Does the Generator Mind its Contexts?
An Analysis of Generative Model Faithfulness under Knowledge Transfer

Anonymous ACL submission

Abstract

The knowledge-augmented generator should generate information grounded on input contextual knowledge despite how the context changes. Many previous works focus on hallucination analysis from static input (e.g., in summarization or machine translation). In this work, we probe faithfulness in generative question answering with dynamic knowledge. We explore whether hallucination from parametric memory exists when contextual knowledge changes and analyze why it happens. For efficiency, we propose a simple and effective measure for such hallucinations. Surprisingly, our investigation reveals that all models only hallucinate previous answers in rare cases. To further analyze the causality of this issue, we conduct experiments and verify that context is a critical factor in hallucination during training and testing from several perspectives.

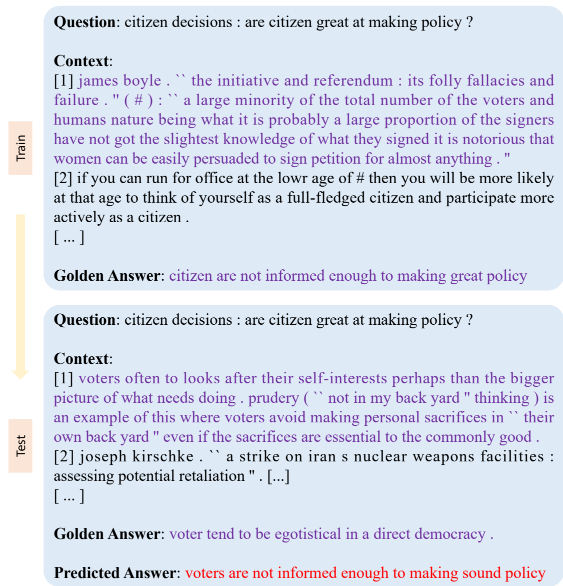


Figure 1: An example of generated hallucination from training memory. The model disregards the transferred contextual knowledge and predicts the out-of-date answer in training data.

1 Introduction

Knowledge-augmented text generation, such as RAG (Lewis et al., 2020), FiD (Izacard and Grave, 2021), and Atlas (Izacard et al., 2022), the paradigm of generating text from external knowledge, has achieved state-of-the-art (SOTA) performance in many NLP tasks. Non-parametric contextual knowledge provides the advantage of plug-and-play, while implicit parametric knowledge stored in models needs to be retrained for updating (Li et al., 2022a). A faithful knowledge-augmented generator should always generate consistent output with the grounded context (Ji et al., 2022). However, hallucination is often generated from parametric memory (Figure 1), making it a hurdle for text generation in real-world applications (Maynez et al., 2020; Zhang et al., 2020b).

The faithfulness of generative models under dynamic knowledge is still under exploration. Many previous works focus on hallucination analysis from statistic input, e.g., for summarization(Pagnoni et al., 2021; Ladhak et al., 2022; Tang et al., 2022) or machine translation(Raunak et al., 2021; Müller et al., 2020). Although many works have attracted the attention of dynamic question answering (Min et al., 2020; Longpre et al., 2021; Zhang and Choi, 2021; Chen et al., 2021; Wang et al., 2022; Liska et al., 2022; Kasai et al., 2022), seldom experiments (Longpre et al., 2021; West et al., 2022) systematically statisticize the extent of model faithfulness and analyze when and why models generate hallucinations under dynamic knowledge. We define *knowledge transfer* as contextual knowledge changes under the same question. Specifically, the generative model is trained on old version knowledge but tested on new ones. Like Longpre et al. (2021), we fall in the scope of analyzing *memory hallucinations*, which are generated from parametric knowledge under knowledge transfer.

