Editing the Mind of Giants: An In-Depth Exploration of Pitfalls of Knowledge Editing in Large Language Models

Cheng-Hsun Hsueh* **Paul Kuo-Ming Huang*** **Tzu-Han Lin*** **Che-Wei Liao***
Hung-Chieh Fang* **Chao-Wei Huang** **Yun-Nung Chen**
National Taiwan University, Taipei, Taiwan
{r12922059,b08902072,r12944034,r09922a25}@csie.ntu.edu.tw
{b09902106,f07922069}@csie.ntu.edu.tw y.v.chen@ieee.org

Abstract

Knowledge editing is a rising technique for efficiently updating factual knowledge in large language models (LLMs) with minimal alteration of parameters. However, recent studies have identified side effects, such as knowledge distortion and the deterioration of general abilities, that have emerged after editing. Despite these findings, evaluating the pitfalls of knowledge editing often relies on inconsistent metrics and benchmarks, lacking a uniform standard. In response, this survey presents a comprehensive study of these side effects, providing a unified perspective on the challenges of knowledge editing in LLMs by conducting experiments with consistent metrics and benchmarks. Additionally, we review related works and outline potential research directions to address these limitations. Our survey highlights the limitations of current knowledge editing methods, emphasizing the need for a deeper understanding of the inner knowledge structures of LLMs and improved knowledge editing methods. To foster future research, we have released the complementary materials publicly[1].

1 Introduction

Recent advancements in large language models (LLMs) have significantly improved NLP applications, enabling LLMs to understand and generate language at a human-like level. However, the mechanisms of knowledge storage in LLMs remain unclear, raising concerns about the reliability of their output, particularly in applications like chatbots. To address these issues, researchers have explored various methods. Traditional methods like fine-tuning, continual learning, and retraining are computationally expensive and may degrade LLM performance. *Knowledge editing* has emerged as a promising alternative, offering efficient adjustments with minimal computational costs and fewer alterations (Cao et al., 2021; Dai et al., 2022; Meng et al., 2022, 2023; Dong et al., 2022; Mitchell et al., 2022a,b; Hartvigsen et al., 2023; Huang et al., 2023; Yu et al., 2024; Zheng et al., 2023; Li et al., 2023; Tan et al., 2024; Gupta et al., 2024b; Wang et al., 2024). This method allows precise LLMs refinement, enhancing their practical and reliable use in real-world applications.

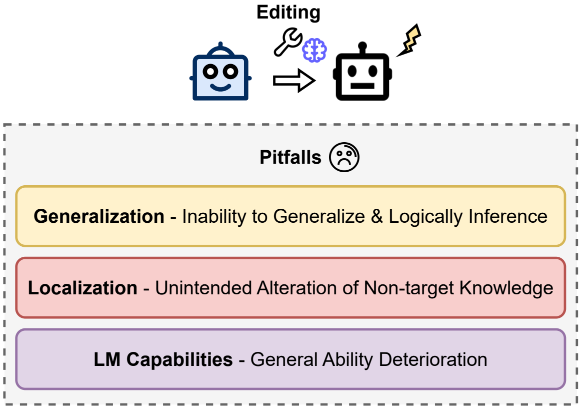
Editing
Pitfalls
Generalization - Inability to Generalize & Logically Inference
Localization - Unintended Alteration of Non-target Knowledge
LM Capabilities - General Ability Deterioration

Figure 1: An overview of pitfalls in current knowledge editing methods. The subsequent sections dive into three key challenges: generalization issues (Section 3.1), locality issues (Section 3.2), and deterioration of general LLM abilities (Section 3.3).

Knowledge editing can be divided into two main categories: parameter-modifying and parameter-preserving. Both aim to refine LLM knowledge efficiently while avoiding the drawbacks of previous tuning methods (Yao et al., 2023). Parameter-modifying methods, including meta-learning (Cao et al., 2021; Mitchell et al., 2022a; Tan et al., 2024) and locate-and-edit techniques (Dai et al., 2022; Meng et al., 2022, 2023; Li et al., 2023; Gupta et al., 2024b), strive to update model parameters effectively. By contrast, parameter-preserving methods introduce external components, like knowledge bases (Mitchell et al., 2022b; Zhong et al., 2023) or extra model parameters (Dong et al., 2022; Huang et al., 2023; Hartvigsen et al., 2023; Yu et al., 2024)

[1] https://github.com/MiuLab/EditLLM-Survey
*Equal contribution.