Entity Embelishment Mitigation in LLMs Output with Noisy Synthetic Dataset

Svitlana Galeshchuk
Arval BNPP/West Ukraine National University

Introduction

Text generation produces text based on an input

Applications include summarization, chatbots, storytelling, and machine translation

Large language models have advanced fluency and diversity of text

However, they are prone to creating factually incorrect, inconsistent, or irrelevant information

Hallucinations can pose ethical risks and loss of trust

Two types: factuality and faithfulness hallucination

Future research is crucial to enhance quality and accuracy of generated text

Objective

Focus on faithfulness problem and context inconsistencies in LLM generated output

- LLM output may be imprecise or untrue compared to user input

Context hallucinations accompanying named entities referred to as entity embellishment

- Paper aims to reduce risk of context hallucination and entity embellishment in foundation models

Example of entity hallucination in Figure 1

- LLM adds information on nationalities of Tesla and Mercedes not mentioned in article

Solution: Use of summarization dataset and perturbated examples for model alignment via DPO procedure

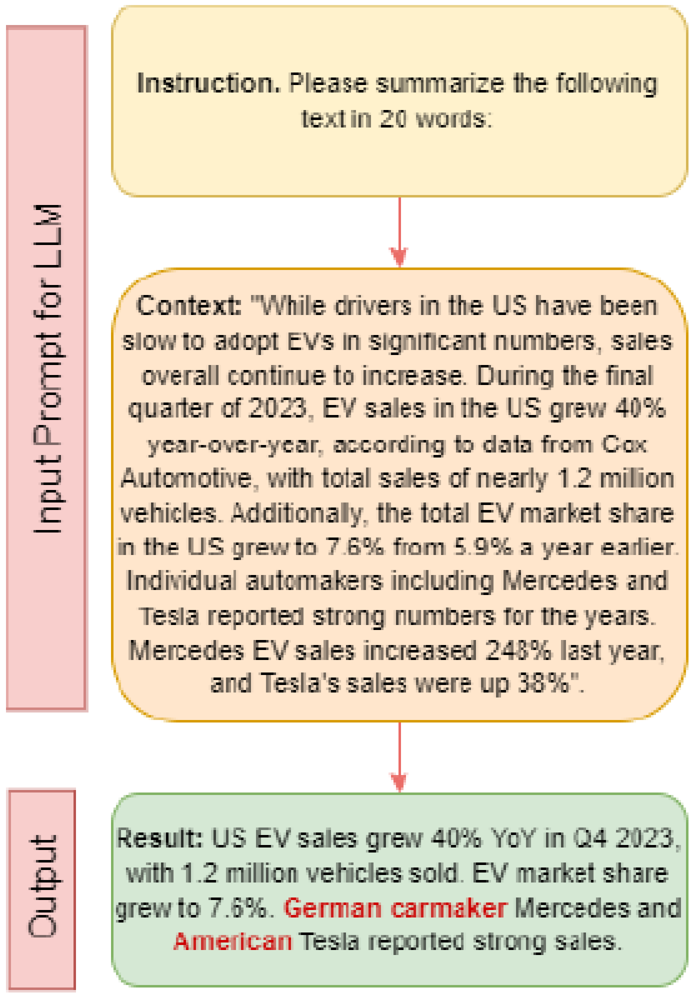

Fig.1. Example of an entity embelishment

Metrics for Hallucination (State-of-the-art)

N-gram based metrics like ROUGE

Calculates ratio of token overlap between generated output and correct answer

Poor correlation with humans, limited usage

Feedback from another LLM

GPT-4 used to collect sentence-level factual consistency annotation for system-generated summaries

High correlation with human annotations

Weakly supervised approaches

Creation of dataset by corrupting golden summaries with paraphrasing, entity swapping, and noise injection

Used as input to LLM alignment phase

Input Data

- Ukrainian part of XL-SUM dataset used for testing
  - Collection of more than 58,000 BBC news articles in Ukrainian
  - Considered a benchmark for comparison and evaluation
- First 10k examples used to fine-tune the model
  - First 3K of test split used as test set
  - Rest of test split used as validation set for alignment

*AI generated image

Experimental Setup

Large Language Model: Llama-2 from Meta

- Trained on 2 trillion tokens from public online sources
- Available in sizes of 7B, 13B, and 70B parameters
- 13B version used in the paper

