Understanding Hallucinations in LLMs: A Graph-Based Reproduction Study

Introduction

Large Language Models (LLMs) have recently surged in popularity and have a wide range of applications in different tasks. However, even though the outcomes of LLMs are fast, relatively reliable, and human-like, they may generate hallucinations in outcomes, which raises concern about the trustworthiness of the information these models provide. The research [1] proposes a graph-based architecture for hallucination detection. In our project, we aim to evaluate the generalizability of the hallucination detection proposed in the research paper “Leveraging Graph Structures to Detect Hallucinations in Large Language Models” by applying it to a new dataset and LLM.

Background

Under the first premise that LLM hallucinations are structured, previous work like SelfCheckGPT [2] uses a sampling-based approach that facilitates fact-checking in a zero-resource fashion to mitigate hallucinations, which is an implicit approach to model consistencies. The new research aims to do it explicitly by exploring semantic correspondences between hallucinations in the latent space. The second premise indicates that entities with similar characteristics are more likely to connect with each other, the new research studies if the degree of hallucination is such a characteristic.

There are numerous alternative approaches for hallucination detection. The Fact Extraction and Verification Shared Task [3] is a retrieval-based hallucination detection approach. Another approach [4] tries to detect hallucinations by training a discriminator on the RelQA LLM-generated question-answering dialogue dataset. Compared with previous work, the new method has advantages in that it does not need access to external knowledge nor to the LLM, avoid biases, and eliminates costs.

Literature Review

The paper "Leveraging Graph Structures to Detect Hallucinations in Large Language Models" [1] introduces a graph-based architecture for hallucination detection in Large Language Models (LLMs). This relational structure was chosen based on the semantic-related nature of the three targeted problems: (1) the existence of shared characteristics of hallucinations, (2) the ability to learn these characteristics, and (3) the feasibility of applying this knowledge to new generations of LLMs.

Training datasets were reformatted from a QA dataset [5] by prompting an LLM to generate sentences that were either hallucinated or not. Two benchmark datasets were also reformatted from labeled fact-checking datasets, SelfCheckGPT [2], and FEVER [3].

The corpus of LLM generations is modeled using a graph where each node represents a data point, with features being the sentence-level embeddings. An edge is inserted between two nodes if their similarity exceeds a certain empirically chosen threshold. This threshold is crucial as it must balance the node degree distribution and graph connectivity and has been shown to be dataset-dependent. A graph attention network is applied to aggregate information from neighboring nodes through message passing. To further enhance performance, additional contrastive learning was included. The inference of output labels is formulated as a regression task to indicate the level of hallucination.

The effectiveness of the graph-based architecture was demonstrated by comparing its performance against a three-level multi-layer perceptron (MLP) and a larger pre-trained language model, DeBERTa [6]. The improvements brought about by contrastive learning were also verified through an ablation study. The enhanced performance supported the conclusion of the first two research problems that hallucinations share common characteristics and these characteristics are learnable. The model's superior performance when generalized to the other two benchmark datasets addressed the third problem, affirming that inference to new LLM generations is practical.

However, the experiments were conducted using only one LLM (Llama2 [7]), and there may be distinctions between different LLMs. Evaluating

and comparing across multiple LLMs could contribute to the generalizability of the findings. Another way to enhance this would be by experimenting with more datasets in different domains, especially since only three datasets were evaluated and the threshold for similarity is dataset-dependent [1].

Experiment Plan

Our objective is to further assess the generalizability of the proposed hallucination detection framework by evaluating its effectiveness across various contexts and model architectures. Depending on the available time, we will choose at least one of the labeled datasets — Climate-FEVER [8], PUBHEALTH [9], or WICE [10] and an open-source LLM Qwen [11], to implement their training and evaluation pipeline. We will compare the results from the following experiments with the original ones:

Experiment	Dataset	LLM
1	FEVER [3] (original)	Qwen [11] (Newly Selected)
2	TBD (Newly Selected)	LLama2 [7] (original)

For each experiment, we will adhere to the established procedure to prompt the LLM to generate four types of statements based on the selected dataset. The preprocessing stage will involve extracting sentences and generating embeddings using the BERT model, resulting in fixed-size vector representations for each sentence. Following this, we will construct a graph where each node represents a sentence, and edges are established based on cosine similarity, using an empirically determined threshold (e.g., $\tau = 0.85$) for connectivity.
To effectively categorize the sentences, we will encode their labels using an ordinal scheme. The dataset will be divided into training (70%), validation (15%), and test (15%) sets to ensure balanced representation.
This comprehensive approach will allow us to evaluate the robustness of the graph-based hallucination detection framework across a variety of LLMs and contexts.

Reference

[1] N. Nonkes, S. Agaronian, E. Kanoulas, and R. Petcu, "Leveraging Graph Structures to Detect Hallucinations in Large Language Models," in Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing, Bangkok, Thailand, Aug. 2024, pp. 93-104. [Online]. Available: https://aclanthology.org/2024.textgraphs-1.7

[2] Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. Preprint, arXiv:2303.08896.

[3] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A Large-Scale Dataset for Fact Extraction and Verification," in NAACL-HLT, 2018.

[4] Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. 2023. Hallucination detection: Robustly discerning reliable answers in large language models. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM '23, page 245–255, New York, NY, USA. Association for Computing Machinery.

[5] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "MS MARCO: A Human-Generated Machine Reading Comprehension Dataset," in CEUR Workshop Proceedings, vol. 1773, CEUR-WS, 2016.

[6] P. He, X. Liu, J. Gao, and W. Chen, "DEBERTA: Decoding Enhanced BERT with Disentangled Attention," in ICLR 2021 - 9th International Conference on Learning Representations, International Conference on Learning Representations (ICLR), 2021.

[7] H. Touvron, L. Martin, et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," Preprint, arXiv:2307.09288, 2023.

[8] T. Diggelmann, J. Boyd-Graber, J. Bulian, M. Ciaramita, and M. Leippold, "CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims," arXiv:2012.00614 [cs], Jan. 2021, Available: https://arxiv.org/abs/2012.00614

[9] "Papers with Code - PUBHEALTH Dataset," Paperswithcode.com, 2022. https://paperswithcode.com/dataset/pubhealth (accessed Oct. 18, 2024).

[10] R. Kamoi, T. Goyal, J. D. Rodriguez, and G. Durrett, "WiCE: Real-World Entailment for Claims in Wikipedia," arXiv.org, 2023. https://arxiv.org/abs/2303.01432 (accessed Oct. 18, 2024).

[11] J. Bai et al., "Qwen Technical Report," arXiv.org, Sep. 28, 2023. https://arxiv.org/abs/2309.16609