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

Xinshuo Hu[1]*, Baotian Hu[1]†, Dongfang Li[1], Xiaoguang Li[2], Lifeng Shang[2]
[1]Harbin Institute of Technology, Shenzhen, [2]Huawei Noah's Ark Lab
yanshek.woo@gmail.com, hubaotian@hit.edu.cn, crazyofapple@gmail.com,
{lixiaoguang11, Shang.Lifeng}@huawei.com

Abstract

The present study introduces the knowledge-augmented generator, which is specifically designed to produce information that remains grounded in contextual knowledge, regardless of alterations in the context. Previous research has predominantly focused on examining hallucinations stemming from static input, such as in the domains of summarization or machine translation. However, our investigation delves into the faithfulness of generative question answering in the presence of dynamic knowledge. Our objective is to explore the existence of hallucinations arising from parametric memory when contextual knowledge undergoes changes, while also analyzing the underlying causes for their occurrence. In order to efficiently address this issue, we propose a straightforward yet effective measure for detecting such hallucinations. Intriguingly, our investigation uncovers that all models exhibit a tendency to generate previous answers as hallucinations. To gain deeper insights into the underlying causes of this phenomenon, we conduct a series of experiments that verify the critical role played by context in hallucination, both during training and testing, from various perspectives.

Keywords: Text Generation, Faithfulness, Question Answering

1. Introduction

Knowledge-augmented text generation method (e.g. RAG (Lewis et al., 2020b), FiD (Izacard and Grave, 2021)), and Atlas (Izacard et al., 2022), have demonstrated state-of-the-art (SOTA) performance across various NLP tasks. The paradigm of generating text using external knowledge offers the advantage of plug-and-play through non-parametric contextual knowledge. In contrast, parametric knowledge embedded within models necessitates retraining for updates (Li et al., 2022a). A faithful knowledge-augmented generator should consistently produce output that aligns with the contextual grounding (Ji et al., 2022). However, the presence of hallucinations originating from parametric memory (see Figure 1) poses a significant challenge for practical text generation applications (Maynez et al., 2020; Zhang et al., 2020b).

The investigation of the faithfulness of generative models in the presence of dynamic contextual knowledge remains an ongoing research area. Previous studies have primarily focused on analyzing hallucinations in scenarios where the input texts during training and testing are independent, such as in summarization(Pagnoni et al., 2021; Ladhak et al., 2022; Tang et al., 2022) or machine translation(Raunak et al., 2021; Müller et al., 2020). While knowledge-dynamic question answering has garnered attention in several works (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; Chen et al., 2023), only a few studies have systematically quantified the extent of model faithfulness or analyzed the circumstances and reasons behind hallucination generation in the

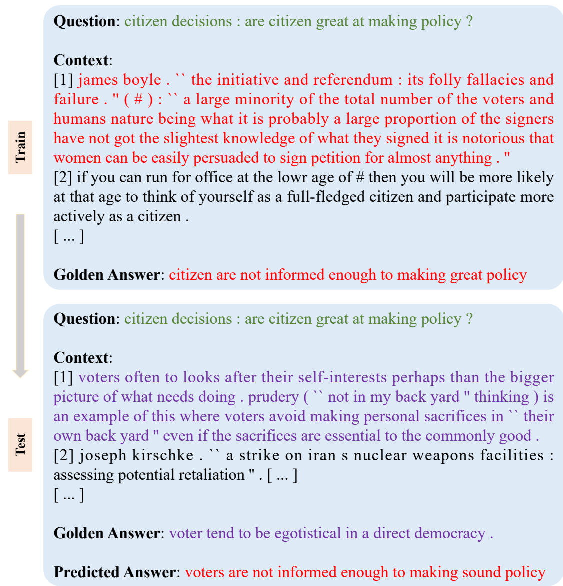

Figure 1: An example of generated hallucination from training memory. The model disregards the transferred contextual knowledge and predicts an out-of-date answer that was present in its original training data when answering the same question. Non-essential details are ignored by [...].

*This work is done when Xinshuo Hu is an intern at Huawei Noah's Ark Lab.
†Corresponding author.

LREC-COLING 2024, pages 5202–5211
20-25 May, 2024.

presence of dynamic contextual knowledge (Longpre et al., 2021; West et al., 2022). In this study, we define *context transfer* as the process of contextual knowledge changing while the question remains the same. Specifically, the generative model is trained on old knowledge but evaluated on new knowledge instances. Our analysis focuses on *memory hallucination* which refers to hallucinations generated by parametric knowledge during context transfer.

