Vinay Venkatesh Ramasesh, Aitor Lewkowycz, and Ethan Dyer. 2022. Effect of scale on catastrophic forgetting in neural networks. In *International Conference on Learning Representations*.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Chenmien Tan, Ge Zhang, and Jie Fu. 2024. Massive editing for large language models via meta learning.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Tianxing Wu, Xudong Cao, Yipeng Zhu, Feiyue Wu, Tianling Gong, Yuxiang Wang, and Shenqi Jing. 2023. Asdkb: A chinese knowledge base for the early screening and diagnosis of autism spectrum disorder.

Tianxing Wu, Haofen Wang, Cheng Li, Guilin Qi, Xing Niu, Meng Wang, Lin Li, and Chaomin Shi. 2020. Knowledge graph construction from multiple online encyclopedias. *World Wide Web*, 23:2671–2698.

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.

Lang Yu, Qin Chen, Jie Zhou, and Liang He. 2023. Melo: Enhancing model editing with neuron-indexed dynamic lora.

Wenhao Yu, Chenguang Zhu, Zhihan Zhang, Shuohang Wang, Zhuosheng Zhang, Yuwei Fang, and Meng Jiang. 2022. Retrieval augmentation for commonsense reasoning: A unified approach. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4364–4377. Association for Computational Linguistics.

Ningyu Zhang, Xin Xie, Xiang Chen, Shumin Deng, Hongbin Ye, and Huajun Chen. 2022. Knowledge collaborative fine-tuning for low-resource knowledge graph completion. *Journal of Software*, 33(10):3531–3545.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023a. Can we edit factual knowledge by in-context learning?

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, Limao Xiong, Lu Chen, Zhiheng Xi, Nuo Xu, Wenbin Lai, Minghao Zhu, Cheng Chang, Zhangyue Yin, Rongxiang Weng, Wensen Cheng, Haoran Huang, Tianxiang Sun, Hang Yan, Tao Gui, Qi Zhang, Xipeng Qiu, and Xuanjing Huang. 2023b. Secrets of RLHF in large language models part I: PPO. *CoRR*, abs/2307.04964.

A Preliminaries of Model Editing

The task of knowledge editing is to effectively modify the initial base model f_θ to the edited model $f_{\theta'}$, with corresponding parameter adjustments for a specific input-output pair (x_e, y_e), where $x_e \in \mathcal{X}_e$ and $f_\theta(x_e) \neq y_e$. Here, $\mathcal{X}_e$ represents the entire set to be edited. Therefore, the current problem formulation for knowledge editing can be broadly categorized into three types:

1. **Single Instance Editing**: Evaluating the performance of the model after a single edit. The model reloads the original weights after a single edit:

$$\theta' \leftarrow \arg\min_{\theta}(\|f_\theta(x_e) - y_e\|) \tag{1}$$

2. **Batch Instance Editing**: Simultaneously modifying N knowledge instances (where $N \ll |\mathcal{X}_e|$) and evaluating the performance of the edited model after processing a batch. The model reloads the original weights after processing a batch of edits:

$$\theta' \leftarrow \arg\min_{\theta} \sum_{e=1}^{N}(\|f_\theta(x_e) - y_e\|) \tag{2}$$

3. **Sequential Editing**: This approach requires sequentially editing each knowledge instance, and evaluation must be performed after all knowledge updates have been applied:

$$\theta' \leftarrow \arg\min_{\theta} \sum_{e=1}^{|\mathcal{X}_e|}(\|f_\theta(x_e) - y_e\|) \tag{3}$$

B Default Hparams Settings

EASYEDIT provides optimal hyperparameters for various editing methods. In addition to common parameters such as learning rate, steps, and regularization coefficients, the location of layers for editing can also be considered as hyperparameters, significantly influencing the robustness of the editing process. The following tables demonstrate the default location settings in EASYEDIT (using **Llama-2-7B** as an example).

Layer with Value Loss
`model.layers.31`
Target Layer for Updating Weights
`model.layers.5.mlp.down_proj`

Table 3: Default Target Modules in **ROME**

Layer with Value Loss
`model.layers.31`
Target Layer for Updating Weights
`model.layers.4.mlp.down_proj`
`model.layers.5.mlp.down_proj`
`model.layers.6.mlp.down_proj`
`model.layers.7.mlp.down_proj`
`model.layers.8.mlp.down_proj`

Table 4: Default Target Modules in **MEMIT** and **PMET**

ROME We follow Meng et al. (2023) in utilizing causal mediation analysis to identify an intermediate layer in the model responsible for recalling facts. The causal traces reveal an early site (5th layer) with causal states concentrated at the last token of the subject, indicating a significant role for MLP states at that specific layer (Table 3).

MEMIT Following Meng et al. (2022), we quantify the average indirect causal effect of MLP modules. The results demonstrate a concentration of intermediate states in LLaMA. The disparity in the effects between MLP severed and hidden states severed becomes significantly reduced after the 8th layer. We choose the entire critical range of MLP layers, denoted as $\mathcal{R} = \{4, 5, 6, 7, 8\}$ (Table 4).

PMET PMET (Li et al., 2024) adopts the localization strategy from MEMIT, designating the corresponding layer as the modification target. Building upon the update of MLP weights, PMET focuses on multi-head self-attention (MHSA), further substantiating the discovery that MHSA encodes specific patterns for general knowledge extraction. (Table 4).

MEND In the context of meta-learning for editing, it is commonly observed that editing MLP layers yields better performance than editing attention

CodeBook Target Modules
model.layers[27].mlp.down_proj.weight

Table 5: Default Target Modules in **GRACE**

Target Layer for Updating Weights
model.layers.29.mlp.gate_proj.weight
model.layers.29.mlp.up_proj.weight
model.layers.29.mlp.down_proj.weight
model.layers.30.mlp.gate_proj.weight
model.layers.30.mlp.up_proj.weight
model.layers.30.mlp.down_proj.weight
model.layers.31.mlp.gate_proj.weight
model.layers.31.mlp.up_proj.weight
model.layers.31.mlp.down_proj.weight

Table 6: Default Target Modules in **MEND**

layers. Typically, MLP weights of the last 3 transformer blocks (totaling 6 weight matrices) are chosen for editing (Mitchell et al., 2022a). EASYEDIT adheres to this default configuration (Table 6).

GRACE Recent studies have revealed the impact of selecting the right layers for fine-tuning (Lee et al., 2023). Similarly, in GRACE (Hartvigsen et al., 2023), we conduct pilot experiments, retaining layers with consistently high edit success rates (Table 5).