Entity Embelishment Mitigation in LLMs Output with Noisy Synthetic Dataset for Alignment

Svitlana Galeshchuk
Arval Leasing Solutions BNP Paribas, West Ukrainian National University
22 r Deux Gares 92500 Rueil Malmaison France, 1 Lvivska str. Ternopil Ukraine
svitlana.galeshchuk@gmail.com

Abstract

The present work focuses on the entity embellishments when named entities are accompanied by additional information that is not supported by the context or the source material. Our paper contributes into mitigating this problem in large language model's generated texts, summaries in particular, by proposing the approach with synthetic noise injection in the generated samples that are further used for alignment of finetuned LLM. We also challenge the issue of solutions scarcity for low-resourced languages and test our approach with corpora in Ukrainian.

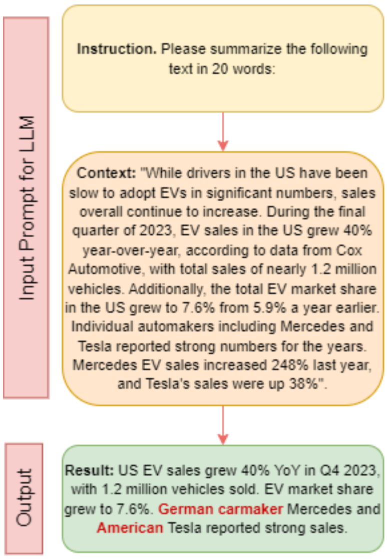
Keywords: large language models, Llama, summarization, Ukrainian NLP

1. Introduction

Text generation is a task that produces text conditioning on an input (a question, an article, an image, etc.). With the increase in number of Transformer models and availability of textual data, we are seeing a rapid growth in the number of text generation applications such as summarization, chatbots, storytelling, and machine translation. The fluency and diversity of automatically produced text has advanced significantly with the introduction of large and very large language models (LLMs). However, LLMs use a probabilistic approach to generate text, which makes these models prone to creating factually incorrect, inconsistent, or irrelevant information that is not supported in the input. This is called hallucination. In real-world applications, hallucinations can pose many problems, ranging from ethical risks to loss of trust from clients. As a result, scholars and practitioners in the field of natural language generation (NLG) have focused their research on mitigating the risk of adding irrelevant information.

Hallucinations problems can be broadly categorized into two types: **factuality hallucination** and **faithfulness hallucination**, as identified by Huang et al. (2023). Factuality hallucination is characterized by a discrepancy between the generated content and real-world facts that can be verified. On the other hand, faithfulness hallucination occurs when there is a deviation of the generated output from the instructions or context provided by the input. This type of hallucination can be further subcategorized into instruction, context, or logic inconsistencies. Future research in this area is crucial to enhance the quality of natural language generation output and to improve the accuracy and relevance of the generated text.

In the paper, we focus on the faithfulness problem, and context inconsistencies in particular when LLM generated output is imprecise or untrue compared to the user's input.



Instruction. Please summarize the following text in 20 words:

Context: "While drivers in the US have been slow to adopt EVs in significant numbers, sales overall continue to increase. During the final quarter of 2023, EV sales in the US grew 40% year-over-year, according to data from Cox Automotive, with total sales of nearly 1.2 million vehicles. Additionally, the total EV market share in the US grew to 7.6% from 5.9% a year earlier. Individual automakers including Mercedes and Tesla reported strong numbers for the years. Mercedes EV sales increased 248% last year, and Tesla's sales were up 38%".

Result: US EV sales grew 40% YoY in Q4 2023, with 1.2 million vehicles sold. EV market share grew to 7.6%. German carmaker Mercedes and American Tesla reported strong sales.

Figure 1: Example of entity hallucination we tackle in the paper.

Figure 1 illustrates the problem when a user asks a LLM to summarize a given article, we find added information on which are nationalities of Tesla and Mercedes that being true (in 2024) is not, however, mentioned in the article but assumed by the LLM as probable to be in the output.

