and improve the specificity and effectiveness of model editing methods.

5.3 Improving Robustness of Knowledge Editing

Even after achieving fair scores on the existing metrics, models may revert to pre-edit versions or provide ambiguous answers if the altered knowledge is conflicted with inherited concepts. Experiments show that more popular knowledge is easier for modified models to revert to (Ma et al., 2024), indicating the lack of robustness in current editing strategies. A deeper understanding of how LLMs store and process interconnected knowledge entities is crucial for more robust editing and warrants future research.

6 Conclusion

Although model editing techniques appear promising for cost-effectively updating knowledge, they still have significant pitfalls. Current editing methods often struggle with making logical inferences based on the edited facts, introducing unintended alterations of non-target knowledge and deterioration in model performance, particularly with parameter-modified methods. By harnessing information retrieval techniques and delving into how models store and process knowledge, deviations in model abilities can be mitigated, and the controllability of edited facts can be enhanced, ultimately leading to greater robustness. We hope our work illuminates potential directions for future improvements in knowledge editing.

Limitations

The field of knowledge editing is advancing at an impressive pace, with numerous innovations in editing methodologies and evaluation metrics being proposed. Despite our efforts to collect and organize previous work, some contributions may not be included in this paper. However, we will continue to monitor the latest developments in this field and update our GitHub repository with recent related works.

Acknowledgments

We thank the reviewers for their insightful comments. This work was financially supported by the National Science and Technology Council (NSTC) in Taiwan, under Grants 111-2222-E-002-013-MY3 and 112-2223-E002-012-MY5. We thank to National Center for High-performance Computing (NCHC) of National Applied Research Laboratories (NARLabs) in Taiwan for providing computational and storage resources.

References

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Conference on Empirical Methods in Natural Language Processing*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022. Calibrating factual knowledge in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5937–5947, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.