In this work, our objective is to assess the faithfulness of generative models in the context of context transfer, focusing on two primary research questions:

RQ 1 *To what extent does the generative model exhibit faithfulness under context transfer?*

RQ 2 *What are the underlying reasons for the occurrence of memory hallucination?*

To address these research questions, we first define the context transfer task and introduce a novel metric for measuring hallucination (§3). Subsequently, we conduct comprehensive experiments involving multiple models to investigate Research Question 1. Our findings indicate that models do not consistently exhibit grounded behavior in the presence of context transfer (§4). To gain deeper insights into the issue raised in Research Question 2, we perform an in-depth analysis of contextual knowledge, revealing that the presence of noisy and irrelevant contexts hinders models from effectively capturing the desired question-context-answer correlation (§5).

2. Related Work

2.1. Faithful Natural Language Generation

Faithful natural language generation (NLG) aims to generate text that is both faithful and consistent with the input information, while avoiding hallucination (Li et al., 2022b; Ji et al., 2022). In recent years, there has been a growing interest in understanding factual errors 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). Additionally, there have been studies focusing on knowledge faithfulness in question answering (Krishna et al., 2021; Mahapatra et al., 2021; Longpre et al., 2021) 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 has been extensively studied, our work focuses on a broader scope by considering non-factoid information, such as debates and opinions.

2.2. Context Transfer

Context transfer in NLG involves models adapting to dynamically provided information rather than relying solely on pre-learned parameters. This aspect has been explored in studies on Wikipedia writing by Prabhumoye et al. (2019) and West et al. (2022), investigating the model grounding ability. Furthermore, several works have addressed question answering in the context of 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), where we transfer the entire knowledge text rather than solely editing entities. All transferred knowledge remains relevant and aligned with the real world, as false contextual information may conflict with pre-learned knowledge and potentially induce hallucinations in the model.

3. Methods

3.1. Task: Question Answering under Context Transfer

Context transfer necessitates the model's ability to generate a novel answer based on newly acquired knowledge for the same question during training. To begin, we employ a dataset D consisting of two partitions, namely D_{train} and D_{test}. Our initial step involves training a knowledge-grounded generative model on the training examples $(q_i, c_i, a_i) \in D_{train}$, where q_i represents the question, c_i consists of contextual sentences comprising positive (c_i^+) and negative (c_i^-) contextual knowledge, and a_i denotes the golden reference answer. Subsequently, the model is evaluated using examples $(q_j, \hat{c_j}) \in D_{test}$, wherein the query q_j can be found in D_{train}, while the contextual knowledge c_j is transferred to $\hat{c_j}$.

Our primary focus lies in abstractive long-form question answering. We consider entity-based question answering to be straightforward, as hallucination can be mitigated or even resolved through extraction-augmentation and post-editing techniques. To construct a relevant benchmark, we utilize query-based summarization data from Debatepedia (Nema et al., 2017), primarily due to its highly abstract nature and natural conditions for context transfer. In contrast to previous research (Longpre et al., 2021), we adopt a more natural setting where the transferred contextual knowledge is factual as well. Furthermore, we ensure that the questions are answerable, considering it a necessary requirement. This precaution is taken because we have observed that models tend to generate hallucinatory responses when the contextual knowledge does not contribute to answering the question effectively.

3.2. Measure: Margin Failure Rate

As illustrated in Figure 1, the trained model exhibits a failure in grounding transferred contextual knowledge, resulting in the generation of answers that are not properly aligned with the given contexts. This phenomenon is referred to as a *grounding failure of context transfer*.

To determine whether a predicted answer $\hat{a}$ represents a grounding failure of context transfer, we introduce the concept of *margin grounding failure* ($\mathcal{MF}$) as follows:

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

where m represents the hyperparameter margin, and Φ is a basic metric (e.g. ROUGE) to measure the similarity between the predicted answer $\hat{a}$ and golden reference r. The reference r comes from either the train or test set (r_{train} from the train set or r_{test} from the test set), which can be the golden answer or the contextual knowledge[1].

It is important to note that grounding failure is a binary label assigned to each case. To statistically probe the faithfulness over the test set, we propose to measure the percentage of grounding failure of context transfer. So the *margin failure rate* ($\mathcal{MFR}$) is defined as:

$$\mathcal{MFR}(\Phi) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{MF}_i(\Phi). \quad (2)$$

In this work, we use BERT-SCORE (Zhang et al., 2020a) as our basic metric Φ. For our experiments, we set the margin m to a value of 1.25 based on intuition, which has a relatively strong correlation with Pearson Correlation of 0.43 with human evaluation on our development set.

