Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. 2024. Model editing can hurt general abilities of large language models. *arXiv preprint arXiv:2401.04700*.

Akshat Gupta, Dev Sajnani, and Gopala Anumanchipalli. 2024. A unified framework for model editing. *arXiv preprint arXiv:2403.14236*.

R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, volume 7, pages 785–794.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. *arXiv preprint arXiv:1706.04115*.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2023a. Pmet: Precise model editing in a transformer. *arXiv preprint arXiv:2308.08742*.

Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. 2023b. Unveiling the pitfalls of knowledge editing for large language models. *arXiv preprint arXiv:2310.02129*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2021. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022. Memory-based model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831. PMLR.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *Advances in neural information processing systems*, 34:11054–11070.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. *arXiv preprint arXiv:2305.13172*.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. Mquake: Assessing knowledge editing in language models via multi-hop questions. *arXiv preprint arXiv:2305.14795*.

A Appendix

A.1 zsRE Compatibility

Two popular datasets are used to evaluate model editing performance - zsRE (Levy et al., 2017) and CounterFact (Meng et al., 2022a). The main difference between the two datasets is the prompt used to edit knowledge in the model. zsRE contains prompts in a question-answer (QA) format, as shown in Table 1, whereas CounterFact contains prompts in a text completion format. Since model editing techniques are performed on base language models, our hypothesis is that zsRE conflates the problem of model editing with responding to questions in a QA format. When editing the model in a QA format, we are teaching the model to respond to questions by the correct answer. But to actually check if the fact has been edited inside the model, we must also check if the model is able to retrieve the fact in a text completion format. Otherwise all we've done is train a QA model and not edited knowledge. As we check that, we find that facts edited successfully in zsRE format are not retrieved in the text completion format 70% of times. Some failure examples are given below (we only show examples that were successfully edited using ROME in GPT2-XL):

- **zsRE Question:** The date of birth of Martha Neumark is?
- **Edited Answer:** 1904
- **Completion Prompt:** Martha Neumark was born on
- **Generated Answer:** Martha Neumark was born on April 15, 1869, in New York City.

- **zsRE Question:** The college Herb Pomeroy attended was what?
- **Edited Answer:** Harvard University
- **Completion Prompt:** Herb Pomeroy attended the college of
- **Generated Answer:** Herb Pomeroy attended the college of Oxford University

A.2 Model Editing Implementation Details

We use the default implementations of FT-C, ROME, MEND and MEMIT for GPT2-XL and GPT-J as used by the authors of Meng et al. (2022b) in `https://github.com/kmeng01/memit`. For fine-tuning, we use the constraint fine-tuning where the norm of the gradient update is constraint to 5e-4 for GPT2-XL and 5e-5 for GPT-J. These are the default hyperparameters used by the authors.

A.3 Downstream Evaluation Details

An important dimension to evaluate model editing, especially at scale, is to evaluate the performance of edited models on downstream performance. In this paper, we evaluate models on four tasks of the glue (Wang et al., 2018) benchmark - sentiment analysis (SST2) (Socher et al., 2013), paraphrase detection (MRPC) (Dolan and Brockett, 2005), natural language inference (NLI) (Dagan et al., 2005; Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009) and linguistic acceptability classification (Warstadt et al., 2019).

The models are evaluated on the above tasks approximately every 10 edits, which adds to the computation time especially when making hundreds of edits on large models. Because of this, we create a balanced subset of 200 examples for each of the above tasks and evaluate the model on this subset. The model performance is measured using the F1 metric.

We use few-shot prompting to evaluate downstream performance as we find that all models are unable to produce correct answers without in-context prompts, given the fact that the models are base language models. We follow the prompt template used by Perez et al. (2021) for our models. The exact prompts used for the different tasks are shown in Tables 4, 5, 6, 7.

Task	Few-Shot Prompt
SST-2	Review : excruciatingly unfunny and pitifully unromantic Sentiment : negative Review : rich veins of funny stuff in this movie Sentiment : positive Review : by far the worst movie of the year Sentiment : negative Review : fashioning an engrossing entertainment out Sentiment : positive Review : INPUT SENTENCE Sentiment :

Table 4: Few-shot template used to measure downstream model performance for the SST-2 task.