We refer to this the type of context hallucinations that accompany named entities as entity embellishment and define mitigating them as the main scope of the paper. This brings us to the objective of the paper that aims at reducing the risk of context hallucination, in particular entity embellishment, in foundation models using summarization dataset

The Third Ukrainian Natural Language Processing Workshop (UNLP) @LREC-COLING-2024, pages 129–134
25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

and pertrurbated examples for model alignment via direct preferential optimization (DPO) procedure. More precisely, the development of LLMs involves two main stages:

- the first stage is **pre-training**, where the models learn general representations and acquire knowledge about the world
- the second stage is **alignment**, where the models are trained to better align with the instructions and preferences of users.

Our approach involves utilizing LLM by fine-tuning it with articles that come with their corresponding golden summaries. We then align the trained model by using generated texts that have been corrupted with injected information on named entities from another LLM, in particular GPT-4. The golden standard is considered as the chosen and preferred answer. During the direct preference optimization (DPO) phase of training, any synthetic response enriched with text from GPT-4 is shown to be rejected and golden summary to be chosen.

The occurrence of hallucinations in LLM output texts is a known issue. However, very few studies have explored how to mitigate hallucination problems in low-resource languages other than English. This is because the most of the pre-training corpora is usually in English for the majority of available LLMs. Consequently, these models may learn information in English and apply it to tasks in other languages. To challenge these limitations, we conducted tests in Ukrainian, a low-resource language, to verify the consistency of results in non-English documents.

The article is organized as follows: Section 2 elaborates on related work and the choice of evaluation metrics. Section 3 focuses on data used to train and align a LLM. Section 4 highlights the experimental setup described in Introduction together with the main challenges. Section 5 presents the results of the study and potential limits.

2. Related Work

Hallucination in text generation is a well-known phenomenon hence we find a plethora of scientific papers on the nature and solutions to LLM embellishments.

2.1. Surveys on hallucination phenomenon and its nature.

We cite several papers that elaborate on the survey analysis of LLM hallucinations. The study by Ji et al. (2023) mainly focuses on the occurrence of hallucinations in pre-trained language models for natural language generation tasks, while not discussing LLMs. The paper of Wang et al. (2023) concentrates on the factuality of LLMs-generated texts. Tonmoy et al. (2024) provides a taxonomy of mitigating approaches against hallucinations, stressing out prompt engineering with retrieval augmented generation and self-refinement through feedback and reasoning as well as prompt-tuning. Yao et al. (2023) demonstrate that nonsense prompts composed of random tokens can also elicit hallucinations in LLMs, suggesting that hallucination may be another view of adversarial examples. Huang et al. (2023), claims that LLMs have been known to create non-existent facts. Current explanations attribute this to the training datasets McKenna et al. (2023). These works argue that noisy data or model overfitting to the training data is responsible for hallucination. The authors believe that alignment, involving supervised fine-tuning and reinforcement learning is crucial for unlocking LLMs capabilities and aligning them with human preferences. However, it introduces the risk of hallucinations due to capability misalignment and belief misalignment, including sycophantic behavior driven by human preferences. Wiggers (2023) suggest that hallucinating models can serve as collaborative creative partners; providing valuable outputs that may not be factual but can lead to novel ideas. While hallucinations can be problematic when factually inaccurate, they can be advantageous in creative or artistic endeavors. In terms of related works for Ukrainian language, we cite Kang et al. (2024) who test multilingual BLOOM for hallucinations finding significant faithfullness issues in generated texts in Ukrainian.

