Chenmien Tan, Ge Zhang, and Jie Fu. 2024. Massive editing for large language models via meta learning. In *International Conference on Learning Representations*.

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. 2023. DyLoRA: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3274–3287, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Yiwei Wang, Muhao Chen, Nanyun Peng, and Kai wei Chang. 2024. Deepedit: Knowledge editing as decoding with constraints. *ArXiv*, abs/2401.10471.

Wanli Yang, Fei Sun, Xinyu Ma, Xun Liu, Dawei Yin, and Xueqi Cheng. 2024. The butterfly effect of model editing: Few edits can trigger large language models collapse.

Yunzhi Yao, Peng Wang, Bo Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. In *Conference on Empirical Methods in Natural Language Processing*.

Lang Yu, Qin Chen, Jie Zhou, and Liang He. 2024. Melo: Enhancing model editing with neuron-indexed dynamic lora. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19449–19457.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, Xiaowei Zhu, Jun Zhou, and Huajun Chen. 2024a. A comprehensive study of knowledge editing for large language models.

Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024b. Unveiling linguistic regions in large language models. Association for Computational Linguistics.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

A Detailed Explanation of Evaluation Metrics and Examples

A.1 Portability / Generalization

Single Edit In the single edit scenario, we modify only one fact in the logical chain with each edit. Let $Z_e = \{(x_e, y_e) \mid f_{\theta_0}(x_e) \neq y_e\}$ be the set where only a single fact is edited in each logical chain. Single edit is conducted as:

$$f_{\theta_e}(x_e) = y_e, \forall (x_e, y_e) \in Z_e \tag{5}$$

This part consists of:

- One-Hop: This setting focuses on evaluating the impact of a single edit on direct, one-hop reasoning tasks.

For one-hop evaluations, we adopt the methods proposed by (Yao et al., 2023). These include:

- **Subject Replace:** This metric tests the model's generalization ability by replacing the subject in the question with an alias or synonym, assessing if the edited attribute is generalized to other descriptions of the same subject.
- **Reversed Relation:** This metric evaluates the model's capability to handle reversed relations by filtering for suitable relations (e.g., one-to-one relation) and asking the reverse question to check if the target entity is also updated.
- **One-Hop Test:** This metric assesses the edited language model's performance on downstream tasks that require one-hop reasoning.

Multiple Edits In the multiple edits scenario, we evaluate the model's performance after applying several logically related edits. Let $Z_e = \{(x_{ei}, y_{ei}) \mid f_{\theta_0}(x_{ei}) \neq y_{ei}\}$ represent a set of logically related facts within a reasoning chain intended to be modified. Multiple edits are performed by altering several facts within this chain:

$$f_{\theta_e}(x_{ei}) = y_{ei}, \forall (x_{ei}, y_{ei}) \in Z_e \tag{6}$$

This part consists of:

- Multi-Hop editing: Evaluate whether the model can infer edited knowledge in multi-hop questions.