Evaluating Factuality in LLM-Based Clinical Text Simplification

Abstract

The goal of my project was to investigate the factuality in the plain language summaries of clinical abstracts generated by open-sourced large language models. In this pursuit, I explored various prompts for clinical text simplification, compared the performance of open-sourced vs proprietary LLMs, used different automatic metrics, and set the foundation for an annotation study and future work with factuality models to compare for correlation with human evaluation.

Prior Work

Prior work in this space includes literature of GPT-3 on Summarizing and Synthesizing medical text (Shaib et al.). In our work, we take the 100 abstracts from the single document dataset of this paper and use them as inputs formatted with the prompts for the four models (GPT-4, LLAMA2, ALPACA, and Flan-T5-XL). Furthermore, prior work includes studying the lack of factuality in open-sourced LLMs compared to proprietary models (Gudibande et al.). We investigate this through comparing the factuality of outputs of open-sourced LLMs (LLAMA2, ALPACA, and Flan-T5-XL) vs. propietary LLMs (GPT-4). Also, prior work includes using ChatGPT as a factuality evaluator (Luo et al.). We employ the ChatGPT-DA method from this paper to automatically evaluate the factuality of each generation with a score from 1-100 (given by ChatGPT).

Research Process

In the process, I generated over 2000 different model generations across many open-sourced LLMs (REDPajama, FALCON, LLAMA, ALPACA, Flan-T5-XL, FlanALPACA) and GPT-3.5 and GPT-4 with various different prompts. I learned how to run pipelines of inference for model generations efficiently and deal with compute limitations. I also learned about the nuances of the styles of different

	ChatGPT-DA	Flesh Kincaid Grade Level	Length (Spacy Tokenizer)
GPT-4 Paper Plain	85.93	9.155	216.77
GPT-4 5th Grade	89.9	10.606	308.73
GPT-4 Summarize	84.9	9.583	183.71
GPT-4 Make it Short	85.205	8.814	111.28
FlanT5-XL Paper Plain	87.53	14.741	47.43
FlanT5-XL 5th Grade	81.44	15.031	28.24
LLAMA2 Paper Plain	81.03	8.218	135.38
ALPACA Complex Passage	88.41	13.308	113.21
ALPACA Medical Abstract	89.277	13.523	101.37
Abstract/Baseline	N/A	11.879	293.74

large language models, like the extractiveness of Flan-T5-XL, chattiness of LLAMA-2 Chat, and the specific format to prompt ALPACA. Furthermore, I evaluated the text generations with different automatic text summarization scores (BERTScore, SARI, ROUGE, BLEU), readability/grade-level scores (Flesch-Kincaid Grade Level), Length (Spacy Tokenizer), and ChatGPT (ChatGPT-DA, ChatGPT-Star, ChatGPT-CoT, ChatGPT-Binary).

Results

Our results include a curated spreadsheet of 400 generations (across 4 model/prompt combinations). I ultimately chose GPT-4 with the Paper Plain prompt modified with Summarize and Make it Short, Flan-T5-XL with the Paper Plain prompt, LLAMA2 with the Paper Plain prompt, and ALPACA with the Complex Passage prompt. All 4 combinations score above 80 on the ChatGPT-DA evaluation (averaged over 100 abstracts). This data is ready to be piloted into the human annotation evaluation for factuality.

Future Work

I am continuing this work with Professor Jessy Li in the Fall to the next steps of this project towards a publication at ACL. The next steps include running an annotation study for the 400 generations, across undergraduate Linguistics and medical students. Furthermore, these generations will provide a foundation for the next step of factuality models and defining factuality through PICO (Population, Intervention, Comparator, and Outcome) elements. We hope to finish the annotation study for factuality and flesh out the novel approach of PICO factuality with LLMs on these generations, and then compare the human annotations vs LLM-based factuality evaluation for correlations.We will then work towards finalizing a writeup to submit to ACL 2024.

References

Gudibande, Arnav, et al. “The False Promise of Imitating Proprietary Llms.” *arXiv.Org*, 25 May 2023, arxiv.org/abs/2305.15717.

Luo, Zheheng, et al. “Chatgpt as a Factual Inconsistency Evaluator for Text Summarization.” *arXiv.Org*, 13 Apr. 2023, arxiv.org/abs/2303.15621.

Shaib, Chantal, et al. “Summarizing, Simplifying, and Synthesizing Medical Evidence Using GPT-3 (with Varying Success).” *arXiv.Org*, 11 May 2023, arxiv.org/abs/2305.06299.