In this work, we try to measure the model faithfulness under knowledge transfer in two-fold:

RQ 1 *Whether is the generative model faithful under knowledge transfer?*

RQ 2 *Why would the memory hallucination take place?*

We clarify the knowledge transfer task and propose a metric for hallucination measurement (§3). Then we conduct experiments on several models for RQ 1. Our investigation reveals that models are not fully grounded on contexts under knowledge transfer (§4), though it is not as severe as in summarization (Maynez et al., 2020). We conduct an in-depth analysis of the contextual knowledge, trying to figure out RQ 2. It is found that noisy irrelevant contexts prevent models from learning the correct question-context-answer correlation(§5).

2 Related Work

2.1 Faithful Natural Language Generation

Recently more and more work has attracted significant interest in understanding the factual error, in summarization (Pagnoni et al., 2021; Ladhak et al., 2022; Tang et al., 2022) and machine translation (Müller et al., 2020; Raunak et al., 2021). There are also works about knowledge faithfulness in question answering (Krishna et al., 2021; Mahapatra et al., 2021; Longpre et al., 2021; Su et al., 2022) and dialogue response generation (Honovich et al., 2021; Dziri et al., 2022). For more details, we refer readers to the surveys (Li et al., 2022b; Ji et al., 2022). Although factoid hallucination is easier to encounter and research, we consider a more general scene with non-factoid information (i.e., debate or opinion in this work).

2.2 Knowledge Transfer

Knowledge transfer requires models to fit in the dynamic given information instead of remembering parametric knowledge. Prabhumoye et al. (2019) and West et al. (2022) researched Wikipedia writing, probing the model grounding ability. There are also lots of works about question answering under dynamic knowledge (Min et al., 2020; Longpre et al., 2021; Zhang and Choi, 2021; Chen et al., 2021; Wang et al., 2022; Liska et al., 2022; Kasai et al., 2022). The most similar work is Longpre et al. (2021), which focused on entity-based knowledge conflict and was under the open-domain setting. However, we investigate long-form question answering (LFQA) and transfer the whole knowledge text rather than just entities. All transferred knowledge is relevant and natural in the real world, since the false contexts may conflict with parametric knowledge and likely encourage the model to generate hallucinations.

3 Methods

3.1 Task: Question Answering under Knowledge Transfer

Knowledge transfer requires the model to generate a new answer grounding on newly transferred knowledge for the same question in training. Given a dataset D with splits D_{train} and D_{test}, we first train a knowledge-grounded generative model on training examples $(q_i, c_i, a_i) \in D_{train}$ (where q_i is the query, c_i is the context sentences including positive (c_i^+) and negative (c_i^-) contextual knowledge, and a_i is the golden answer, respectively). Then the model is benchmarked on examples $(q_j, \hat{c}_j) \in D_{test}$, where the query q_j can be found in D_{train}, but the contextual knowledge c_j is transferred to $\hat{c}_j$.

We use query-based summarization data, Debatepedia (Nema et al., 2017), to construct the relevant benchmark. The detailed data construct can be found in Appendix B.

3.2 Measure: Marginal Error Ratio

As shown in Figure 1, when the trained model is benchmarked on transferred contextual knowledge, it fails to generate a new answer grounded on given contexts but hallucinates from memory. We treat it as a *grounding failure of knowledge transfer*. Inpired by Factual Ablation (West et al., 2022), we propose *margin grounding failure (MF)* that enforces a significant gap:

$$MF(\Phi) = \begin{cases} 1, \Phi(\hat{a}, r_{train}) > m \cdot \Phi(a, r_{test}) \\ 0, \Phi(\hat{a}, r_{train}) \leq m \cdot \Phi(a, r_{test}) \end{cases} \tag{1}$$

where m denotes the margin, and Φ is any evaluation metric with the predicted answer $\hat{a}$ and golden reference r as inputs. The reference r comes from either the test or train set[1], which can be the golden answer or the contextual knowledge.

Note that the grounding failure is a binary label for each case. To statistically probe the faithfulness

[1] For cases with more than one reference, we calculate their scores separately and take the maximum one.

