Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models

Yue Zhang[♠*], **Yafu Li**[♢], **Leyang Cui**[♡†], **Deng Cai**[♡], **Lemao Liu**[♡]
Tingchen Fu[☆], **Xinting Huang**[♡], **Enbo Zhao**[♡], **Yu Zhang**[♠], **Yulong Chen**[♢]
Longyue Wang[♡], **Anh Tuan Luu**[☼], **Wei Bi**[♡], **Freda Shi**[☾], **Shuming Shi**[♡]

[♡] Tencent AI lab [♠] Soochow University [♢]Zhejiang University
[☆] Renmin University of China [☼] Nanyang Technological University
[☾] Toyota Technological Institute at Chicago

Abstract

While large language models (LLMs) have demonstrated remarkable capabilities across a range of downstream tasks, a significant concern revolves around their propensity to exhibit *hallucinations*: LLMs occasionally generate content that diverges from the user input, contradicts previously generated context, or misaligns with established world knowledge. This phenomenon poses a substantial challenge to the reliability of LLMs in real-world scenarios. In this paper, we survey recent efforts on the detection, explanation, and mitigation of hallucination, with an emphasis on the unique challenges posed by LLMs. We present taxonomies of the LLM hallucination phenomena and evaluation benchmarks, analyze existing approaches aiming at mitigating LLM hallucination, and discuss potential directions for future research.

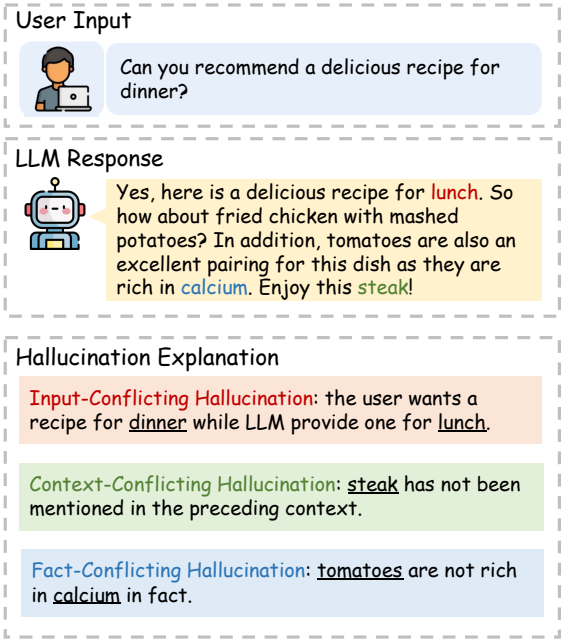

Figure 1: Three types of hallucinations occurred in LLM responses (best viewed in color).

1 Introduction

Large language models (LLMs), particularly characterized by their substantial number of parameters, have arisen as a promising cornerstone for the development of natural language processing (NLP) and artificial intelligence (Zhao et al., 2023c). With proper alignment techniques, such as supervised finetuning (SFT; Zhang et al., 2023b) and reinforcement learning from human feedback (RLHF; Ouyang et al., 2022; Fernandes et al., 2023), recent LLMs (OpenAI, 2023a; Touvron et al., 2023b; OpenAI, 2023b, *inter alia*) have exhibited strong capabilities in solving various downstream tasks.

Nonetheless, as exemplified in Figure 1, LLMs, despite their remarkable success, occasionally produce outputs that, while seemingly plausible, deviate from user input (Adlakha et al., 2023), previously generated context (Liu et al., 2022), or factual knowledge (Min et al., 2023; Muhlgay et al., 2023; Li et al., 2023a)—this phenomenon is commonly referred to as hallucination, which significantly undermines the reliability of LLMs in real-world scenarios (Kaddour et al., 2023). For instance, LLMs can potentially fabricate erroneous medical diagnoses or treatment plans that lead to tangible real-life risks (Umapathi et al., 2023).

While hallucination in conventional natural language generation (NLG) settings has been widely studied (Ji et al., 2023), understanding and addressing the hallucination problem within the realm of LLMs encounters unique challenges introduced by

1. **Massive training data**: in contrast to carefully curating data for a specific task, LLM pre-

* This survey paper was completed during Yue Zhang (yzhang21@stu.suda.edu.cn), Yafu Li, Tingchen Fu, and Yu Zhang's internships at Tencent AI Lab.
† Corresponding author (leyangcui@tencent.com).