2.2. Strategies to overcome hallucinations

Decoding strategies. Lango and Dušek (2023) highlight decoding strategies as techniques designed to target the generation phase of a model. With regards to hallucination, these techniques aim to reduce its occurrence in the generated outputs by guiding the generation phase toward producing authentic or context-specific content, Shi et al. (2023), expand their study to context-aware decoding relying on the intuition that a contrastive output distribution amplifies the difference between the output probabilities when a model is used with and without context. Choi et al. (2023) introduce a method called Knowledge-Constrained Decoding (KCD) that uses a token-level detection system to identify hallucinations and improve the generation process by adjusting the token distribution based on a more an accurate estimate of future knowledge groundedness. **Knowledge base strategies**. Zhang et al. (2023) address the issue of knowledge alignment by introducing MixAlign, a framework that interacts with both the user and the knowledge base to clarify the relationship between the user question and the information stored in the knowledge base. This approach while being effective for factual inconsistencies is not designed for faithfulness problems. **Training strategies.** DRESS: (Chen et al. (2023), propose using critique and refinement of natural language feedback to improve alignment with human preferences and tackle hallucination issues. This the approach allows us to define the setup of the paper that exploits the alignment stage to "show" the model the right and "wrong", corrupted samples with hallucinations.

2.3. Metric for hallucination

According to Azaria and Mitchell (2023),Ji et al. (2023), LLMs are capable of determining the factual accuracy of statements, even when the false statements are generated by the models themselves. The statement brings us to investigate the potential capabilities of LLMs to judge the faithfulness of generated texts without a need of a human annotator. Here are the metrics considered in our research:

- **N-gram**, (calculates the ratio of token overlap between the generated output and the correct answer) based metrics like ROUGE and PARENT-T assesses faithfulness but show poor correlation with humans thus their usage is very limited (Ji et al. (2023), Maynez et al. (2020)).

- **Feedback from another LLM**: Feng et al. (2023) proposes to employ GPT-4 to collect sentence-level factual consistency annotation for system-generated summaries. They make a comparison between GPT-4 and human annotations prove high correlation of the feedbacks.

- **Weekly supervised classifier finetuning**: , Kryściński et al. (2019) create a data set by corrupting golden summaries with paraphrasing, entity swapping, and noise injection. Similarly, Dziri et al. (2021) develop perturbated samples by replacing up to two verbs with verbs of the same tense or extracting all mentioned entities from different dialogue examples using the SpaCy NER tagger and corrupting them.

The overview of the literature helps define our experimental strategy by creating a dataset of adversarial summaries to golden summaries for news articles inspired by weekly-supervised approaches presented that are used as an input to LLM alignment phase rather than fine-tuning that is advocated by Chen et al. (2023). We then apply GPT-4 to assess faithfulness of generated texts as this method reflects human feedback (Feng et al. (2023)) and can account for the abstractiveness of generated answers.

3. Input Data

We test our approach on summarization task. Considering the scope of experimentation is low-resource languages we use the Ukrainian part of XL-SUM dataset.

The Ukrainian part of the XL-SUM dataset is a collection of more than 58,000 BBC news articles in Ukrainian, introduced by Hasan et al. (2021)[1]. It is used as a training resource for summarization in Ukrainian and is considered a benchmark for comparison and evaluation in related studies. No human evaluation was provided for the Ukrainian language, as the authors focus mainly on the top 10 spoken languages. The data is used to train language model. However, due to the lack of computational resources we use only the first 10k examples to fine-tune the model, first 3K of test split as a test set and the rest of the test split (around 2.6K articles) as validation set for the alignment as described in the following chapter.

4. Experimental Setup

4.1. Large Language Model

Since the introduction of ChatGPT to public use, LLMs models became popular not only among researchers and data scientists for particular applications but also to the general public that accelerated development of LLMs. One of the first open-sourced models released was Llama from Meta. We use Llama-2 as a language model for the set-up. Llama 2 is a freely available large language model that has been trained on 2 trillion tokens from public online sources. They include also Wikipedia dumps from the June-August 2022 period part of which is in Ukrainian. The model thus may be applied to texts in Ukrainian, however, Meta researchers warn they do not run tests of Llama with languages other than English. It is available in sizes of 7B, 13B, and 70B parameters. We use the 13B version in the paper.