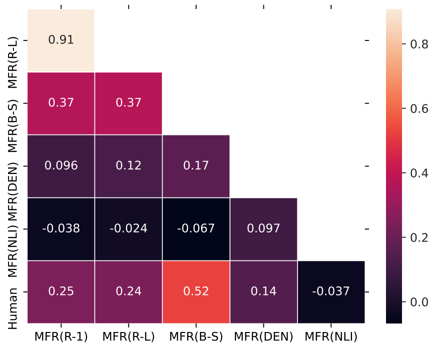


Figure 2: The Pearson correlation of margin failure ratio from each metrics and human evaluation.

Model	Experimental Data	
	Original	Extractive
BART-Base	4.01	0.00(↓4.01)
BART-Large	2.51	0.00(↓2.51)
BART-Large-xsum	3.18	0.00(↓3.18)
FiD(BART-Base)	3.85	0.84(↓3.01)
FiD(BART-Large)	2.84	0.50(↓2.34)
FiD(BART-Large-xsum)	6.52	0.50(↓6.02)
T5-Small	2.68	0.00(↓2.68)
T5-Base	2.34	0.00(↓2.34)
FiD(T5-Small)	3.01	0.50(↓2.51)
FiD(T5-Base)	3.68	0.50(↓3.18)

Table 1: MFR(BERT-Score) from different models. Extractive Data denotes the extractiveness-augmentation from Original Data in §5.

over the test set, we propose to measure the percentage of grounding failure of knowledge transfer. So the *margin failure rate (MFR)* is defined as:

$$MFR(\Phi) = \frac{1}{N}\sum_{i=1}^{N} MF_i(\Phi). \qquad (2)$$

Note that The margin m in this measure is adjustable. In this work, we tune this hyperparameter via the golden labels to search for the best-correlated measure with human (§4).

4 Results

We manually evaluate some results on a small scale and then use these labeled data to tune the MFR. With the adjusted MFR measure, we present results for BART and T5, the state-of-the-art seq2seq pretrained models in both QA tasks. The FiD (Izacard and Grave, 2021) architecture is also applied due to its effective and efficient utilization of extensive documents. The experimental setting is attached in Appendix C.

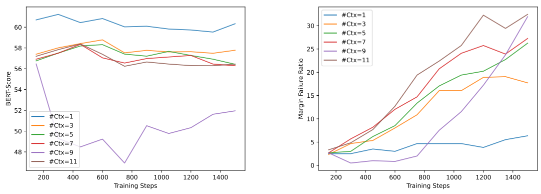
MFR(BERT-Score) can be a reliable alternative for human evaluation. We ask human judges for hallucination assessments. We provide the human evaluation details and some case studies in Appendix D.

We take the metrics Φ from two perspectives: the similarity with golden answers; the faithfulness to contextual knowledge. Concretely, for answer similarity metrics, we use ROUGE(-1/L) and BERT-SCORE (Zhang et al., 2020a); for knowledge faithfulness metrics, we use Density(Grusky et al., 2018) and NLI-Score[2]. For each metric Φ in MFR, we search its specific margin from 1.00 to 2.00 with the step of 0.01, by maximizing its Pearson correlation with human labels. The final tuned margin of ROUGE-1, ROUGE-L, BERT-Score, Density and NLI score are 1.93, 1.89, 1.41, 1.3, and 1.96.

We measure the Pearson correlation between each version of MFR and human evaluation. As depicted in Figure 2, all automatic metrics are little related to each other, except MFR(ROUGE-1) and MFR(ROUGE-L). There is even little relationship between MFR of MFR(NLI-Score) and human evaluation. MFR(BERT-Score) performs best correlatively with human evaluation, so we mainly take MFR(BERT-Score) as the main measure for the following experiments.

All models have memory hallucination under knowledge transfer, but only in rare cases. Table 1 represents the MFR(BERT-Score) of different models under knowledge transfer. The Original Data column denotes the primitively constructed benchmark in Appendix B. It is found that all models have the issue of generating memory hallucinations, though different models expose issues to different extents. However, such issues are not that severe. We also observe that models tend to generate answers which are lexically like memory while are, in fact, faithful to contexts (some case studies in Table 3). Generative models seem to be underestimated (Longpre et al., 2021; Kasai et al., 2022) due to the poor knowledge retriever. It is reasonable that models tend to generate hallucination when retrieved knowledge is irrelevant to the question. It is more convincing that we always provide relevant knowledge of answers eliminating the confounder of the retriever.

