Knowledge Editing for Large Language Models

Ningyu Zhang, Yunzhi Yao, Shumin Deng
Zhejiang University, National University of Singapore
zhangningyu@zju.edu.cn, yyztodd@zju.edu.cn, shumin@nus.edu.sg

Abstract

Even with their remarkable capabilities, Large Language Models (LLMs) like ChatGPT are not without challenges, particularly in maintaining factual accuracy and logical consistency. A primary concern is the ability to efficiently update these LLMs to rectify inaccuracies without undergoing comprehensive retraining or continuous training processes, which can be resource-intensive and time-consuming. The ability to edit LLMs presents a promising solution, allowing for modifications in specific areas of interest while preserving the model's overall performance across various tasks. This tutorial is designed to familiarize NLP researchers with the latest advancements and emerging techniques in editing LLMs. Our goal is to offer a thorough and up-to-date review of state-of-the-art methodologies, complemented by practical tools, and to highlight new avenues for research within the community. All referenced resources are available at https://github.com/zjunlp/KnowledgeEditingPapers.

Keywords: Knowledge Editing, Large Language Model

1. Introduction

Large Language Models (LLMs) have demonstrated impressive potential in generating text that closely resembles human writing, as evidenced by numerous studies. However, despite their advanced capabilities, models such as ChatGPT can sometimes struggle to maintain factual accuracy or logical coherence. There's also the risk of them generating content that could be considered harmful or offensive, compounded by their inability to recognize events occurring after their last training update. Addressing these issues without resorting to comprehensive retraining or ongoing training processes—both of which require substantial resources and time—presents a significant challenge. In response, the concept of **knowledge editing for LLMs** has emerged as a promising solution. This approach offers an efficient means to adjust the model's behavior in targeted areas without detrimentally affecting its performance across other tasks.

In this tutorial, our goal is to familiarize researchers with the latest advancements and emerging strategies in the realm of knowledge editing for LLMs. We aim to provide a systematic and comprehensive overview of state-of-the-art methods, enriched with practical tools, and to explore new avenues of research for our audience. The session will begin with an introduction to the tasks associated with knowledge editing for LLMs, alongside relevant evaluation metrics and benchmark datasets. We will then progress to discussing a range of knowledge editing methodologies, with a particular emphasis on those that maintain the original parameters of LLMs. These methods typically adjust the model's responses in specific instances by integrating an auxiliary network that works in tandem with the unmodified core model. The discussion will shift towards techniques that directly modify the parameters of LLMs, targeting the adjustment of model parameters linked to undesirable outputs. Throughout the tutorial, we aim to share insights from various research communities involved in knowledge editing, introduce open-source tools such as EasyEdit[1], and delve into both the challenges and opportunities presented by knowledge editing for LLMs. This session seeks to provide valuable knowledge to the community, underlining potential issues and uncovering prospects in the field of knowledge editing. The detailed schedule and content structure of the tutorial are outlined in the referenced schedule Table 1.

Our tutorial is grounded in the exploration of principles that guide the encapsulation of knowledge within pre-trained language models, drawing upon a range of pivotal studies such as those by Geva et al. (2021); Haviv et al. (2023); Hao et al. (2021); Hernandez et al. (2023b); Yao et al. (2023a); Cao et al. (2023b). These works provide foundational insights into how language models store and process information. The practice of knowledge editing, which includes the manipulation of a model's external knowledge, shares commonalities with knowledge augmentation techniques. This is because updating a model's stored knowledge essentially involves infusing it with new, relevant information. Additionally, we view knowledge editing as a nuanced form of lifelong learning (Biesialska et al., 2020) and unlearning (Wu et al., 2022; Tarun et al., 2021), where models are designed to dynamically incorporate and adjust new knowledge, while also shedding outdated or incorrect data. This approach is crucial for enhancing the model's relevance and accuracy over time. Moreover, by enabling models to discard harmful or toxic

[1] https://github.com/zjunlp/EasyEdit

LREC-COLING 2024 Tutorials, pages 33–41
20-25 May, 2024.

information, knowledge editing presents a viable strategy for addressing the security and privacy challenges that accompany the use of Large Language Models (Geva et al., 2022). In our tutorial, we will explore these dimensions in depth, offering insights into how knowledge editing contributes to the ongoing evolution of language models. We will also suggest possible future directions for research in this area. Attendees will find all related materials and slides available at https://github.com/zjunlp/KnowledgeEditingPapers, ensuring they have access to a comprehensive set of resources to further their understanding and application of knowledge editing techniques.