4. Results

In this study, we present the outcomes obtained from two prominent state-of-the-art sequence-to-sequence (seq2seq) pre-trained models, namely BART (Lewis et al., 2020a) and T5 (Raffel et al., 2020), in the context of question answering (QA) tasks. Besides the vanilla transformer architecture, we also incorporate the FiD method (Izacard and Grave, 2021) owing to its efficient and effective utilization of extensive document collections. The model selection process is based on the ROUGE-L score achieved on the development set.

[1] In cases where there are multiple references, individual scores are calculated, and the maximum score is selected.

Model	Decoding Strategy	
	Greedy	Beam Search
T5_{small}	7.69	8.19
T5_{base}	7.53	6.19
BART_{base}	9.20	10.87
BART_{large}	7.86	8.36
$\text{BART}_{large-xsum}$	8.03	7.19
FiD (T5_{small})	11.37	9.53
FiD (T5_{base})	11.04	10.03
FiD (BART_{base})	13.88	12.71
FiD (BART_{large})	10.03	8.86
FiD ($\text{BART}_{large-xsum}$)	15.38	14.55

Table 1: The $\mathcal{MFR}$(BERT-Score) results of different models. We generate text by greedy and beam search (beam=4) decoding strategy.

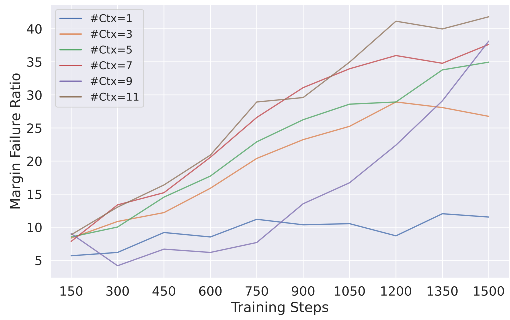

Figure 2: The influence of the scale of contextual knowledge and training step on $\mathcal{MFR}$(BERT-Score).

All models have memory hallucination under context transfer. The $\mathcal{MFR}$(BERT-Score) results of various models under context transfer are presented in Table 1. It is observed that all the models exhibit the phenomenon of memory hallucination during context transfer, albeit to varying degrees. The choice of decoding strategies does not appear to have a significant impact on the generation of hallucinations. Specifically, the FiD method demonstrates a higher occurrence of context transfer grounding failure compared to the vanilla transformer. This can be attributed to the fact that FiD has a tendency to memorize the question-answer pairs, as the questions are duplicated for each context.

5. Analysis

In this section, we endeavor to elucidate the intricate interplay between causality and its impact

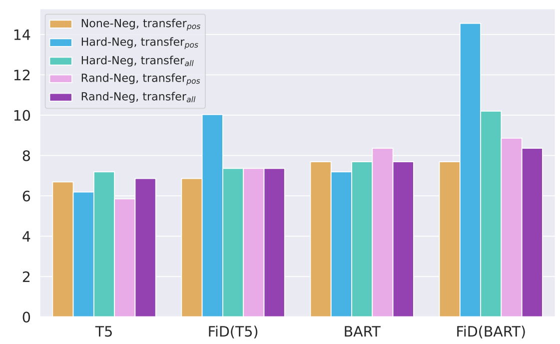

Figure 3: The $\mathcal{MFR}$(BERT-Score) results over different settings of contexts.

on model faithfulness within the realm of context transfer. To this end, we embark upon a series of rigorous experiments, wherein we manipulate contextual factors from various perspectives, in order to derive meaningful insights. We conduct all the analysis on FiD(BART$_{large-xsum}$).

Impact of Contextual Knowledge Scale We examine the effect of varying the scale of contextual knowledge on the performance of FiD (BART$_{large-xsum}$) as measured by the $\mathcal{MFR}$(BERT-Score). It becomes evident that the $\mathcal{MFR}$ value increases proportionally with the expansion of the context scale (Figure 2). This surplus of noisy contexts hampers the model's ability to ground itself in accurate knowledge and introduces confusion during the generation process, as elaborated upon later in Figure 3. Therefore, it becomes crucial to strike a balance between the quantity of information retrieved and the presence of noise, particularly in practical applications where obtaining more knowledge through an imperfect retriever holds significance. Furthermore, it is worth noting that training the model for an extended duration may lead to overfitting on question-answer spurious correlations. Notably, the $\mathcal{MFR}$(BERT-Score) can reach as high as 20 after a mere 600 training steps, equivalent to approximately four epochs.