[2] We take the entailment probability from the RoBERTa-Large classifier fine-tuned on MNLI as NLI-Score.



(a) BERT-Score (b) MFR(BERT-Score)

Figure 3: The influence of the scale of contextual knowledge and training step on BERT-Score and MFR(BERT-Score).

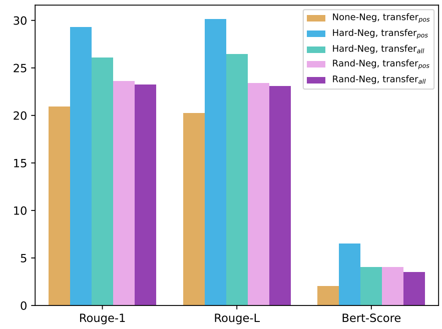


Figure 4: The MFR results over different settings of contexts. Detailed context setting is available in Appendix E

5 Analyzing the Original of Hallucination

In this section, we try to figure out the causality affecting model faithfulness under knowledge transfer. We conduct experiments by manipulating contexts from several perspectives.

Abstractiveness prevents the model to learn to ground on contexts. Abstractiveness measures the lexical overlap extent between contexts and answers. The training answer is evidently too abstractive to lead the model to learn the grounding ability. So we augment the oracle data by appending golden answers to the contextual knowledge to construct fully extractive QA data [3]. Results of different models trained on this data are also presented in Table 1. Models handle the augmented data with little faithfulness problem. It is also evident that models are underestimated on extractive data (Longpre et al., 2021; Kasai et al., 2022). Nevertheless, how to generate abstractive but faithful results still remains challenging (Dreyer et al., 2021; Ladhak et al., 2022).

The larger scale of contextual knowledge increases the burden of grounded generation. We take FiD(BART-Large-xsum) as an example, and evaluate the BERT-Score and MFR(BERT-Score) under different scale settings of contextual knowledge. It is obvious that the MFR increases as the context scale grows (Figure 3). More contexts bring more information but also more irrelevant noise. The noisy contexts prevent the model to ground on correct knowledge and confuse the model during generation (analyzed later in Figure 4). It is necessary to consider the information and noise trade-off, since it is meaningful in real application to retrieve more knowledge with an imperfect retriever. Moreover, training more steps also encourages the model overfitted on question-answer-only spurious correlation.

Irrelevant noisy context affects faithful generation during both training and testing. Also with FiD(BART-Large-xsum), we adopt different settings of contextual knowledge for experiments. During training, we provide negative contexts through retrieval (Hard Neg) or random sampling (Rand-Neg). During testing, we can transfer only the positive context with negative contexts unchanged (transfer$_{pos}$), or also transfer negative contexts by random ones (transfer$_{all}$). Detailed information can be found in Appendix E. The final comparative results are presented in Figure 4. Providing negative contexts significantly increases margin grounding failure. Comparing transfer$_{pos}$ with transfer$_{all}$, it is concluded that the model is unintendedly grounded on irrelevant knowledge, since transferring negative contexts would cause the generated answer to change, which is not expected. Hard Neg is a tough confounding that may induce models to learn spurious correlation, since retrieved knowledge is much more relevant to the question than sampled ones.

6 Conclusion

In this work, we research the memory hallucination under knowledge transfer. We benchmark several models and find they might be unfaithful to contextual knowledge in rare cases. Furthermore, we also reveal that context is a critical factor in hallucination during both training and testing. Although memory hallucination seems like a needle in a haystack, it is still an important issue hurdling faithful natural language generation into the real application, that needs to be solved out.

[3] The extractive fragment coverage of training data is upgraded from 0.61 to 1.00, and the extractive fragment density is enhanced from 1.00 to 9.26 after augmentation.