Set-up Steps:

- Fine-tune Llama-2 model on training data
- Generate summaries using fine-tuned Llama-2 model on validation set
- Corrupt generated summaries by adding information not given in input text
- Align fine-tuned Llama-2 with golden summaries to choose and reject noisy synthetic text
- Apply both fine-tuned and aligned versions on test set
- Assess level of faithfulness hallucinations in generated texts using GPT-4 and Rouge-L, and human evaluation on a small subset

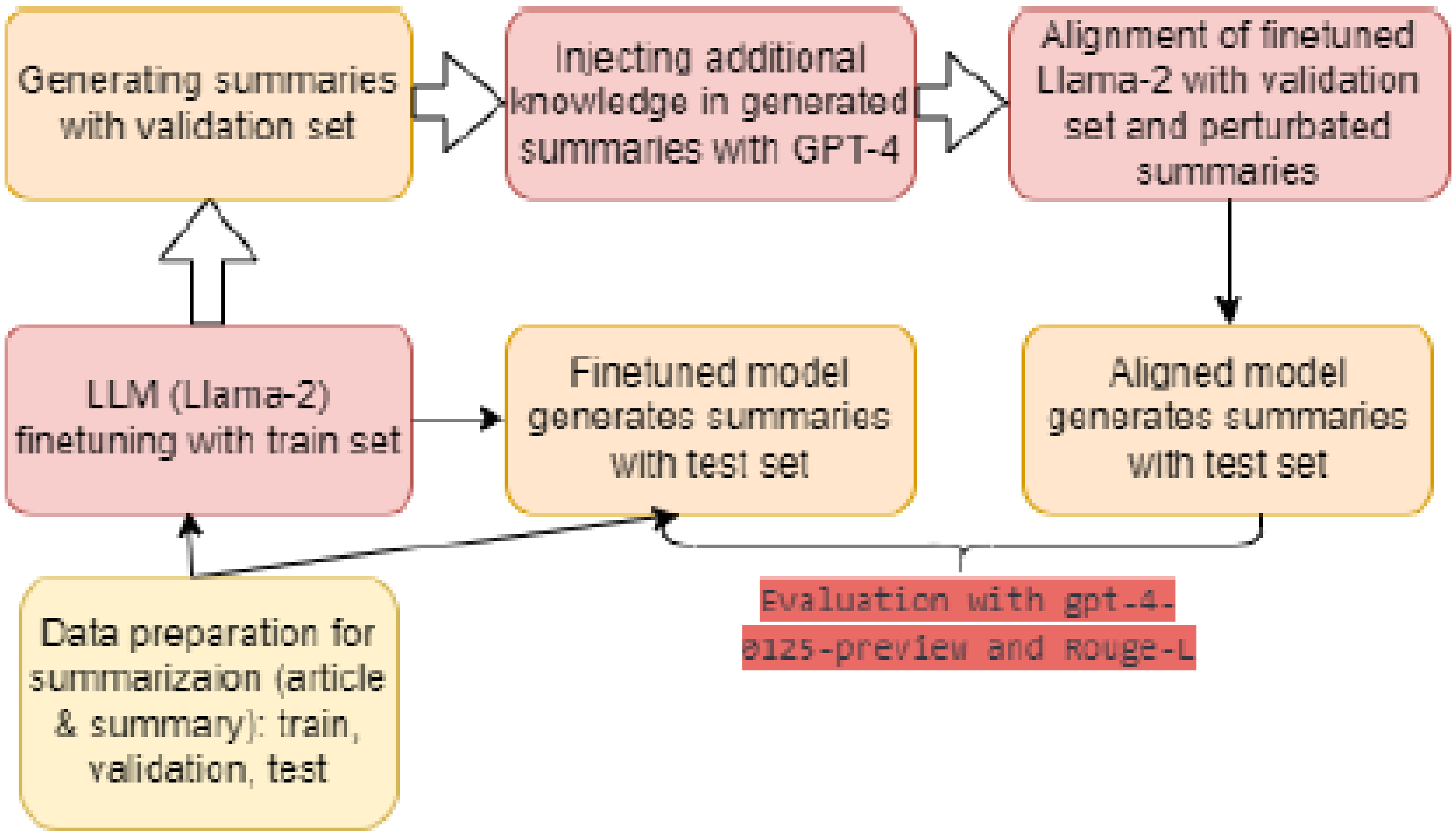

Figure 2: Illustration of the proposed approach.