2. Target Audience

This tutorial is designed to appeal to a broad spectrum of participants, including academics like researchers and students, as well as industry professionals engaged in the fields of Natural Language Processing (NLP) and Artificial Intelligence (AI). It is structured to be accessible and informative for anyone with a basic understanding of NLP and AI principles. Furthermore, participants with a foundational knowledge of neural networks will find the content particularly advantageous. For those already familiar with LLMs and techniques for parameter-efficient tuning, this tutorial will significantly enrich their learning experience, providing deeper insights and practical applications in these areas.

3. Outline

The tutorial mainly consists of the following parts, as shown in Table 1.

1. Introduction (15 minutes)

- Background
- Why knowledge editing for LLMs?

2. Preliminaries (15 minutes)

- Pre-trained language models
- Definition of knowledge editing for LLMs
- Metrics and benchmark datasets

3. Knowledge Editing for LLMs

- Knowledge editing methods of preserving LLMs' parameters (40 minutes)

Coffee Break (30 minutes)

- Knowledge editing methods of modifying LLMs' Parameters (40 minutes)

4. Extensions (40 minutes)

- Knowledge editing for multilingual, multimodal LLMs
- Knowledge fairness, bias and security issues

5. Open-sourced Tools (30 minutes)

6. Discussion on Main Issues & Opportunities (30 minutes)

4. Suggested Duration

Half day (4 hours, including 30-minute break)

5. History

The presenters have organized the following tutorials:

- AACL 2023[2]: Editing Large Language Models (3-hour tutorial)
- IJCAI 2023[3]: Open-Environment Knowledge Graph Construction and Reasoning: Challenges, Approaches, and Opportunities (3-hour tutorial)
- AACL 2022[4]: Efficient and Robust Knowledge Graph Construction (3-hour tutorial)
- The 18th Reasoning Web Summer School[5]: Cross-Modal Knowledge Discovery, Inference, and Challenges (3-hour tutorial)

6. Diversity Considerations

The presenting team comprises individuals from two academic institutions, featuring a diverse mix of roles such as professors, a research fellow, and a Ph.D. candidate. Among the four speakers, one is a woman, highlighting the team's commitment to inclusivity and diversity in academic representation.

7. Estimated Number of Participants

LLMs are increasingly being applied across a wide array of tasks. Given the need for frequent post-training adjustments to correct errors and mitigate

[2]Resources will be available at https://github.com/zjunlp/KnowledgeEditingPapers.

[3]https://openkg-tutorial.github.io/.

[4]https://github.com/NLP-Tutorials/AACL-IJCNLP2022-KGC-Tutorial.

[5]https://2022.declarativeai.net/events/reasoning-web/rw-lectures.

Presentation Topic	Presenter	Time
Introduction	Ningyu Zhang	15min
Preliminaries	Ningyu Zhang	15min
Methods for Preserve LLMs' Parameters	Yunzhi Yao	40min
Coffee break	-	30min
Methods for Modify LLMs' Parameters	Yunzhi Yao	40min
Extensions	Shumin Deng	40min
Open-sourced Tools	Yunzhi Yao	30min
Discussion on Main Issues & Opportunities	Ningyu Zhang	30min

Table 1: Tutorial Schedule

undesirable behaviors in many of these applications, there is a rising interest in methods for efficient and immediate model modifications. Consequently, we expect this tutorial to attract an audience of more than 100 attendees, reflecting the growing focus on adaptable and flexible approaches to enhancing LLM performance.

8. Ethical Considerations

Knowledge editing involves techniques designed to modify the behavior of pre-trained models. It's crucial, however, to acknowledge the potential risks: if misapplied, knowledge editing could cause models to produce harmful or inappropriate content. Thus, prioritizing safe and responsible practices in the application of knowledge editing is imperative. Ethical guidelines should steer the use of these techniques, accompanied by robust safeguards to deter misuse and prevent the generation of damaging outcomes.