References

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Markus Dreyer, Mengwen Liu, Feng Nan, Sandeep Atluri, and Sujith Ravi. 2021. Analyzing the abstractiveness-factuality tradeoff with nonlinear abstractiveness constraints. *CoRR*, abs/2108.02859.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar R. Zaïane, and Siva Reddy. 2022. 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, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5271–5285. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 708–719. Association for Computational Linguistics.

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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7856–7870. 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, EACL 2021, Online, April 19 - 23, 2021*, pages 874–880. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *CoRR*, abs/2208.03299.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *CoRR*, abs/2202.03629.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir R. Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2022. Realtime QA: what's the answer right now? *CoRR*, abs/2207.13332.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4940–4957. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen R. McKeown. 2022. Faithful or extractive? on mitigating the faithfulness-abstractiveness tradeoff in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1410–1421. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022a. A survey on retrieval-augmented text generation. *CoRR*, abs/2202.01110.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022b. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *CoRR*, abs/2203.05227.

Adam Liska, Tomás Kociský, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien de Masson d'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-McMahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 13604–13622. PMLR.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7052–7063. Association for Computational Linguistics.

Suchismit Mahapatra, Vladimir Blagojevic, Pablo Bertorello, and Prasanna Kumar. 2021. New methods & metrics for LFQA tasks. *CoRR*, abs/2112.13432.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5783–5797. Association for Computational Linguistics.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas, AMTA 2020, Virtual, October 6-9, 2020*, pages 151–164. Association for Machine Translation in the Americas.

Preksha Nema, Mitesh M. Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1063–1072. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4812–4829. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Shrimai Prabhumoye, Chris Quirk, and Michel Galley. 2019. Towards content transfer through grounded text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2622–2632. 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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1172–1183. Association for Computational Linguistics.

Dan Su, Xiaoguang Li, Jindi Zhang, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Read before generate! faithful long form question answering with machine reading. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 744–756. Association for Computational Linguistics.

Liyan Tang, Tanya Goyal, Alexander R. Fabbri, Philippe Laban, Jiacheng Xu, Semih Yahvuz, Wojciech Kryscinski, Justin F. Rousseau, and Greg Durrett. 2022. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. *CoRR*, abs/2205.12854.

Jiexin Wang, Adam Jatowt, and Masatoshi Yoshikawa. 2022. Archivalqa: A large-scale benchmark dataset for open-domain question answering over historical news collections. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3025–3035. ACM.

Peter West, Chris Quirk, Michel Galley, and Yejin Choi. 2022. Probing factually grounded content transfer with factual ablation. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3732–3746. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Michael J. Q. Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into QA. In

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, pages 7371–7387. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yuhao Zhang, Derek Merck, Emily Bao Tsai, Christopher D. Manning, and Curtis P. Langlotz. 2020b. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5108–5120. Association for Computational Linguistics.

A Limitation and Future Work

Benchmark dataset It is hard to find so many datasets for long-form abstractive QA under knowledge transfer. Although Debatepedia is suitable for this experiment, its data scale and quality may not be fully guaranteed. It limits us to research the elements influencing faithfulness, including the data scale and the abstractiveness of the answer to contexts. Actually, we conclude four levels of transfer in knowledge-augmented text generation: (i) training on the general domain, then testing on a specific domain; (ii) training on one specific domain, then testing on another specific domain; (iii) training on one subclass of a specific domain, then testing on another subclass of the same domain; (iv) training on old version knowledge, then testing on new ones. All of these scenarios are realistic due to data scarcity or training cost. We hope more domains and more level knowledge transfer would be researched in future work.

Evaluation metrics Existing automatic evaluation metrics still correlate poorly with human evaluation (§4). It is necessary to propose an alternative method to systematically evaluation large scale results, trying to reduce the variance in small scale data.

Faithfulness improvement The final goal of faithfulness probing is to build an faithful generative model. This work lacks methods to improve generative model faithfulness. We will take a further step to research on hallucination causality and propose methods to solve this issue.

