is hallucinated).
- **Generic**: The response is vague and does not contain factual claims. Such responses are often used simply to convey receptiveness to what the interlocutor has to say.
- **Uncooperative**: The response is entailed but violates the Gricean maxims of conversation (Grice, 1989). These responses may be perceived as rude, purposefully misleading, or showing a general unwillingness to cooperate with the interlocutor for effective communication.

E Expert Annotation

The two experts were students with linguistics background, fluent in English, and were trained for the task by exchanging rigorous discussions with the authors. As part of this stage, they were required to write justifications for 50 samples elaborating the reasoning for the provided ratings. The collected justifications were helpful in understanding the reasoning experts used to reach their ratings and in laying the groundwork for designing the second round of annotations.

F Inter-annotator Agreement on Gold Responses

Table 4 contains the Fleiss kappa scores for CMU_DoG and TopicalChat.

G AMT Human Annotation

Task Design To streamline the process for raters we break down the task into hierarchical (yes/no) questions. We summarize this procedure below, and provide the exact questions in §H. First, we ask annotators to judge whether the response contain information that is not supported by the source. If yes, we ask them to indicate the type of the unsupported information (e.g., unsupported opinion, unsupported fact, etc). In a followup question, we ask them to indicate whether there are any supported information besides the hallucinated content. If the response was not hallucinated, we present them with two follow-up questions about whether the response is entailing the source or generic. Finally, if the response entails the source, we ask whether it is cooperative with respect to the previous utterance in the context.

AMT Data Quality To access the preliminary staging round in AMT, workers had to be located in the United States and Canada and pass a qualifying test, correctly answering 14 questions about BEGIN and VRM. Once they are granted access to the task, we use a process of manual quality control where we monitor worker performance in two stages. First, workers would have access only to a small staging round (batch size $\sim$ 30 HITs). In this round, one of the authors acts as an inspector who would meticulously check each of the annotators submissions for compliance with the instructions. For any observed errors, the inspector would pro-

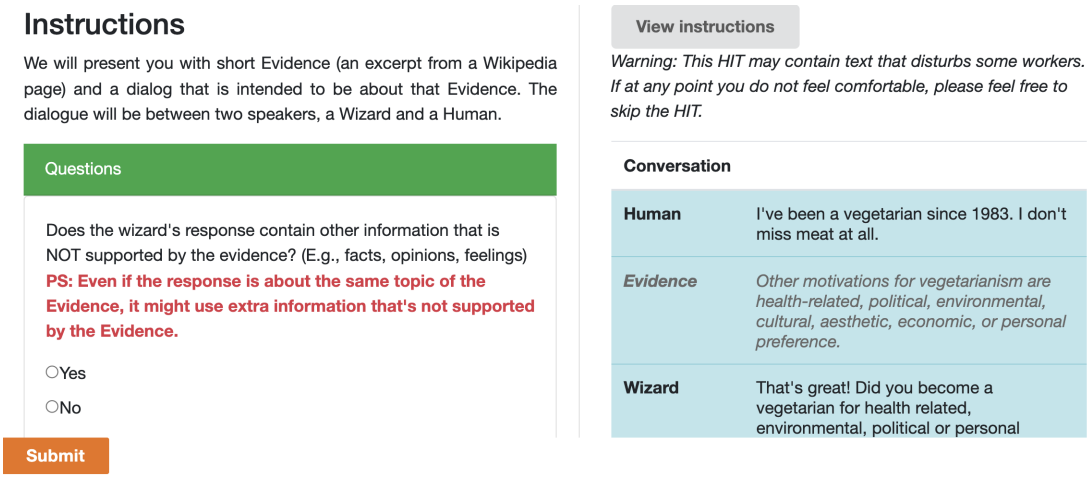


Figure 3: AMT Annotation interfaces for determining BEGIN and VRM classes (1).