9. Reading list

- "A Comprehensive Study of Knowledge Editing for Large Language Models", (Zhang et al., 2024)
- "Editing Large Language Models: Problems, Methods, and Opportunities", (Yao et al., 2023b)
- "Detoxifying Large Language Models via Knowledge Editing", (Wang et al., 2024a)
- "Editing Conceptual Knowledge for Large Language Models", (Wang et al., 2024b)
- "Evaluating the Ripple Effects of Knowledge Editing in Language Models", (Cohen et al., 2023a)
- "Can We Edit Multimodal Large Language Models?", (Cheng et al., 2023a)
- "Unveiling the Pitfalls of Knowledge Editing for Large Language Models", (Li et al., 2023)
- "Editing Personality for LLMs", (Mao et al., 2023)
- "Editing Language Model-based Knowledge Graph Embeddings", (Cheng et al., 2023b)
- "Memory-Based Model Editing at Scale", (Mitchell et al., 2022c)
- "Calibrating Factual Knowledge in Pretrained Language Models", (Dong et al., 2022)
- "Transformer-Patcher: One Mistake worth One Neuron", (Huang et al., 2023)
- "Can We Edit Factual Knowledge by In-Context Learning?", (Zheng et al., 2023)
- "Editing Factual Knowledge in Language Models", (Cao et al., 2021)
- "Fast Model Editing at Scale", (Mitchell et al., 2022a)
- "Knowledge Neurons in Pretrained Transformers", (Dai et al., 2022a)
- "Locating and Editing Factual Associations in GPT", (Meng et al., 2022a)
- "Mass-Editing Memory in a Transformer", (Meng et al., 2023)
- "MQUAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions", (Zhong et al., 2023)
- "Can LMs Learn New Entities from Descriptions? Challenges in Propagating Injected Knowledge", (Gupta et al., 2023)
- "Detecting Edit Failures In Large Language Models: An Improved Specificity Benchmark", (Hoelscher-Obermaier et al., 2023)
- "Editing Commonsense Knowledge in GPT", (Gupta et al., 2023)

10. Presenters

Ningyu Zhang is an associate professor/doctoral supervisor at Zhejiang University, leading the group about KG and NLP technologies. He has supervised to construct a information extraction toolkit named DeepKE[6] (2.8K+ stars on Github). His research interest include knowledge graph and natural language processing. He has published many papers in top international academic conferences and journals such as Natural Machine Intelligence, Nature Communications, NeurIPS, ICLR, AAAI, IJCAI, WWW, KDD, SIGIR, ACL, ENNLP, NAACL, and IEEE/ACM Transactions on Audio Speech and Language. He has served as Area Chair for ACL/EMNLP 2023, ARR Action Editor, Senior Program Committee member for IJCAI 2023, Program Committee member for EMNLP, NAACL, NeurIPS, ICLR, ICML, WWW, SIGIR, KDD, AAAI, and reviewer for TKDE, TKDD.

Email: zhangningyu@zju.edu.cn

Homepage: https://person.zju.edu.cn/en/ningyu

Yunzhi Yao is a Ph.D candidate at at School of Computer Science and Technology, Zhejiang University. His research interests focus on Editing Large Language Models and Knowledge-enhanced Natural Language Processing. He has been research intern at Microsoft Research Asia supervised by Shaohan Huang, and research intern at Alibaba Group. He has published many papers in ACL, EMNLP, NAACL, SIGIR. For tutorial experience, he has given talks at AI-TIME to deliver his recent works. Moreover, he is the first author of the paper "**Editing Large Language Models: Problems, Methods, and Opportunities**" and one of the developers of the knowledge editing framework EasyEdit, which is related to this tutorial.

