Improving the Faithfulness of LLM-based Abstractive Summarization with Span-level Unlikelihood Training

Sicong Huang Qianqi Yan Shengze Wang Ian Lane
University of California, Santa Cruz
{shuan213, qyan79, shengze, ialane}@ucsc.edu

Abstract

Abstractive summarization using large language models (LLMs) has become an essential tool for condensing information. Despite their ability to generate fluent summaries, these models often produce texts that are unfaithful to the original documents. Current approaches to mitigating unfaithfulness typically involve post-processing corrections or contrastive learning from synthetically generated negative samples, which do not fully address the spectrum of errors that can arise in LLM-generated summaries. In this paper, we introduce a novel approach to fine-tune LLMs specifically to reduce the occurrence of unfaithful spans of text in generated summaries. We first annotate span-level hallucinations in LLM-generated summaries using automatic labeling with GPT-4. We then fine-tune the LLM using both summaries with no hallucinations and spans of hallucinated text to improve the faithfulness of the model. This paper introduces a dataset labeled to distinguish between faithful and unfaithful content and compare the performance of three techniques: *gradient ascent*, *unlikelihood training*, and *task vector negation*. Our experimental results show that unlikelihood training can effectively use span-level annotations to enhance summary faithfulness, reducing the number of summaries with hallucinations from 31% to 13%, a reduction of 58% on the CNN summarization dataset and from 33% to 20%, a reduction of 39% on the SAMSum dataset.

1 Introduction

Abstractive summarization aims to condense text by distilling key information from the source text and rewriting it concisely. Recent advances in large language models (LLMs) have significantly enhanced the capabilities of summarization systems, with retrieval-augmented generation (RAG) (Lewis et al., 2020) further emphasizing its importance in interactive natural language systems. However, LLMs still struggle with hallucination, or unfaithfulness in summarization, where generated summaries contain information not grounded in the source document. This limits the practical deployment of summarization systems.

Various approaches have attempted to address unfaithfulness: **Post-processing methods**: These edit and correct factual inaccuracies after generation or employ critique-and-refine processes. While effective, they increase latency and computational demands. **Learning from synthetic negative samples**: This involves creating synthetic unfaithful summaries for training. Even though it does not increase the inference latency, this approach still faces challenges: 1. Human reviewers often prefer LLM summaries over standard references, questioning the quality of training data. 2. Synthetic samples may not accurately replicate actual errors in model-generated summaries.

To address these issues, we propose annotating span-level hallucinations in LLM-generated summaries and updating the model using this fine-grained information. Our contributions include: 1. Constructing a dataset with span-level labeled faithful and unfaithful summaries. 2. Comparing three approaches (gradient ascent (Yao et al., 2024), unlikelihood training (Welleck et al., 2020), and task

vector negation (Ilharco et al., 2023)) that use span-level information to improve faithfulness. 3. Demonstrating that span-level unlikelihood training is the most effective of the three approaches.

2 Approach

Our approach of leveraging span-level labeling to improve summary faithfulness involves two steps: (1) Constructing a dataset of LLM-generated summaries with span-level hallucination annotations. (2) Comparing three fine-tuning techniques that can make use of span-level hallucination information: gradient ascent, unlikelihood training, and task vector negation.

We created a dataset using summaries generated by Llama-2-7b-chat for the CNN and SAMSum datasets. We use GPT-4 to automatically label spans of text inconsistent with the source documents. The resulting dataset contains 11,096 training samples and 200 test samples, evenly distributed between positive (faithful) and negative (unfaithful) examples.

We studied three methods for fine-tuning LLMs using both positive and negative samples:

1. **Gradient Ascent**: Reversing the sign of the cross-entropy loss for hallucinated spans.
2. **Unlikelihood Training**: Maximizing the complement of the probability of generating hallucinated tokens.
3. **Task Vector Negation**: Negating task vectors to reduce the influence of undesired traits.

We then evaluate the effectiveness of these techniques using automatic metrics, i.e. GPT-4 span labeling, G-Eval (Liu et al., 2023), and AlignScore (Zha et al., 2023). GPT-4 span labeling essentially follows the same process as creating the dataset, labeling spans of unfaithful text with GPT-4, then compute the percentage of tokens that are faithful out of all generated tokens. In addition, we also perform human evaluation on a subset of the generated summaries to validate the results given by automatic metrics.

3 Results

We compare our approaches against a baseline system that was trained only on positive examples with cross entropy loss. Our experiment results indicate that unlikelihood training is the most effective method for improving summary faithfulness. On the SAMSum dataset, unlikelihood training reduced hallucinated summaries from 33% to 20% (39% reduction). On CNN, it reduced hallucinated summaries from 31% to 13% (58% reduction). Task vector negation showed moderate improvements, reducing SAMSum hallucinated summaries by 15%, lower than that of unlikelihood training. It also reduced CNN hallucinated summaries by 58%, matching that of unlikelihood training. Gradient ascent, on the contrary, performed poorly, often severely deteriorating model performance, making its performance in reducing hallucination irrelevant.

These results were consistent across multiple automatic evaluation metrics (GPT-4 span labeling, G-Eval, and AlignScore). Human verification on the SAMSum summaries generated by the baseline systems and the unlikelihood trained system further validated these results.

4 Discussion and Conclusion

Our study introduces several key contributions: 1. A dataset of LLM-generated summaries with span-level hallucination annotations. 2. Comparison of three fine-tuning techniques that leverage span-level labeling for improving summary faithfulness. 3. Demonstration of the effectiveness of span-level unlikelihood training in reducing hallucinations.

Our work demonstrates the potential of leveraging span-level annotations and unlikelihood training to significantly improve the faithfulness of LLM-generated summaries, addressing a critical challenge in the field of abstractive summarization. Future work could explore the generalizability of this approach to other tasks and investigate the potential of GPT-4 as a fine-grained faithfulness metric.

References

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Patrick 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*, volume 33, pages 9459–9474. Curran Associates, Inc.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2024. Large language model unlearning.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.