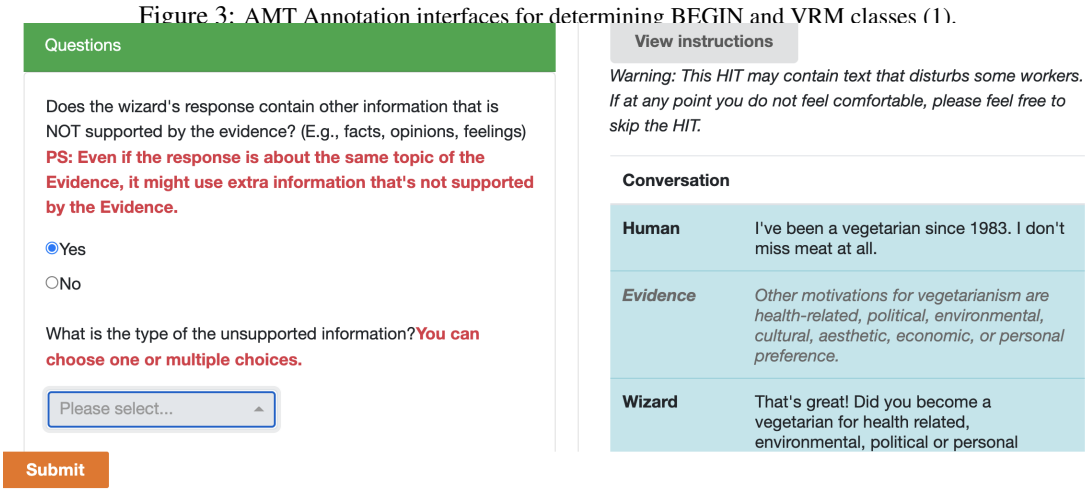


Figure 4: AMT Annotation interfaces for determining BEGIN and VRM classes (2).

vide direct feedback to the worker, explaining any misunderstandings and encouraging the worker to engage in an open discussion concerning these misunderstandings via email. At the end of this round, we revoke access for workers who provide poor quality annotations. Next, we launch the main annotation stage which is larger (batch size $\sim$ 400 HITs) and more efficient. We perform daily manual inspection and we send detailed feedback to workers who commit persistent error patterns. We reject poor quality work in this stage and repeated rejections result in the worker being blocked from the task entirely. In total, we ended up with 4 workers annotating the 4k responses. The workers were informed that their annotations would be used for research purposes and their workers ID would be anonymous in case we decide to release the data. To compute the inter-annotator agreement between the workers, we assign three workers per dialogue response in a secondary task, and ask each of them to judge 500 responses. Reported Fleiss' Kappa agreements were 0.75 for BEGIN and 0.61 for VRM. Although substantial, the agreement is lower than the experts' one and this is expected as they have stronger linguistic background and they pay careful attention to the task.

H AMT Human Instructions

AMT Human annotation interfaces are depicted in Figure 3 and Figure 4. We pay workers an hourly wage around 18-20 USD which is above the minimum wage rate. Workers were asked the following questions:

1. Does the Wizard's response contain other information that is NOT supported by the evidence? (E.g., facts, opinions, feelings)?
   (a) If the response is hallucinated, what is the type of the unsupported information? (expressing a personal experience, expressing an opinion, expressing feelings, expressing unsupported facts, giving advice, acknowledging with information from the human)

BEGIN	VRM	Example	
Entailment	Disclosure	Knowledge:	A dragon is a legendary creature, typically scaled or fire-spewing and with serpentine, reptilian or avian traits, that features in the myths of many cultures around world.
		History:	Dragons are so fascinating, I wonder where they originated from.
		Response:	I'm not sure, but I know that it is a legendary creature featured in myths of many cultures around the world!
Hallucination	Edification Disclosure	Knowledge:	The central premise for these stories oftentimes involves changing history, either intentionally or by accident, and the ways by which altering the past changes the future and creates an altered present or future for the time traveler when they return home.
		History:	One of my favorite forms of science fiction is anything related to time travel! I find it fascinating.
		Response:	It's not quite sci-fi, but my favorite version of time travel is in Harry Potter and the Prisoner of Azkaban. Breaks zero logical rules.
Partial hallucination	Confirmation	Knowledge:	Love encompasses a variety of different emotional and mental states, typically strongly and positively experienced, ranging from the deepest interpersonal affection to the simplest pleasure.
		History:	Romance can be small acts, like making breakfast for your significant other. But it's portrayed as grand gestures which are unattainable.
		Response:	I agree it has been portrayed as impossible actions . Love consists of a variety of emotional and mental states.
Generic	Question	Knowledge:	Hard rock is a loosely defined subgenre of rock music that began in the mid-1960s, with the garage, psychedelic and blues rock movements.
		History:	yea i have heard of the American hard rock band Van Halen.
		Response:	Are you a fan of hard rock?.
Uncooperative	Edification	Knowledge:	Marketing is used to create, keep and satisfy the customer.
		History:	Well, I wish I could. I think my training was a little off. I can't find a job and thinking about going back to school to be a counselor. What do you do?
		Response:	Marketing is used to create a customer and then retain the customer and keep them happy.

Table 5: Examples from Wizard of Wikipedia (Dinan et al., 2018) showing the BEGIN breakdown and different VRM linguistic phenomena for each response.

(b) Besides unsupported information, does the Wizard's response contain thoughts/opinions/feelings/facts that are supported by the Evidence?

2. If the response is not hallucinated, is it faithful to the source or generic? (Faithful, Generic)

3. If the response if faithful, is it cooperative with the Human's response?

I Limitation

The main goal of our paper is to present a data quality audit by gaining an in-depth understanding of the various types of hallucination in both gold and machine-generated responses. We do not investigate the root causes of hallucination in the models. Also, we limit our analysis to only English Benchmarks. Future studies can extend our work to explore the main causes of hallucination in the models and study the problem of hallucination in multilingual datasets.