Email: yyztodd@zju.edu.cn

Homepage: https://scholar.google.ch/citations?user=nAagIwEAAAAJ

Shumin Deng is a research fellow at Department of Computer Science, School of Computing (SoC), National University of Singapore. She have obtained her Ph.D. degree at School of Computer Science and Technology, Zhejiang University. Her research interests focus on Natural Language Processing, Knowledge Graph, Information Extraction, Neuro-Symbolic Reasoning and LLM Reasoning. She has been awarded 2022 Outstanding Graduate of Zhejiang Province, China; 2020 Outstanding Intern in Academic Cooperation of Alibaba Group. She is a member of ACL, and a member of the Youth Working Committee of the Chinese Information Processing Society of China. She has serves as a Research Session (Information Extraction) Chair for EMNLP 2022, and a Publication Chair for CoNLL 2023. She has been a Journal Reviewer for many high-quality journals, such as TPAMI, TASLP, TALLIP, WWWJ, ESWA, KBS and so on; and serves as a Program Committee member for NeurIPS, ICLR, ACL, EMNLP, EACL, AACL, WWW, AAAI, IJCAI, CIKM and so on. She has constructed a billion-scale Open Business Knowledge Graph (OpenBG), and released a leaderboard[7] which has attracted thousands of teams and researchers.

Email: shumin@nus.edu.sg

Homepage: https://231sm.github.io/

[6] https://github.com/zjunlp/DeepKE.

[7] https://tianchi.aliyun.com/dataset/dataDetail?dataId=122271&lang=en-us.

11. Bibliographical References

Ahmed Alajrami and Nikolaos Aletras. 2022. How does the pre-training objective affect what large language models learn about linguistic properties? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 131–147. Association for Computational Linguistics.

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona T. Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *CoRR*, abs/2204.06031.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Iz Beltagy, Arman Cohan, Robert L. Logan IV, Sewon Min, and Sameer Singh. 2022. Zero- and few-shot NLP with pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022 - Tutorial Abstracts, Dublin, Ireland, May 22-27, 2022*, pages 32–37. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Magdalena Biesialska, Katarzyna Biesialska, and Marta R. Costa-jussà. 2020. Continual lifelong learning in natural language processing: A survey. In *COLING*, pages 6523–6541. International Committee on Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *NeurIPS*.

Boxi Cao, Hongyu Lin, Xianpei Han, and Le Sun. 2023a. The life cycle of knowledge in big language models: A survey. *CoRR*, abs/2303.07616.

Boxi Cao, Qiaoyu Tang, Hongyu Lin, Xianpei Han, Jiawei Chen, Tianshu Wang, and Le Sun. 2023b. Retentive or forgetful? diving into the knowledge memorizing mechanism of language models. *arXiv preprint arXiv:2305.09144*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6491–6506. Association for Computational Linguistics.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Xiaodong Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting training data from large language models. In *USENIX Security Symposium*.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2778–2788. ACM.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons.

Siyuan Cheng, Bozhong Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Ningyu Zhang. 2023a. Can we edit multimodal large language models? *arXiv preprint arXiv:2310.08475*.

Siyuan Cheng, Ningyu Zhang, Bozhong Tian, Zelin Dai, Feiyu Xiong, Wei Guo, and Huajun Chen. 2023b. Editing language model-based knowledge graph embeddings. *CoRR*, abs/2301.10405.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023a. Evaluating the ripple effects of knowledge editing in language models. *CoRR*, abs/2307.12976.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023b. Evaluating the ripple effects of knowledge editing in language models. *arXiv preprint arXiv:2307.12976*.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023c. Evaluating the ripple effects of knowledge editing in language models. *ArXiv*, abs/2307.12976.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022a. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2022b. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. *CoRR*, abs/2212.10559.

Joe Davison, Joshua Feldman, and Alexander M. Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1173–1178. Association for Computational Linguistics.

Shumin Deng, Chengming Wang, Zhoubo Li, Ningyu Zhang, Zelin Dai, Hehong Chen, Feiyu Xiong, Ming Yan, Qiang Chen, Mosha Chen, Jiaoyan Chen, Jeff Z. Pan, Bryan Hooi, and Huajun Chen. 2023. Construction and applications of billion-scale pre-trained multimodal business knowledge graph. In *ICDE*. IEEE.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Trans. Assoc. Comput. Linguistics*, 10:257–273.

Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022. Calibrating factual

knowledge in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5937–5947. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. 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 30–45. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5484–5495. Association for Computational Linguistics.

Anshita Gupta, Debanjan Mondal, Akshay Krishna Sheshadri, Wenlong Zhao, Xiang Lorraine Li, Sarah Wiegreffe, and Niket Tandon. 2023. Editing commonsense knowledge in GPT. *CoRR*, abs/2305.14956.

Y. Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. In *Proc. of AAAI*.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *ArXiv*, abs/2301.04213.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264, Dubrovnik, Croatia. Association for Computational Linguistics.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2023a. Inspecting and editing knowledge representations in language models.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2023b. Measuring and manipulating knowledge representations in language models. *CoRR*, abs/2304.00740.

Jason Hoelscher-Obermaier, Julia Persson, Esben Kran, Ioannis Konstas, and Fazl Barez. 2023. Detecting edit failures in large language models: An improved specificity benchmark. *CoRR*, abs/2305.17553.

Ari Holtzman, Peter West, and Luke Zettlemoyer. 2023a. Generative models as a complex systems science: How can we make sense of large language model behavior? *CoRR*, abs/2308.00189.

Ari Holtzman, Peter West, and Luke Zettlemoyer. 2023b. Generative models as a complex systems science: How can we make sense of large language model behavior? *ArXiv*, abs/2308.00189.

Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023. Transformer-patcher: One mistake worth one neuron. In *The Eleventh International Conference on Learning Representations*.

Jacques Thibodeau. 2022. But is it really in rome? an investigation of the rome model editing technique.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8:423–438.

Nikhil Kandpal, H. Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2022. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. 2023. Unveiling the pitfalls of knowledge editing for large language models. *CoRR*, abs/2310.02129.

Yuxi Ma, Chi Zhang, and Song-Chun Zhu. 2023. Brain in a vat: On missing pieces towards artificial general intelligence in large language models. *CoRR*, abs/2307.03762.

Shengyu Mao, Ningyu Zhang, Xiaohan Wang, Mengru Wang, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2023. Editing personality for llms. *CoRR*, abs/2310.02168.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Zaiqiao Meng, Fangyu Liu, Ehsan Shareghi, Yixuan Su, Charlotte Collins, and Nigel Collier. 2022b. Rewire-then-probe: A contrastive recipe for probing biomedical knowledge of pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4798–4810. Association for Computational Linguistics.

Jack Merullo, Carsten Eickhoff, and Elizabeth-Jane Pavlick. 2023a. Language models implement simple word2vec-style vector arithmetic. *ArXiv*, abs/2305.16130.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2023b. Language models implement simple word2vec-style vector arithmetic. *CoRR*, abs/2305.16130.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022a. Fast model editing at scale. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022b. Fast model editing at scale. In *International Conference on Learning Representations*.

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

Shikhar Murty, Christopher D. Manning, Scott M. Lundberg, and Marco Túlio Ribeiro. 2022. Fixing model bugs with natural language patches. 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 11600–11613. Association for Computational Linguistics.

Yasumasa Onoe, Michael J. Q. Zhang, Shankar Padmanabhan, Greg Durrett, and Eunsol Choi. 2023. Can lms learn new entities from descriptions? challenges in propagating injected knowledge. *CoRR*, abs/2305.01651.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 2023. What in-context learning" learns" in-context: Disentangling task recognition and task learning. *arXiv preprint arXiv:2305.09731*.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5368–5393. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. In chatgpt we trust? measuring and characterizing the reliability of chatgpt.

Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitry Pyrkin, Sergei Popov, and Artem Babenko. 2020. Editable neural networks. In *International Conference on Learning Representations*.

Hao Sun, Zhexin Zhang, Jiawen Deng, Jiale Cheng, and Minlie Huang. 2023. Safety assessment of chinese large language models. *CoRR*, abs/2304.10436.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. olmpics - on what language model pre-training captures. *Trans. Assoc. Comput. Linguistics*, 8:743–758.

Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan S. Kankanhalli. 2021. Fast yet effective machine unlearning. *IEEE transactions on neural networks and learning systems*, PP.

