7 Ethical considerations

Although we test a language model's ability to serve as a knowledge base, we do not find these models to be a reliable source of knowledge, and we caution readers that a LLM should not be used as an authoritative source of facts. Our memory-editing methods shed light on the internal mechanisms of models and potentially reduce the cost and energy needed to fix errors in a model, but the same methods might also enable a malicious actor to insert false or damaging information into a model that was not originally present in the training data.

8 Acknowledgements.

Thanks to Jaden Fiotto-Kaufmann for building the demonstration at memit.baulab.us. This project was supported by an AI Alignment grant from Open Philanthropy. YB was also supported by the Israel Science Foundation (grant No. 448/20) and an Azrieli Foundation Early Career Faculty Fellowship.

9 Reproducibility

The code and data for our methods and experiments are available at memit.baulab.info.

All experiments are run on workstations with NVIDIA A6000 GPUs. The language models are loaded using HuggingFace Transformers (Wolf et al., 2019), and PyTorch (Paszke et al., 2019) is used for executing the model editing algorithms on GPUs.

GPT-J experiments fit into one 48GB A6000, but GPT-NeoX runs require at least two: one 48GB GPU for running the model in `float16`, and another slightly smaller GPU for executing the editing method. Due to the size of these language models, our experiments will not run on GPUs with less memory.

References

Oshin Agarwal and Ani Nenkova. Temporal effects on pre-trained models for language processing tasks. *Transactions of the Association for Computational Linguistics*, 10:904–921, 2022.

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031*, 2022.

James A Anderson. A simple neural network generating an interactive memory. *Mathematical biosciences*, 14(3-4):197–220, 1972.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pp. 722–735. Springer, 2007.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL `https://doi.org/10.5281/zenodo.5297715`.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. Gpt-neox-20b: An open-source autoregressive language model, 2022.

Kurt Bollacker, Robert Cook, and Patrick Tufts. Freebase: A shared database of structured general human knowledge. In *AAAI*, volume 7, pp. 1962–1963, 2007.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4762–4779, 2019.