J Hallucination in CMU-DoG and TopicalChat

Figure 5 shows the hallucination breakdown in CMU-DoG and TopicalChat benchamrks.

K Hallucinated Human-Human Responses

Table 7 contains hallucinated gold responses from WoW, CMU_DoG and TopicalChat.

L Breakdown of BEGIN and VRM in Machine-generated Responses

Figure 6, 7 and 8 display the distribution of BEGIN and VRM in GPT2, DoHA and CTRL trained on the three benchmark.

M Machine-generated Responses

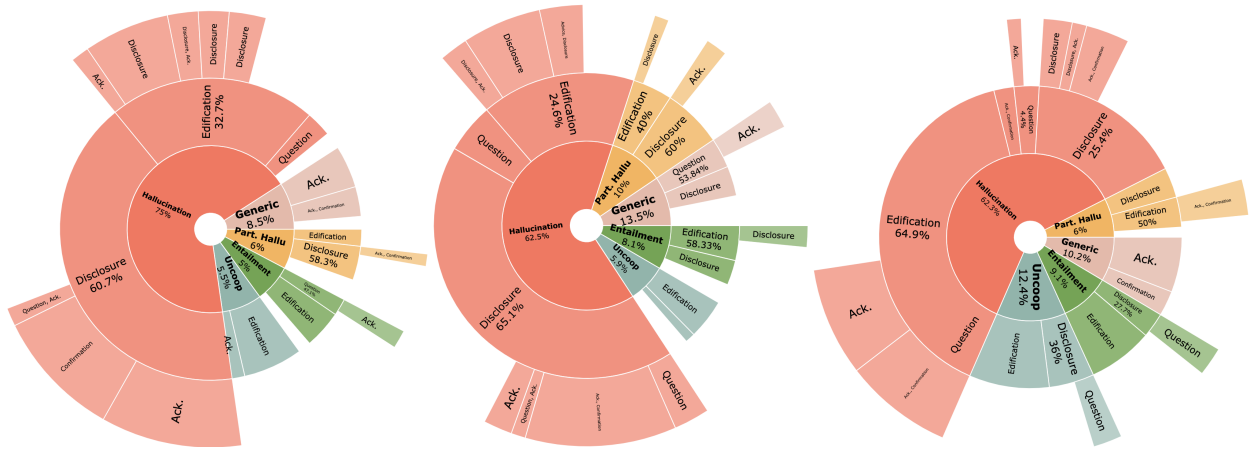
Table 6 contains a sample of generated responses from GPT2, DoHA and CTRL on the WoW and CMU-DoG.

(a) CMU-DoG responses (b) Topical responses

Figure 5: BEGIN and VRM breakdown of responses from CMU-DoG and TopicalChat. The inner circle shows the breakdown of BEGIN classes and the outer shows the VRM types in each BEGIN type: Hallucination (red), Entailment (green), Partial Hallucination (yellow), Generic (pink), and Uncooperative (blue).

(a) GPT2 responses (b) DoHA responses (c) CTRL responses

Figure 6: Breakdown of BEGIN classes and VRM speech acts on WoW machine-generated responses.

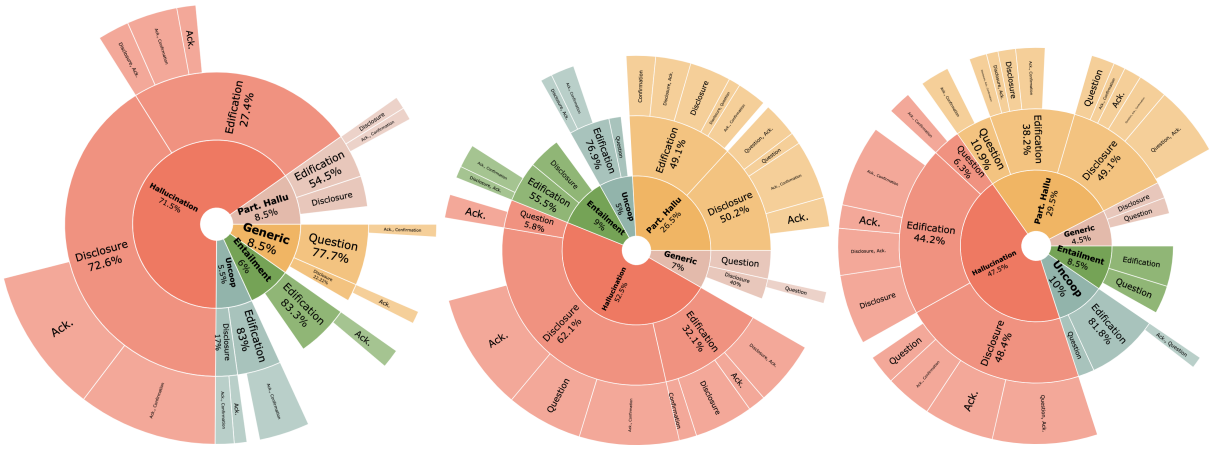


(a) GPT2 responses (b) DoHA responses (c) CTRL responses

Figure 7: Breakdown of BEGIN classes and VRM speech acts on CMU_DoG machine-generated responses.

