面向大语言模型事实幻觉的可解释多模型修正方法研究

Research on Explainable Multi-Model Correction Methods for Hallucinations in Large Language Models

黄世彬[1]，张准[*]
[1]华南师范大学 光电科学与工程学院
[*] zhhzun@scnu.edu.cn

【摘要】 大语言模型在教育与专业领域具有广泛应用潜力，但事实性幻觉限制了其在高可靠场景中的应用。针对现有方法多孤立处理幻觉检测或校正且缺乏可解释评估的问题，本文构建带有噪声注入与路径标注的事实幻觉数据集，以追踪幻觉生成过程。在此基础上提出可解释的多模型检测与校正框架，通过多模型协同识别幻觉类型并生成修正结果。同时引入基于逆向推理的“黄金校正路径”，用于指导校正并作为评估参考。实验表明，该方法教育知识生成任务中能够提升文本事实一致性与修正过程的可解释性。

【关键词】 大语言模型；事实性幻觉；检测与校正；可解释框架；知识生成

Abstract: *Large language models have broad application potential in education and professional fields, but factual illusions limit their use in high-reliability scenarios. Addressing the issue that existing methods often handle illusion detection or correction in isolation and lack interpretable evaluation, this paper constructs a factual illusion dataset with noise injection and path annotation to trace the illusion generation process. Based on this, an interpretable multi-model detection and correction framework is proposed, which identifies illusion types and generates correction results through multi-model collaboration. Simultaneously, a "golden correction path" based on inverse reasoning is introduced to guide correction and serve as an evaluation reference. Experiments show that this method can enhance textual factual consistency and the interpretability of the correction process in educational knowledge generation tasks.*

Keywords: LLM, Factual Hallucination, Detection and Correction, Interpretable Framework, knowledge generation

1. 前言

近年来，大语言模型在自然语言生成与知识应用场景中取得显著进展，但其生成内容中仍频繁出现与真实事实不一致的事实性幻觉，在教育与科研等高可靠性场景中可能降低系统可信度并误导用户。现有研究主要通过幻觉检测、检索增强或自我反思等方法进行缓解，但多数工作仅关注输出结果正确性，缺乏对幻觉产生过程及其推理偏移机制的系统建模。

2. 论文贡献

1. 构建可解释事实幻觉数据集 FactHMT，通过噪声传播路径标注刻画幻觉生成机制。
2. 提出黄金修正路径，通过逆向推理获得标准化修正路径并支持路径级评估。
3. 设计多模型自适应融合校正框架，实现针对不同幻觉类型的可解释协同修正。

3. TPH-ECF 系统框架

本文提出的整体系统框架称为 **TPH-ECF**，该框架主要包括四个阶段：数据构建、幻觉检测与分类、多模型协同修正以及结果评估。框架图如下图 1 所示。

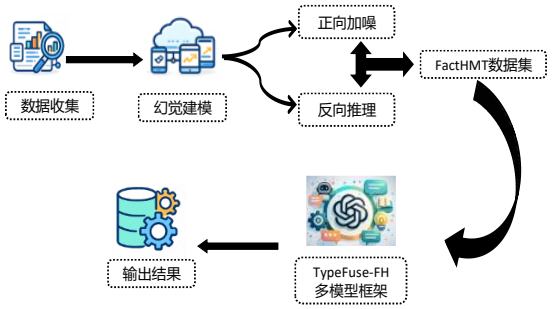

图 1 基于事实性幻觉诊断与修正 TPH-ECF 系统总框架

3.1. 可解释事实幻觉数据集

本文基于 Wikipedia 等权威知识源构建多领域基础事实样本，并在推理链关键阶段注入规则化噪声以模拟大模型生成中的事实偏差，同时标注噪声在推理过程中的传播路径，构建可追溯的事实幻觉数据集 FactHMT（现已达 7000 条幻觉建模数据）。根据幻觉表现形式，本文将事实性幻觉划分为三类：事实约束违背、实体建模错误和关系推理错误。

3.2. 黄金修正路径

本文提出黄金修正路径，通过对噪声传播路径进行逆向推理，从错误结果逐步回溯至正确事实链，形成标准化修正步骤序列。该路径既可用于指导幻觉修正，也可作为评估模型修正过程合理性的路径级参考（Chen et al., 2024）。

3.4. 多模型协同修正

在幻觉类型识别后，本文提出 TypeFuse-FH 多模型自适应融合框架。系统并行调用不同策略模型，通过不同策略模型对不同幻觉类型有不同贡献值，依此进行权重划分，加权融合生成最终修正结果，同时生成修正理由与修改路径，以实现可解释的协同修正。

3.5. 实验设计与评估维度

基于上述框架，本文构建了完整的实验流程。对于每个输入样本，系统输出包括：幻觉存在性判定（Fact-ACC）、幻觉类型识别（FC-ACC）、修正路径匹配度（Path Match）以及修正后的事实准确率（Accuracy）。本文选取主流大模型作为基线，对所提出方法进行初步验证。实验结果表明，所提出框架在事实准确率与路径匹配度等指标上均优于基线模型，显示出更强的幻觉识别与修正能力。同时，不同模型融合结果在同一幻觉类型下表现出更高的一致性，表明多模型协同机制能够有效提升校正过程的稳定性与可靠性。上述结果验证了该方法在事实幻觉检测与可解释修正任务中的有效性。

4. 结论

本文从推理链视角将大语言模型的事实性幻觉建模为可溯源的推理偏移过程，并构建带有噪声路径标注的事实幻觉数据集。在此基础上提出“黄金修正路径”和面向幻觉类型的多模型自适应融合校正框架，实现可解释的协同修正。实验结果表明，该方法能够提升生成内容的事实一致性与修正稳定性。进一步将该框架集成至学习资源生成系统，实现生成、检测与校正一体化应用。未来将扩展数据规模并优化权重策略，以提升多领域场景下的泛化能力。

参考文献

Chen, X., Song, D., Gui, H., Wang, C., Zhang, N., Jiang, Y., Huang, F., Lv, C., Zhang, D., & Chen, H. (2024). FactCHD: Benchmarking fact-conflicting hallucination detection. In arXiv preprint arXiv:2310.12086v3.