Bozhong Tian, Siyuan Cheng, Xiaozhuan Liang, Ningyu Zhang, Yi Hu, Kouying Xue, Yanjie Gou, Xi Chen, and Huajun Chen. 2024. Instructedit: Instruction-based knowledge editing for large language models. *arXiv preprint arXiv:2402.16123*.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. In *NeurIPS*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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

Ben Wang and Aran Komatsuzaki. 2021b. Gpt-j-6b: A 6 billion parameter autoregressive language model.

Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi Yang, Jindong Wang, and Huajun Chen. 2024a. Detoxifying large language models via knowledge editing. *arXiv preprint arXiv:2403.14472*.

Peng Wang, Ningyu Zhang, Xin Xie, Yunzhi Yao, Bozhong Tian, Mengru Wang, Zekun Xi, Siyuan Cheng, Kangwei Liu, Guozhou Zheng, and Huajun Chen. 2023. Easyedit: An easy-to-use knowledge editing framework for large language models. *CoRR*, abs/2308.07269.

Xiaohan Wang, Shengyu Mao, Ningyu Zhang, Shumin Deng, Yunzhi Yao, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024b. Editing conceptual knowledge for large language models. *arXiv preprint arXiv:2403.06259*.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022. Finding skill neurons in pre-trained transformer-based language models. 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 11132–11152. Association for Computational Linguistics.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4602–4625. Association for Computational Linguistics.

Ga Wu, Masoud Hashemi, and Christopher Srinivasa. 2022. Puma: Performance unchanged model augmentation for training data removal. In *AAAI Conference on Artificial Intelligence*.

Weiqi Wu, Chengyue Jiang, Yong Jiang, Pengjun Xie, and Kewei Tu. 2023. Do plms know and understand ontological knowledge? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3080–3101. Association for Computational Linguistics.

Yang Xu, Yutai Hou, and Wanxiang Che. 2022. Language anisotropic cross-lingual model editing. *ArXiv*, abs/2205.12677.

Yang Xu, Yutai Hou, Wanxiang Che, and Min Zhang. 2023. Language anisotropic cross-lingual model editing. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5554–5569. Association for Computational Linguistics.

Yunzhi Yao, Shaohan Huang, Ningyu Zhang, Li Dong, Furu Wei, and Huajun Chen. 2022. Kformer: Knowledge injection in transformer feed-forward layers. In *Natural Language Processing and Chinese Computing*.

Yunzhi Yao, Peng Wang, Shengyu Mao, Chuanqi Tan, Fei Huang, Huajun Chen, and Ningyu Zhang. 2023a. Knowledge rumination for pre-trained language models. *CoRR*, abs/2305.08732.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen,

and Ningyu Zhang. 2023b. Editing large language models: Problems, methods, and opportunities. *CoRR*, abs/2305.13172.

Chenyu Zhang, Benjamin Van Durme, Zhuowan Li, and Elias Stengel-Eskin. 2022a. Visual commonsense in pretrained unimodal and multimodal models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5321–5335. Association for Computational Linguistics.

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. 2024. A comprehensive study of knowledge editing for large language models. *CoRR*, abs/2401.01286.

Ningyu Zhang, Jintian Zhang, Xiaohan Wang, Honghao Gui, Kangwei Liu, Yinuo Jiang, Xiang Chen, Shengyu Mao, Shuofei Qiao, Yuqi Zhu, Zhen Bi, Jing Chen, Xiaozhuan Liang, Yixin Ou, Runnan Fang, Zekun Xi, Xin Xu, Lei Li, Peng Wang, Mengru Wang, Yunzhi Yao, Bozhong Tian, Yin Fang, Guozhou Zheng, and Huajun Chen. 2023. Knowlm: An open-sourced knowledgeable large langugae model framework.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2022b. GreaseLM: Graph REASoning enhanced language models. In *International Conference on Learning Representations*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: enhanced language representation with informative entities. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1441–1451. Association for Computational Linguistics.

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. *CoRR*, abs/2303.18223.

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

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

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9733–9740. AAAI Press.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix X. Yu, and Sanjiv Kumar. 2020. Modifying memories in transformer models. *ArXiv*, abs/2012.00363.