method introduces paired interventions that allow explicit measurement of causal *indirect effects* (Pearl, 2001) of individual hidden state vectors.

Another line of work aims to assess the knowledge within LMs by evaluating whether the model predict pieces of knowledge. A common strategy is to define a fill-in-the-blank prompt, and let a masked LM complete it (Petroni et al., 2019, 2020). Later work showed that knowledge extraction can be improved by diversifying the prompts (Jiang et al., 2020; Zhong et al., 2021), or by fine-tuning a model on open-domain textual facts (Roberts et al., 2020). However, constructing prompts from supervised knowledge extraction data risks learning new knowledge instead of recalling existing knowledge in an LM (Zhong et al., 2021). More recently, Elazar et al. (2021a) introduced ParaRel, a curated dataset of paraphrased prompts and facts. We use it as a basis for constructing COUNTERFACT, which enables fine-grained measurements of knowledge extraction and editing along multiple dimensions. Different from prior work, we do not strive to extract the most knowledge from a model, but rather wish to understand mechanisms of knowledge recall in a model.

Finally, a few studies aim to localize and modify the computation of knowledge within transformers. Geva et al. (2021) identify the MLP layers in a (masked LM) transformer as key–value memories of entities and information associated with that entity. Building on this finding, Dai et al. (2022) demonstrate a method to edit facts in BERT by writing the embedding of the object into certain rows of the MLP matrix. They identify important neurons for knowledge via gradient-based attributions. De Cao et al. (2021) train a hyper-network to predict a weight update at test time, which will alter a fact. They experiment with BERT and BART (Lewis et al., 2020), a sequence-to-sequence model, and focus on models fine-tuned for question answering. Mitchell et al. (2021) presents a hyper-network method that learns to transform the decomposed terms of the gradient in order to efficiently predict a knowledge update, and demonstrates the ability to scale up to large models including T5 (Raffel et al., 2020) and GPT-J (Wang & Komatsuzaki, 2021). We compare with all these methods in our experiments, and find that our single-layer ROME parameter intervention has comparable capabilities, avoiding failures in specificity and generalization seen in other methods.

5 Conclusion

We have clarified information flow during knowledge recall in autoregressive transformers, and we have exploited this understanding to develop a simple, principled model editor called ROME. Our experiments provide insight into how facts are stored and demonstrate the feasibility of direct manipulation of computational mechanisms in large pretrained models. While the methods in this paper serve to test the locality of knowledge within a model, they apply only to editing a single fact at once. Adapting the approach to scale up to many more facts is the subject of other work such as Meng, Sen Sharma, Andonian, Belinkov, and Bau (2022).

Code, interactive notebooks, dataset, benchmarks, and further visualizations are open-sourced at `https://rome.baulab.info`.

6 Ethical Considerations

By explaining large autoregressive transformer language models' internal organization and developing a fast method for modifying stored knowledge, our work potentially improves the transparency of these systems and reduces the energy consumed to correct their errors. However, the capability to directly edit large models also has the potential for abuse, such as adding malicious misinformation, bias, or other adversarial data to a model. Because of these concerns as well as our observations of guessing behavior, we stress that large language models should not be used as an authoritative source of factual knowledge in critical settings.

Acknowledgements

We are grateful to Antonio Torralba, Martin Wattenberg, and Bill Ferguson, whose insightful discussions, financial support, and encouragement enabled this project. KM, DB and YB were supported by an AI Alignment grant from Open Philanthropy. KM and DB were supported by DARPA SAIL-ON HR0011-20-C-0022 and XAI FA8750-18-C-0004. YB was supported by the ISRAEL SCIENCE FOUNDATION (grant No. 448/20) and an Azrieli Foundation Early Career Faculty Fellowship.