‘Do-It-Yourself’ Tactics for Large Language Models: A Survey on Self-Hallucination-Management Mechanisms

Jeeseun Baik, Sounman Hong, Kyungho Yoon, Won-Yong Shin

Yonsei University

jsbpa73@yonsei.ac.kr, sounman_hong@yonsei.ac.kr, yoonkh@yonsei.ac.kr, wy.shin@yonsei.ac.kr

Abstract

This survey explores the 'self-hallucination-management' mechanism in large language lodels (LLMs), a self-regulatory approach that addresses the challenge of 'hallucination'—the generation of misleading or nonsensical content despite seeming fluency. Departing from traditional methods that depend on extensive data and computation, this novel mechanism functions in a zero-resource black-box setting without additional training or databases. They introduce internal feedback loops for LLMs to self-detect and correct hallucinations, enhancing credibility. We introduce several mechanisms, which exemplify the application of this concept, illustrating a shift towards more autonomous and resilient LLMs. The paper discusses new directions for self-hallucination-management strategies and evaluation metrics for further research direction.

Ⅰ. Introduction

As with the accelerating progress of large language models (LLMs), a growing concern for the issue of ‘hallucination’ may hinder their credibility and availability. Although LLM’s hallucinated responses have considerable significance to be dealt with, the current perception of hallucination in LLMs still remains unresolved, with only few approaches existing to address the challenge. We propose a novel mechanism, ‘self-hallucination-management’, which relies solely on the LLM itself to administer hallucination in a zero-resource black-box setting.

Ⅱ. Definition

In the field of natural language processing (NLP), tracing back its origin from the field of psychology, the term ‘hallucination’ is referred to as a phenomenon where generated content appears to be articulate and natural in spite of being unfaithful or nonsensical to the provided source [1]. Prior studies in NLP categorize hallucination into two main types: 1) intrinsic hallucination and 2) extrinsic hallucination, depending upon whether the generated output can be verified through the provided source content. However, in the era of LLMs, the existing typology appears to have certain limitations due to the characteristics of LLMs-versatility, user-centric interactions, connectivity with facticity, etc. Therefore, we introduce a more fine-grained categorization based on the fundamental work by [2]. The details are elaborated below:

▪ **Factuality hallucination,** where LLMs generate content inconsistent with the real-world knowledge;

▪ **Input-faithfulness hallucination,** where LLMs generate content deviating from the source input;

▪ **Context-faithfulness hallucination,** where LLMs generate content inconsistent with its formerly generated content;

▪ **Reasoning-faithfulness hallucination,** where LLMs generate content that logically contradicts its reasoning process or the final output.

Ⅲ. Self-Hallucination-Management Mechanism

Numerous research efforts have been conducted to detect, correct, and mitigate the hallucination in LLMs, indicating ‘hallucination management’. Conventional hallucination management mechanisms rely mainly on external knowledge sources or model refinement strategies according to the LLMs' life cycle, which require access to massive data, considerable computational costs, and additional human creative processes. In order to surmount the limitations, there arises a need for the concept of 'self-hallucination-management', a mechanism for LLMs to address the hallucination utilizing internal feedback iteratively. Certain conditions for the establishment of the concept are as follows: ① Only **a single model** is used for the mechanism. ② A **black-box approach** is applied for the mechanism. ③ **No additional training or external database** is required.

Several recent studies have employed the mechanism.

SelfCheckGPT [2] is a sampling-based approach for hallucination detection, which compares multiple stochastically-sampled responses and calculates the hallucination score by measuring the consistency between them. When the sampled responses are divergent and contradictory, which means 'inconsistent', the response is more likely to be hallucinated.

SELF-FAMILIARTIY [3] introduced another black-box model with a zero resource setting to pre-detect and mitigate the hallucination in LLMs. The main concept of input instruction is extracted and an LLM is prompted to generate explanations for the concept. According to the explanation, the model is asked to recreate the original concept, and the probability score of a response sequence serves as a familiarity score, thus operating as a yardstick to judge hallucination.

Self-Detection Method [4], another sampling-based approach, suggested the notion of 'self-inconsistency', referring to a phenomenon where LLMs provide divergent or contradictory responses to semantically analogous questions-that is, hallucinated-, which indicates the non-factuality of the model. Once a set of paraphrased questions is created, the method examines the inconsistency score between the corresponding answers, leading to detect hallucination.

IV. Open Problems and Future Directions

Reasoning and self-Hallucination-management The self-hallucination-management mechanism is intimately related to reasoning-faithfulness hallucination. Combined with the area of natural language reasoning, new prospects for the mechanism can be open.

What is new standard for hallucination detection? Prior studies focus primarily on the consistency of instructions, queries, or responses as a standard for self-hallucination-management. However, since the consistency-based method is not able to determine the hallucination between semantically similar but incorrect cases, the need for alternative standards emerges as a necessity.

Is hallucination a self-adversarial attack for LLMs? A few recent research has suggested to view hallucination as a kind of adversarial attack. Inducing LLMs to be hallucinated intentionally can lead to comprehensive defense and evaluation of LLMs in the process of self-management.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C3004345, No. RS-2023-00220762).

REFERENCES

[1] Filippova, K. (2020). Controlled Hallucinations: Learning to Generate Faithfully from Noisy Data. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 864-870).

[2] Manakul, P., Liusie, A., & Gales, M. J. (2023). Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. arXiv preprint arXiv:2303.08896.

[3] Luo, J., Xiao, C., & Ma, F. (2023). Zero-resource hallucination prevention for large language models. arXiv preprint arXiv:2309.02654.

[4] Zhao, Y., Yan, L., Sun, W., Xing, G., Meng, C., Wang, S., Cheng, Z., Ren, Z., & Yin, D. (2023). Knowing What LLMs DO NOT Know: A Simple Yet Effective Self-Detection Method. arXiv preprint arXiv:2310.17918.