Impact of Irrelevant Noisy Context The presence of irrelevant noisy context can have a detrimental effect on faithful generation during both the training and testing phases. In our experiments, we explore different settings of contextual knowledge using the T5$_{base}$ and BART$_{large-xsum}$. During the training process, we introduce negative contexts using two different methods: retrieval-based methods (referred to as Hard-Neg) nd random sampling (referred to as Rand-Neg). For testing, we consider two scenarios: transferring only the positive context while keeping the negative contexts unchanged (referred to as transfer$_{pos}$), or transferring both the positive and negative contexts by replacing the latter with random ones (referred to as transfer$_{all}$). The detailed settings are as follows:

1) **None Negative Contexts (None-Neg):** Only positive contextual knowledge is provided during training. During testing, we transfer only the positive knowledge (**transfer**$_{pos}$).

2) **Hard Negative Contexts (Hard-Neg):** In this setting, we provide the positive contextual knowledge along with retrieved hard negative knowledge using BM25. This setting is more realistic as it involves retrieving external knowledge in an open domain. During testing, **transfer**$_{pos}$ refers to transferring only the positive knowledge, while **transfer**$_{all}$ refers to transferring both the positive and negative knowledge, with the negative knowledge being randomly sampled.

3) **Random Negative Contexts (Rand-Neg):** Similar to the Hard-Neg setting, we provide the positive contextual knowledge, but pair it with randomly sampled negative knowledge. The testing scenarios (**transfer**pos and **transfer**$_{all}$) remain the same as in the Hard-Neg setting.

The final comparative results are presented in Figure 3. Notably, there is a drop on $\mathcal{MFR}$(BERT-Score) for the FiD architecture when tested on transfer$_{all}$, specially trained on hard negative contexts. The presence of hard negative contexts poses a challenging confounding factor, as it may induce models to learn spurious correlations, given that retrieved knowledge is often more relevant to the question than sampled knowledge. Furthermore, our findings align with the conclusions drawn from Figure 2, indicating that the inclusion of negative contexts significantly increases the occurrence of margin grounding failure. However, it is worth noting that the vanilla transformer architecture exhibits robustness against negative contexts, displaying insensitivity to contextual disturbance. Upon comparing transfer$_{pos}$ with transfer$_{all}$, we observe that the model unintentionally grounds its answers on irrelevant knowledge when negative contexts are transferred, leading to unexpected changes in the generated answers.

6. Conclusion

This study endeavors to explore the phenomenon of memory hallucination in the realm of context transfer. Our investigation entails the comprehensive examination of multiple models, unveiling potential deficiencies in their ability to faithfully align contextual knowledge. Furthermore, our research emphasizes the pivotal role played by context in the manifestation of hallucinations during both training and testing phases. Despite the apparent rarity of memory hallucination, it represents a critical concern that demands attention for the attainment of veracious natural language generation in practical settings. We anticipate that this research will contribute to a more profound comprehension of the faithfulness of generative models.

Limitations

Benchmark Dataset Acquiring suitable datasets for long-form abstractive Question Answering (QA) in the context of context transfer poses a significant challenge. Although Debatepedia may initially seem appropriate for such experiments, the reliability of its data scale and quality is questionable, thereby limiting our ability to investigate the factors that influence answer faithfulness. We anticipate that future research will explore additional domains and levels of context transfer, expanding the scope of investigation.

Evaluation Metrics Existing automatic evaluation metrics Existing automatic evaluation metrics demonstrate limited correlation with human evaluations. Therefore, it is crucial to propose an alternative methodology for systematically assessing large-scale results, with the aim of reducing the variance inherent in small-scale data.

Evaluation Models Owing to constraints in resources, comprehensive experimentation on the prevalent large language models, has not been undertaken. Nonetheless, we have intentions to incorporate experiments pertaining to large language models in our future endeavors, contingent upon the feasibility thereof.

Faithfulness Improvement The primary goal of faithfulness probing is to establish a generative model that faithfully incorporates and aligns with the provided context. Nevertheless, this work lacks methodologies to enhance the faithfulness of generative models. Consequently, we try to advance this investigation by exploring the causal factors behind hallucination and proposing viable solutions to address this intricate challenge.