Data	Knowledge & History	Responses
WoW	**Knowledge**: Cinematography (also called "Direction of Photography") is the science or art of motion-picture photography by recording light or other electromagnetic radiation, either electronically by means of an image sensor, or chemically by means of a light-sensitive material such as film stock. **History**: Hi buddy, What you think about cinematography?	**GPT2**: Cinematography is another title I'm currently in school for. I've been doing the art of photography for over 15 years now. (Hallucination) **DoHA**: I do enjoy the science and art of motion-picture photography. (Partial Hallucination) **CTRL**: You're right. Cinematography is an art of motion -picture photography by recording light or other electromagnetic radiation, either electronically by means of an image sensor, or chemically by means of a light-sensitive material such as film stock. What about you ? (Entailment)
CMU-DoG	**Knowledge**: Mean Girls is a 2004 American teen comedy film directed by Mark Waters and written by Tina Fey. The film is partially based on Rosalind Wiseman's 2002 non-fiction self-help book Queen Bees and Wannabes, which describes female high school social cliques and the damaging effects they can have on girls. Genre: Comedy. MovieName: Mean Girls rating: Rotten Tomatoes: 84 % and average: 6.9/10. **History**: I think Rachel McAdams had an even better role as Regina George however!	**GPT2**: Mean Girls was a classic. (Hallucination) **DoHA**: Yes I agree! she was great and she was better than Mark Waters! (Hallucination) **CTRL**: I see. Mean Girls is a comedy movie that was directed by Mark Waters in 2004. (Entailment)

Table 6: Generated responses from different models based on Wizard of Wikipedia (Dinan et al., 2018) and CMU-DoG (Zhou et al., 2018) test samples.



(a) GPT2 responses (b) DoHA responses (c) CTRL responses

Figure 8: Breakdown of BEGIN classes and VRM speech acts on Topical machine-generated responses.

Data	Knowledge & History	Hallucinated Gold Response
WoW	**Knowledge**: Science fiction often explores the potential consequences of scientific and other innovations, and has been called a 'literature of ideas'. **History**: And that's difficult to do when dealing with time travel . I actually haven't seen the latest Harry Potter movies. Guess it's time to check them out!	If you really want a look at the potential negative consequences of scientific innovation, what you should check out is the TV show Fringe. Incredibly well written.
CMU-DoG	**Knowledge**: Movie: The Social Network. In October 2003, 19-year-old Harvard University student Mark Zuckerberg is dumped by his girlfriend Erica Albright. Returning to his dorm, Zuckerberg writes an insulting entry about Albright on his LiveJournal blog and then creates a campus website called Facemash by hacking into college databases to steal photos of female students, then allowing site visitors to rate their attractiveness. After traffic to the site crashes parts of Harvard's computer network, Zuckerberg is given six months of academic probation. However, Facemash's popularity attracts the attention of Harvard upperclassmen and twins Cameron and Tyler Winklevoss and their business partner Divya Narendra. The trio invites Zuckerberg to work on Harvard Connection, a social network featuring the exclusive nature of Harvard students and aimed at dating. **History**: The movie is The Social Network. I personally do not like Facebook as a company.	The movie portrays the founding of social networking website Facebook and the resulting lawsuits. It even has Justin Timberlake in it, I don't think I've ever seen him act.
TopicalChat	**Knowledge**: Wikipedia: first paragraph in https://en.wikipedia.org/wiki/Google Reddit facts: A single Google search requires more computing power than it took to send Neil Armstrong and eleven other astronauts to the moon. Google Maps calculates traffic by tracking how fast Android devices are moving on the road instead of hiring people to mow the lawns around their headquarters. Google uses hundreds of live goats. On 16th August 2013, Google went down for about five minutes, and took 40% of web traffic with it. When there is a disputed border, Google maps tailors its maps to the claims of each country where the Internet browser is located. **History**: Google provides online related services and products, which includes online ads, search engine and cloud computing.	Yeah, their services are good. I 'm just not a fan of intrusive they can be on our personal lives.

Table 7: Hallucinated responses from different benchmarks: Wikipedia (Dinan et al., 2018), CMU_DoG (Zhou et al., 2018) and TopicalChat (Gopalakrishnan et al., 2019). Text highlighted in red indicates hallucinated content.