to maintain the integrity of pre-trained LLMs while updating their knowledge.

Despite the success of knowledge editing, challenges remain. Recent studies have revealed side effects that can harm the general capabilities and intrinsic structures of LLMs. We categorize these pitfalls into three main areas: (1) the inability to perform logical inference (Cohen et al., 2023; Li et al., 2024; Zhong et al., 2023; Hua et al., 2024; Yao et al., 2023), (2) the unintended modification of non-target knowledge (Cohen et al., 2023; Li et al., 2024; Yao et al., 2023; Meng et al., 2022; Hoelscher-Obermaier et al., 2023), and (3) the deterioration of general LLM abilities (Gupta et al., 2024a; Gu et al., 2024; Yang et al., 2024). Although various side effects have been identified, they are evaluated using inconsistent metrics and benchmarks in different studies, which lack a uniform standard. As a result, this survey aims to provide a comprehensive overview of the current issues in the knowledge editing paradigm and to establish a fair platform for comparing the side effects of different editing methods. Additionally, we encourage further investigation into the pitfalls and underlying knowledge structures of LLMs. A brief overview of the discussed pitfalls is shown in Figure 1.

This paper is organized as follows: Section 2 introduces the definition and methods of knowledge editing. Section 3 discusses current challenges and corresponding benchmarks. In Section 4, we present experimental results evaluating different editing methods. Finally, Section 5 explores related studies and future research directions. We summarize our contributions as follows:

1. We are the first to provide a comprehensive analysis of the side effects associated with existing knowledge editing techniques.
2. We systematically organize previous research and conduct experiments to benchmark the side effects of knowledge editing, providing a unified perspective on this issue.
3. We discuss related studies and potential directions to address existing challenges, encouraging further exploration in this field.

2 Overview of Knowledge Editing

2.1 Problem Definition

Knowledge editing for LLMs entails modifying the output of LLMs in response to specific edit queries, with the aim of minimizing alterations to their original behavior (Yao et al., 2023; Mazzia et al., 2023; Zhang et al., 2024a). In this section, we follow the notation from Mazzia et al. (2023).

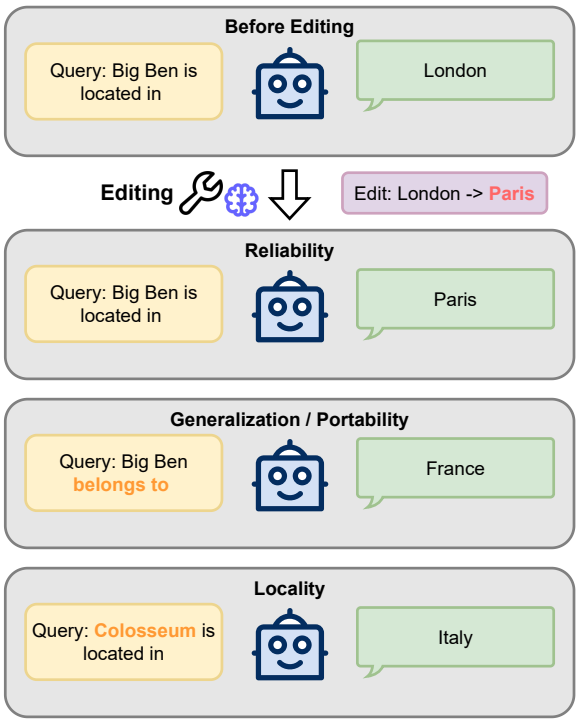

Figure 2: Illustration of properties that knowledge editing methods should satisfy: reliability, generalizability/portability, and locality.

We denote the input and output space as $\mathbb{X}$ and $\mathbb{Y}$, respectively. The function space $\mathbb{F} : \mathbb{X} \to \mathbb{Y}$ is estimated by the base model f_{θ_0} parameterized by $\theta_0 \in \Theta$. Finally, let $Z_e = \{(x_e, y_e) \mid f_{\theta_0}(x_e) \neq y_e\}$ be the set of edit queries we would like to apply to the base model. The goal of knowledge editing is to efficiently derive the edited model f_{θ_e} from the base model that satisfies the following:

$$f_{\theta_e}(x_e) = y_e, \forall (x_e, y_e) \in Z_e \tag{1}$$

The ideal edited model f_{θ_e} should satisfy three properties: **reliability**, **generalization**, and **locality**. An illustration is shown in Figure 2.

Reliability Given an edit query (x_e, y_e), the edited model f_{θ_e} should output the target answer y_e when given the target input x_e, i.e. $f_{\theta_e}(x_e) = y_e$. The reliability of a editing method is measured by calculating the average edit success rate:

$$\mathbb{E}_{(x'_e, y'_e) \sim Z_e} \mathbb{1}\{f_{\theta_e}(x'_e) = y'_e\} \tag{2}$$

Generalization The edited model should generalize the edited knowledge to relevant instances.