B Benchmark Construction

Unlike previous work (Longpre et al., 2021) , we follow the more natural setting where the transferred contextual knowledge is also factual. Besides we make the question answerable as a necessary condition. Because we find the models prefer to generate hallucination when given contextual knowledge does not contribute to answer the question.

To construct long-form QA data, we reuse Debatepedia(Nema et al., 2017), an abstractive summarization data, to supply our experiments. We choose this data due to its high abstractiveness and natural knowledge transfer condition. We observe that there are lots of lexically similar examples, so we deduplicate examples whose Levenshtein distance is less than 4. This filtered dataset satisfies the format of (q_i, c_i^+, a_i), and there are lots of questions paired with different contextual knowledge and answer. The examples with the same question are gathered, and one of them with the most distinctive answer is splited into development set. To enrich the contextual information of every cases, we apply BM25 to retrieve negative knowledge c_i^- from the whole dataset contexts via the question. Both relevant c_i^+ and irrelevant c_i^- contexts are merged into c_i. Because if there is only c_i^+, the question q_i is meaningless to position the positive context. In our basic setting, the contexts consists of 1 positive c_i^+ plus 4 negative c_i^-. The final processed dataset contains 2,549 training examples, 631 validation examples, and 598 test examples.

C Experimental Setting

Parameter	Value
Learning Rate	5×10^{-5}
Batch Size	16
Accumulation Steps	1
Total Step	4500
Warmup Step	150
Evaluate Step	150
Weight Decay	0.0
Input Maximum Length	512
Output Maximum Length	100
Beam Size	4

Table 2: The experimental setting details. *Beam Size is the hyper-parameter of text generation in development and testing, while other parameters contribute to model training.

We implement all the models using Pytorch (Paszke et al., 2019) and Transformers (Wolf et al., 2020) toolkit. The training and evaluation hyper-parameters are presented in Table 2. We use Adam optimizer(Kingma and Ba, 2015) with linear scheduler. All the training is started from the same random seed for a single round. We choose the best model by ROUGE-L score on development set.

All the models are trained on a single NVIDIA V100 GPU with 32GB memory. Training BART-Large, BART-Large-xsum, FiD(BART-Large), FiD(BART-Large-xsum), T5-base, FiD(T5-base) takes approximately 3 hours. Training BART-base, FiD(BART-base), T5-small, FiD(T5-small) takes less than 1 hour.

D Human Evaluation

We ask two postgraduate students who major in natural language processing to manually evaluate the results. We also explain to them about memory hallucination under knowledge transfer. We choose to label the generated results from FiD(BART-Large-xsum), as we observe this model hallucinates more than others. Human evaluation for more models is planned for future work.

For efficiency we only label the examples whose generated answers get ROUGE-1 score more than 40 with the references in training data, rather than all the examples in test set. We believe only these cases could be hallucinated memory from training data. Notice that we only consider memory hallucination which comes from training(fine-tuning phrase), while other hallucination may also occur but not taken into account. The final labeled data consist of 598 items with only 22 memory hallucination. Some case studies are presented in Table 3.

E Context analysis settings

None Negative contexts (None-Neg): Only the positive contextual knowledge is given. During testing, transfer$_{pos}$ denotes transferring the only given positive knowledge.

Hard Negative contexts (Hard-Neg): The positive contextual knowledge is given, paired with retrieved hard negative knowledge via BM25. This is the more real setting, as we need to retrieve external knowledge under open domain. During testing, transfer$_{pos}$ denotes transferring the given positive knowledge, and transfer$_{all}$ denotes not only transferring the given positive knowledge but also substitute the negative knowledge by randomly sampled ones.

Random Negative contexts (Rand-Neg): The positive contextual knowledge is given, paired with randomly sampled negative knowledge. During testing, transfer$_{pos}$ denotes transferring the given positive knowledge, and transfer$_{all}$ denotes not only transferring the given positive knowledge but also substitute the negative knowledge by newly sampled ones.