Acknowledgments

We thank Shaobo Li for his insightful suggestions and invaluable feedback. This work is jointly supported by grants: Natural Science Foundation of China (No. 62376067), and Guangdong Basic and Applied Basic Research Foundation (2023A1515110078).

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*.

Ziyang Chen, Dongfang Li, Xiang Zhao, Baotian Hu, and Min Zhang. 2023. Temporal knowledge question answering via abstract reasoning induction. *CoRR*, abs/2311.09149.

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.

Tanya Goyal, Jiacheng Xu, Junyi Jessy Li, and Greg Durrett. 2022. Training dynamics for text summarization models. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2061–2073. 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.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1967–1972. The 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 trade-off 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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. 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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. 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. 2020b. 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.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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.

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5602–5609. AAAI Press.

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. 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 the contextual knowledge does not contribute to answering 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 context 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 split into the development set. To enrich the contextual information of every case, 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 consist of 1 positive c_i^+ plus four negative c_i^-. The final processed dataset contains 2,549 training examples, 631 validation examples, and 598 test examples.

B. 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 the 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 the 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,

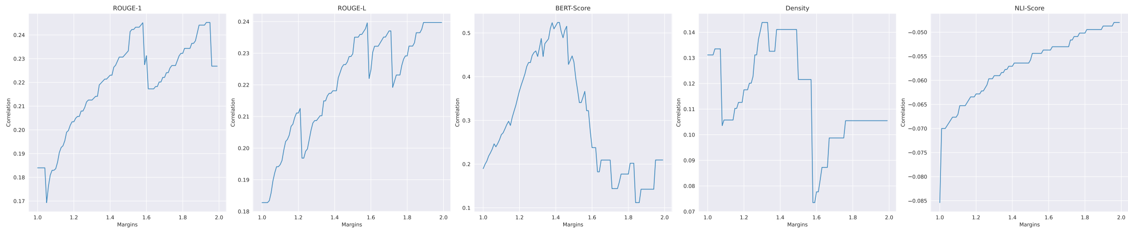

Figure 4: The Pearson correlation of margin failure ratio from basic metrics with different margins.

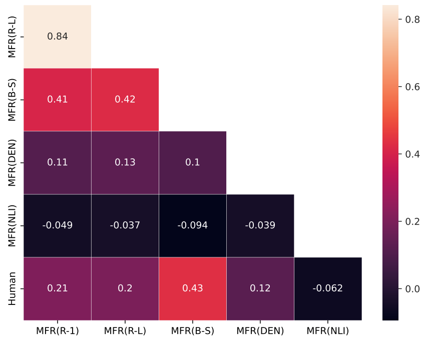

Figure 5: The Pearson correlation of margin failure ratio from each metric and human evaluation.

FiD(BART-base), T5-small, FiD(T5-small) takes less than 1 hour.

C. Meta Evaluation of MFR

We manually evaluate the grounding failure of context transfer on a small scale from test data in order that we can measure the Pearson Correlation between MFR and human labels. 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 context 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. We only label the examples whose generated answers get ROUGE-1 score of more than 40 with the references in training data rather than all the examples in the 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 hallucinations may also occur but are not taken into account. The final labelled data consist of 598 items with only 22 memory hallucination. Some case studies are presented in Table 3.

We measure the Pearson correlation between different versions of MFR and human evaluation. We take the basic metrics Φ from two perspectives: the similarity with golden answers; the faithfulness to contextual knowledge. Concretely, for basic metrics of answer similarity, we use ROUGE(-1/L) and BERT-SCORE (Zhang et al., 2020a); for basic metrics of knowledge faithfulness, we use Density(Grusky et al., 2018) and NLI-Score[2]. As depicted in Figure 5, all automatic metrics are only a little related to each other, except MFR(ROUGE-1) and MFR(ROUGE-L). There is even little relationship between MFR(NLI-Score) and human evaluation. MFR(BERT-Score) performs best correlatively with human evaluation, so we take MFR(BERT-Score) as the main measure in this work.

We also measure the influence of the margin m. For each metric Φ in MFR, we experiment with its margin varying from 1.00 to 2.00 with a stripe of 0.01. As shown in Figure 4, the margin m has a great impact on the human correlation of MFR and different basic metrics achieve the best performance at different margins. Although the intuitively chosen margin $m = 1.25$ is not the perfect hyperparameter of BERT-Score, it still has a relatively strong correlation with Pearson Correlation of 0.43.

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

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.