The set-up for our approach foresees the following steps depicted on Fig. 2:

1. Fine-tune Llama-2 model on training data.

2. Generate summaries using fine-tuned Llama-2 model on validation set.

3. Corrupt generated summaries by adding information not given in input text.

[1] Downloaded from `https://huggingface.co/datasets/csebuetnlp/xlsum`

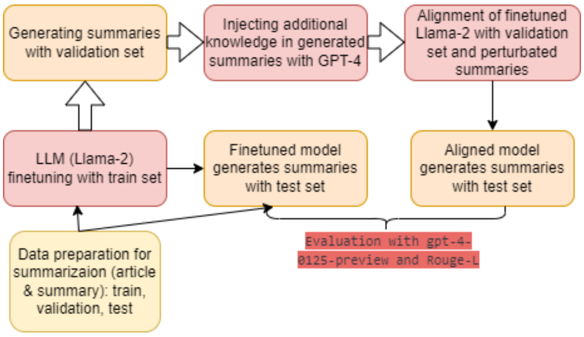

Generating summaries with validation set
Injecting additional knowledge in generated summaries with GPT-4
Alignment of finetuned Llama-2 with validation set and perturbated summaries
LLM (Llama-2) finetuning with train set
Finetuned model generates summaries with test set
Aligned model generates summaries with test set
Data preparation for summarizaion (article & summary): train, validation, test
Evaluation with gpt-4-0125-preview and Rouge-L

Figure 2: Illustration of the proposed approach.

4. Align fine-tuned Llama-2 with golden summaries to choose and noisy synthetic text from Step 3 to be rejected.

5. Apply both fine-tuned and aligned versions on test set.

6. Assess level of faithfulness hallucinations in generated texts using GPT-4 and Rouge-L, and human evaluation on a small subset.

4.2. LLM Finetuning

We use open-source Python packages for LLM finetuning using Lora adapters for faster training (transformers, trl, perf). The following training arguments ensure the results of the paper: learning-rate=2e-4, warm-up ratio = 0.03, maximum number of tokens = 512, truncate otherwise, 5 epochs. Lora perf arguments: rank = 32, lora-alpha=16, dropout = 0.1. As mentioned in Section 3.1., the first 10k of XL-Sum train split's articles has been used for finetuning. We used A100 40G GPU in the experiment. The training uses the prompt format:

Article to summarize in 26 words delimited with triple backticks: Article : “‘{article}“‘, Summary : “‘{summary}“‘.

4.3. Alignment with data perturbation

After finetuning the model generates summaries for 1239 articles out of the validation set that the LLM has not seen during training. These 1239 are chosen with the following logic: the average length of the golden summary is 26 words. We want to make sure that during alignment model does not prefer golden summaries because they are shorter than generated. For this, we adjust the training prompt format for inference. But more importantly we filter out rows with golden summaries of less than 20 words. We find 1239 articles after filtering from initial almost 2.6K set.

The generated summaries are further corrupted with added noise from GPT-4. Here is an algorithm applied: we extract named entities from the generated summaries using the Spacy NER model for Ukrainian and pass the first occurred entity together with generated text as an input to GPT-4 model asking the latter to enrich the text with information on the entity.

Prompt used for data corruption: *Instruction: You are a newspaper editor with much of encyclopedic knowledge. You have an entity and a text in Ukrainian. Then please insert in the phrase information of up to 4 words about the entity. Context: the text: {text }, entity: {entity }. Input: Your answer shall contain this text in Ukrainian enriched with your information in Ukrainian. Please add information about the entity as mentioned in the instruction.* . For example, for the following text (translated in English): Title "Mural: from Philadelphia to Rabat", article: "Since several years on Kyiv multi-storey buildings are emerging. . . " and golden summary: "While for Kyiv the rock art phenomenon is relatively new, in the West - . . . " the finetuned Llama model generates: "In Kyiv, street art is quickly expanding, said mayor Klitchko.". Corrupted sample is: "In Kyiv, street art is quickly expanding, said mayor Klitchko, a former boxer".

We used DPO for model alignment with the following parameters: learning-rate = 2e-6, beta = 0.5, batch = 2. Beta is relatively high to use the model knowledge.