Alignment with Data Perturbation

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 a text (translated in English) the golden summary is: "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".

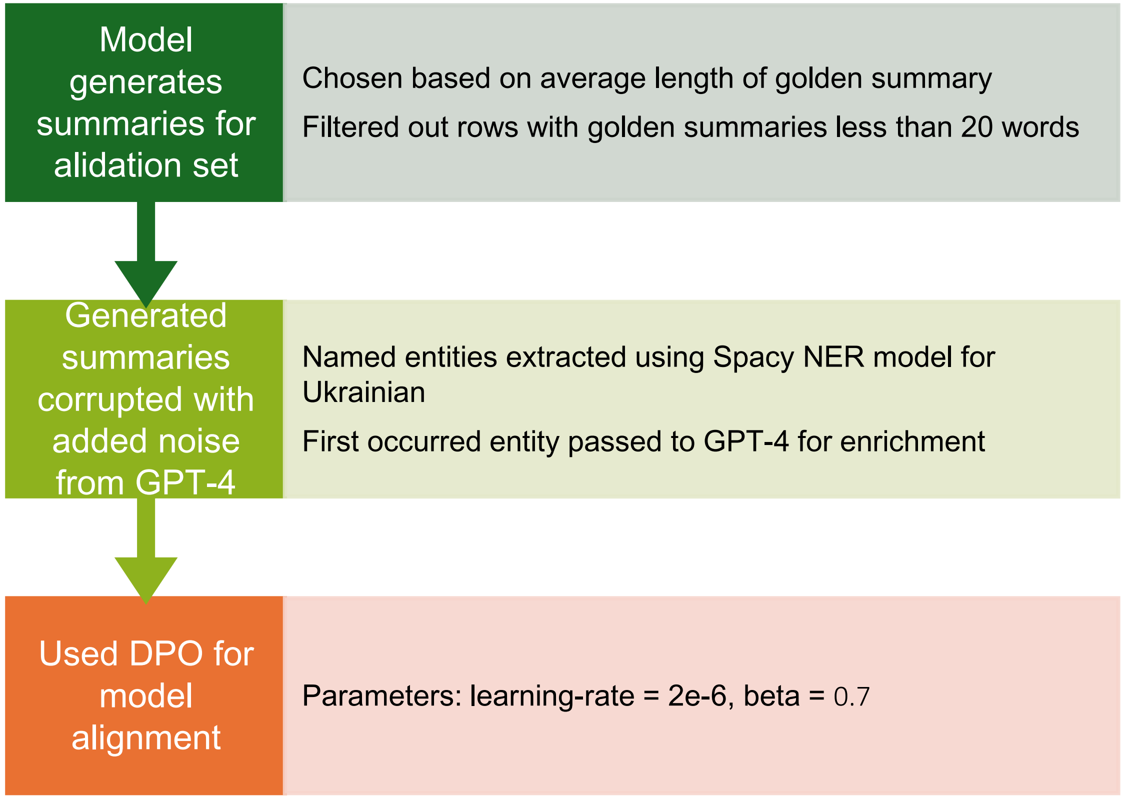

Fig.3. Zoom on data pertrurbation

Evaluation and Results

- LLM model evaluation approach based on Feng et al. (2023)
  - Using GPT-4 to evaluate summary consistency with article
- Results show increase in Rouge-L and GPT-verified evaluation scores after alignment with synthetically generated texts
  - Random sampling of 50 articles showed reduction in entity embellishment in aligned LLM model

User Prompt: 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

Table 1. Results on test set

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

Where to find

- Finetuned version: SGaleshchuk/Llama-2-13b-hf_uk_rank-32_ft at main (huggingface.co)
- Aligned version: SGaleshchuk/Llama-2-13b-summarization_uk_dpo · Hugging Face

Limitations

Test Set Size

- Bigger test set might have shown more accurate results

Language Experimentation

- Experiment with other language could prove coherence of our set-up

Automatic Evaluation

- Automatic evaluation with LLM model may imbibe issues and biases of evaluating model and might not always be correct

Rouge-L Score

- Rouge-L score has many limits

Human Evaluation

- Human evaluation of bigger sample would show more accurate evaluation of results

Experimentation

- Experimenting with more prompts and Llama-specific syntax could deliver improvements