Unintended edits to unrelated facts may occur because a single edit can implicitly change the predictive distribution among objects associated with the same (*subject* - *relation*) pair. After multiple consecutive edits, these alterations can accumulate and distort the stored knowledge. To evaluate this condition, the concept of Knowledge Distortion has been introduced by Li et al. (2024), which estimates the Jensen–Shannon divergence of the object set distribution before and after editing. This can be further extended to metrics such as the Ignore Rate, measuring how objects other than the target in the object set are neglected after editing, and the Failure Rate, which measures the proportion of instances where over half of the objects in the set are overlooked.

3.3 Deterioration of General LLM Abilities

Current evaluation metrics are primarily limited to scenarios where editing is performed only once or infrequently, prompting some studies to extend evaluations to the outcomes after consecutive edits. A study by Gupta et al. (2024a) discovers that post-edit models exhibit susceptibility to both gradual forgetting and catastrophic forgetting in sequential editing scenarios. Notably, their findings indicate that the extent of knowledge forgetting is more pronounced in meta-learning-based methods compared to locate-and-edit methods. Additionally, models with parameters modified successively show a decline in performance across various downstream NLP tasks (Gu et al., 2024). Furthermore, perplexity is found to increase after consecutive edits across all parameter-modified methods and different LLMs, and is proposed as another metric to indicate model collapse (Yang et al., 2024). These findings further corroborate that model editing aimed at modifying parameters adversely affects the general capabilities of the original LLMs.

4 Experiments

The experiments are done to evaluate robust generalization and locality (Section 4.1.1 as well as deterioration of general LLM abilities (Section 4.1.2 across different editing methods.

4.1 Experimental Setup

Given the variety of benchmarks addressing different challenges in knowledge editing, systematically comparing model performance becomes difficult. To address this, we select the most widely used datasets for each category of pitfalls, ensuring a fair and transparent comparison.

4.1.1 Robust Generalization and Locality

We use GPT-J (Wang and Komatsuzaki, 2021) as the baseline model for editing and implement six distinct editing methodologies to assess robust generalization and locality: MEND (meta-learning), ROME and MEMIT (locate-and-edit), SERAC (external memory), and IKE (prompting).

Given the overlap in benchmarks for robust generalization and locality, we select a subset for our experiments. The evaluation is divided into two settings: *single edit*, where only one fact in a reasoning chain is modified, and *multiple edits*, where several logically connected facts in the chain are altered simultaneously. A detailed description is provided in the Appendix A). Single edit metrics include Subject-Replace, Reverse-Replace, and One-Hop reasoning (Yao et al., 2023). Multiple edit metrics include multi-hop editing accuracy (Zhong et al., 2023), and Conflict Score and Conflict Magnitude for Reverse Conflict and Composite Conflict respectively (Li et al., 2024). For locality, single edit metrics include Other-Attribution, Distract-Neighbor, and Other-Task (Yao et al., 2023), while multiple edit metrics encompass Success Rate, Distortion, Ignore Rate, and Failure Rate (Li et al., 2024).

4.1.2 Deterioration of General LLM Abilities

Following the settings of (Gu et al., 2024), we assess deterioration of general LLM abilities post-editing using six methodologies: ROME, MEMIT, SERAC, MEND, KN, and GRACE. We evaluate general abilities across four NLP downstream tasks: open-domain question answering, sentiment analysis, reasoning, and summarization. These tasks are assessed after 10 to 40 edits on the Zero-Shot Relation Extraction (ZsRE) dataset(Levy et al., 2017), comparing the results against pre-editing benchmarks. More details on the selected downstream tasks are in Appendix B.

4.2 Experimental Results and Discussion

In general, current editing methodologies show suboptimal performance in both robust generalization and locality. Regarding robust generalization (Table 3), IKE, which leverages prompt demonstrations, excels in single edit but declines with multiple edits. This suggests that prompt demonstrations may become confused when editing multiple log-