Testing Data	Training Data	R-L	Label
QUESTION: genocide ? can the violence in darfur be considered genocide ? CONTEXT: joschka fischer . former german foreign minister and vice chancellor from 1998 to 2005 . “ the eu must act in darfur . targeted sanctions would be a real step towards stopping the killing . ” april 19th 2007 - “ ... there insufficient political will for an international force [in darfur] ... ” GOLDEN ANSWER: there is insufficient political will for military intervention in darfur PREDICTED ANSWER: the violence in darfur could be considered genocide.	QUESTION: genocide ? can the violence in darfur be considered genocide ? CONTEXT: genocide is defined by most to include the systematic murders of a group of peoples as well as deliberate displacement and abuse . more than # # people have died since # with other estimates ranging up to # # according to amnesty international and the un . over # million people have become displaced and many are in danger of starvation due to lack of water and food . conclusively darfur is the worst humanitarian abuse in africa . to the extent that the janjaweed is systematically overseeing this mass-murder and to the extent that the government is involved in supporting the janjaweed darfur ’s crisis can be considered a genocide . GOLDEN ANSWER: the violence in darfur could be considered genocide	$22.22/100.00$	True
QUESTION: changing menus : will mandatory calorie counts compel restaurants to improve menus ? CONTEXT: restaurants that get caught under-reporting calories on their menus may face not only fines from the government but also significant pr problems as stories of their manipulations reach and turn-off their customers . GOLDEN ANSWER: restaurants will not under-report calories and risk pr backlash . PREDICTED ANSWER: restaurants under-report calories on menus	QUESTION: changing menus : will mandatory calorie counts compel restaurants to improve menus ? CONTEXT: “ calorie disclosures fail to weigh whole enchilada ” . wall street journal . july 8 2009 : “ scripps television stations sent several menu items to testing labs and found some big deviations from posted calorie content most of them making menu items appear healthier than they are . for example two tests of applebee ’s cajun-lime tilapia meal found about 400 calories compared with the posted total of 310 . ” this means that restaurants may simply choose to lower their reporting of calories instead of actually lower the calories in the foods they are serving . GOLDEN ANSWER: restaurants frequently under-report calories on menus	$42.86/90.91$	False
QUESTION: wealthy : is a progressive tax system fair to the wealthy ? CONTEXT: david n. mayer . “ wealthy americans deserve real tax relief on principle ” . ashbrook center . october # - “ there is no correlation between the amount of taxes an american pays and whatever benefits if any he receives ; indeed a wealthy person may get fewer government services than a poorer person . ” GOLDEN ANSWER: the rich do not necessarily benefit more from taxes/system PREDICTED ANSWER: progressive tax system unfairly benefits the wealthy	QUESTION: wealthy : is a progressive tax system fair to the wealthy ? CONTEXT: it is unfair that people who earn more should pay at a progressive rate . even on a standard rate they already pay more tax because they have a higher taxable income . therefore progressive tax rates are a form of double taxation as higher earners pay tax on more income and then at a high level . this is further unfair to them since high earners are the least likely group to benefit from much taxpayer-funded activity e.g . welfare . GOLDEN ANSWER: flat tax fairly has wealthy pay proportionally more in taxes .	$12.50/23.53$	True
QUESTION: militia : does the # nd amendment secure an individual right to form an independent militia ? CONTEXT: an armed citizenry empowers citizens to protect themselves so that a big government does n’t have to . GOLDEN ANSWER: in order to form a militia citizens require guns and a right to own them PREDICTED ANSWER: the # nd amendment secured an individual right to bear arm for the purpose of self-defense	QUESTION: militia : does the # nd amendment secure an individual right to form an independent militia ? CONTEXT: an armed citizen can places a checking on inappropriate cops power and the emergence of a cops state . GOLDEN ANSWER: # nd amendment secured equally the right of the militia and the individual to arms .	$14.29/42.86$	False

Table 3: Case study of human evaluation. The X/Y in R-L denotes the ROUGE-L score of predicted answer with the golden answer in testing(X) or training(Y) data. And Label denotes the human label for memory hallucination under knowledge transfer.