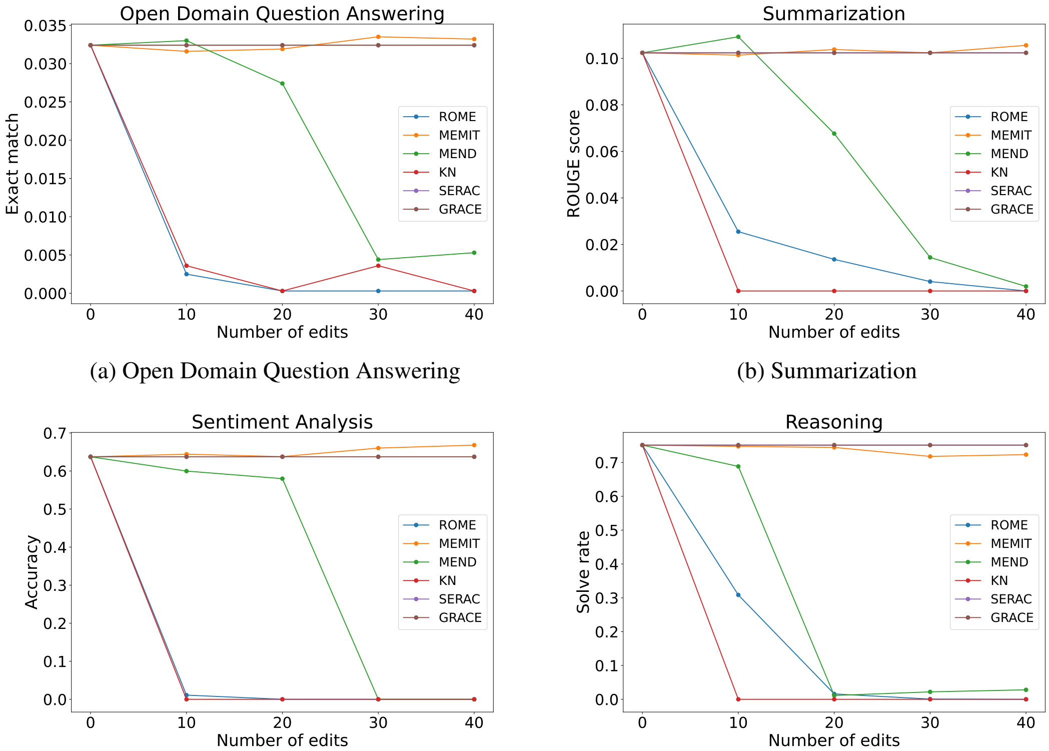

(a) Open Domain Question Answering

(b) Summarization

(c) Sentiment Analysis

(d) Reasoning

Figure 4: The experimental results for the deterioration of general abilities were obtained by editing GPT-J with various editing algorithms, including ROME, MEMIT, MEND, KN, SERAC, and GRACE, each applied 10 to 40 times. The edited models were subsequently evaluated on four downstream tasks, including open-domain question answering, sentiment analysis, summarization, and reasoning. The results for SERAC and GRACE are overlapping.

5 Future Prospects

5.1 Leveraging Information Retrieval and External Memory

Research shows that using external knowledge bases, rather than relying solely on internal knowledge, benefits LLMs by guiding content generation based on predefined facts. External knowledge sources, such as text corpora, structured tables, or key-value databases, can be utilized either to finetune LLMs for improved information retrieval or to employ prompting techniques for querying these sources. These approaches separate factual knowledge from inference process, thus preserves the original model parameters and minimizes post-editing damage. Moreover, they ensure that generated content aligns with predefined knowledge bases, thereby enhancing accountability and accuracy.

5.2 Improving Understandings of LLMs' Internal Knowledge Structures

While identifying where factual knowledge is stored in LLMs has been extensively explored (Meng et al., 2022, 2023; Dai et al., 2022; Hernandez et al., 2024; Geva et al., 2021), the correlation between knowledge location and editing success remains low (Hase et al., 2023). Additionally, despite evidence suggesting a strong connection between factual knowledge and the feed-forward network layers (Meng et al., 2022; Geva et al., 2021, 2022), recent findings indicate that updates to multi-head self-attention layers also improve outcomes (Li et al., 2023). This suggests that locating fact storage alone doesn't fully explain knowledge structures in LLMs. Further research is needed to understand how knowledge locations interact with model predictions in order to enhance LLM interpretability and controllability.

Preserving LLMs' general capabilities is also crucial for model editing, as discussed in Section 3.3. Recent breakthroughs in identifying regions within models that correlate with general linguistic abilities have opened up a direction for future research in model editing (Zhang et al., 2024b). By making targeted modifications, we can potentially prevent the deterioration of general abilities