5. Evaluation and Results

Recall from Section 2 that we build on Feng et al. (2023) approach to use one LLM model to evaluate the results of another. The following prompt is the input of GPT-4 model that shall define which summary contains irrelevant information:

Verify if summary is not consistent with the corresponding article. Provide the answer "Yes" if consistent or "No" if not consistent. The article: {article}; the summary: {summary}

The results of GPT-4 evaluation together with Rouge-L score are given in the Table 1. GPT-4 metric contains a percentage of texts found without hallucinations due to GPT-4. We can observe an increase of both Rouge-L and GPT-verified evaluation scores after alignment with synthetically generated texts with added noise. Apart from GPT-4 classification we randomly sampled 50 articles from the test set and asked human annotators to check for entity embellishments in summaries generated by finetuned and alighned LLama-2 versions presented in the paper. The rule for annotation is the following: if at least one embellishment found, label the article as 1, else 0. Out of 50 summaries produced by fine-tuned LLM, 11 contained faithfulness problems; out of 50 summaries produced by aligned LLM, only 6 contained entity embellishment.

Metric	Finetuned	Aligned
Rouge-L	23.4	29.7
GPT-4	72.1	81.5

Table 1: Results on test dataset with 3K news articles for finetuned model vs finetuned&aligned model with synthetic data corrupted with entities information (II)

The reduction in entity hallucinations is quite significant in case of human check but the sample is too small to be used as a proxy for all test data. Based on the results we may claim that our approach to alignment input data is experimentally tested.

Having obtained positive results to attain our objective, we shall recognize limitations of our study: 1. Bigger test set might have shown more accurate results. 2. Experiment with other language could prove coherence of our set-up. 3. Automatic evaluation with LLM model may imbibe issues and biases of evaluating model and might be not always correct. Rouge-L score has many limits (see Section2). 4. Human evaluation of bigger sample would show more accurate evaluation of results. 5. Experimenting with more prompts and Llama-specific syntax could deliver improvements. Thus, we foresee using the same algorithm with more data in Ukrainian and make comparison with other languages in future research to avoid stochastic biases.

We release the following versions of the Llama-2 model on HuggingFace Hub as described in the paper:

* finetuned model [2];

* aligned with noisy synthetic data [3].

HuggingFace dataset hub also contains the test data with golden and corrupted synthetic summaries [4].

6. Bibliographical References

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2023. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. *arXiv preprint arXiv:2311.10081*.

Sehyun Choi, Tianqing Fang, Zhaowei Wang, and Yangqiu Song. 2023. Kcts: knowledge-constrained tree search decoding with token-level hallucination detection. *arXiv preprint arXiv:2310.09044*.

Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. *arXiv preprint arXiv:2104.08455*.

Huawen Feng, Yan Fan, Xiong Liu, Ting-En Lin, Zekun Yao, Yuchuan Wu, Fei Huang, Yongbin Li, and Qianli Ma. 2023. Improving factual consistency of text summarization by adversarially decoupling comprehension and embellishment abilities of llms. *arXiv preprint arXiv:2310.19347*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Ziwei Ji, YU Tiezheng, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating llm hallucination via self reflection. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Haoqiang Kang, Terra Blevins, and Luke Zettlemoyer. 2024. Comparing hallucination detection metrics for multilingual generation. *arXiv preprint arXiv:2402.10496*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Mateusz Lango and Ondřej Dušek. 2023. Critic-driven decoding for mitigating hallucinations in data-to-text generation. *arXiv preprint arXiv:2310.16964*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

[2] https://huggingface.co/SGaleshchuk/Llama-2-13b-hf_uk_rank-32_ft

[3] https://huggingface.co/SGaleshchuk/Llama-2-13b-sum_ukr_dpo

[4] https://huggingface.co/datasets/SGaleshchuk/XL_SUM_ukr_synthetic_hallucinations

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. *arXiv preprint arXiv:2305.14552*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Kyle Wiggers. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2023. Mitigating language model hallucination with interactive question-knowledge alignment. *arXiv preprint arXiv:2305.13669*.