Editing Large Language Models

Ningyu Zhang[♣], Yunzhi Yao[♣], Shumin Deng[♠]
[♣]Zhejiang University, China [♠]National University of Singapore, Singapore
{zhangningyu,yyztodd}@zju.edu.cn, shumin@nus.edu.sg
https://github.com/zjunlp/KnowledgeEditingPapers

Abstract

Even with their impressive abilities, Large Language Models (LLMs) such as ChatGPT are not immune to issues of factual or logically consistent. Concretely, the key concern is how to seamlessly update those LLMs to correct mistakes without resorting to an exhaustive retraining or continuous training procedure, both of which can demand significant computational resources and time. Thus, the capability to edit LLMs offers an efficient solution to alter a model's behavior, notably within a distinct area of interest, without negatively impacting its performance on other tasks. Through this tutorial, we strive to acquaint interested NLP researchers with recent and emerging techniques for editing LLMs. Specifically, we aim to present a systematic and current overview of cutting-edge methods, supplemented with practical tools, and unveil new research opportunities for our audiences. All the valuable resources can be accessed at https://github.com/zjunlp/KnowledgeEditingPapers.

1 Introduction

Motivation: Large Language Models (LLMs) have showcased their immense potential in generating human-like text as demonstrated by numerous studies (Brown et al., 2020; OpenAI, 2023; Chen et al., 2022; Qiao et al., 2023; Zhao et al., 2023; Ma et al., 2023). Despite their remarkable capabilities, LLMs such as ChatGPT can sometimes falter in terms of factuality or logical consistency. They might inadvertently generate content that is harmful or offensive, and they are unaware of events that have occurred post their training phase. The challenge at hand, quite intuitively, remains how to effectively update these LLMs or rectify their errors without resorting to a complete retraining or continuous training process, both of which can be significantly resource-intensive and time-consuming. To this end, the notion of **model editing** has been proposed (Sinitsin et al., 2020; De Cao et al., 2021) which provide an efficient way to modify a model's behavior, particularly within a specified domain of interest, without compromising its performance on other inputs.

Recently, editing LLMs has garnered substantial attention due to its impressive capability to rectify errors, adapt to diverse scenarios, or customize to fulfill particular use cases. In this tutorial, we aim to offer a holistic view on the process of editing LLMs. We will explore common challenges and discuss potential solutions associated with various methodologies. Our exploration will encompass strategies for preserving LLMs' parameters and techniques for modifying them. Additionally, we will introduce a range of open-sourced tools designed for specific applications, thus providing a more comprehensive understanding of the field and its practical aspects.

Note that our tutorial is related to understanding the principles that govern how pre-trained language models encapsulate knowledge (Geva et al., 2021, 2022; Haviv et al., 2023; Hao et al., 2021; Hernandez et al., 2023; Yao et al., 2023a; Cao et al., 2023). Techniques of model editing, which involve the manipulation of external knowledge, bear similarities to knowledge augmentation approaches, since updating a model's knowledge can also be viewed as infusing new knowledge into the model. Moreover, model editing can be interpreted as a specific instance of lifelong learning (Biesialska et al., 2020) and unlearning (Wu et al., 2022; Tarun et al., 2021), wherein the model is capable of adaptively and continuously incorporating and adjusting new knowledge, while simultaneously discarding outdated and incorrect information. Furthermore, model editing could help mitigate the production of toxic or harmful language by LLMs, indicating that such techniques could be instrumental in addressing the security and privacy concerns associated with these models (Geva et al., 2022). We plan

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Tutorial Abstracts), pages 23–28
November 1–4, 2023. ©2023 Association for Computational Linguistics

to delve into these aspects and propose potential future directions in this area.

Tutorial Content: We will initiate this tutorial by defining the tasks involved in editing LLMs, and introducing evaluation metrics and benchmark datasets. Subsequently, we will provide an overview of various model editing methodologies. Initially, our focus will be on methods that preserve the parameters of LLMs. These methods operate by manipulating the model's output for specific cases through the integration of an auxiliary network with the original, unaltered model. We will then transition to methods that modify the parameters of LLMs, which aim to alter the model parameters responsible for undesirable output. Throughout the tutorial, we aim to share insights gleaned from the diverse communities engaged in LLM editing research and introduce open-sourced plug-and-play model editing toolkits. Further, we will delve into potential issues as well as opportunities associated with editing LLMs, with the goal of imparting valuable insights to the NLP community. The schedule and content outline of the tutorial can be found in Table 1 and Section 3. All tutorial slides will be available at `https://github.com/zjunlp/KnowledgeEditingPapers`.

Relevance to AACL: LLMs, such as ChatGPT, have significantly enriched a wide array of crucial NLP tasks, thereby commanding substantial influence across the entire community. A 2022 ACL tutorial, "Zero- and Few-Shot NLP with Pretrained Language Models" (Beltagy et al., 2022), introduces recent advancements and emerging techniques in zero- and few-shot settings based on pretrained language models. More recently, another ACL tutorial titled "Everything You Need to Know about Multilingual LLMs: Towards Fair, Performant, and Reliable Models for Languages of the World" provides an overview of Multilingual LLMs with a focus on fairness, performance, and reliability. However, currently, there is no tutorial specifically dedicated to editing LLMs. Our aim is to deliver this tutorial at AACL, where we will focus on a specific, pertinent problem for LLMs, and present an in-depth exploration of the latest work within the span of a concise 3-hour session.

2 Type of this Tutorial

This tutorial will encompass **cutting-edge** methodologies pertaining to the editing of LLMs. However, our content will also include **introductory material** on LLMs, designed to cater to a broad audience within the NLP community. Additionally, we plan to introduce pertinent tools beneficial for both beginners and developers in the field.

3 Outline

The tutorial mainly includes the following parts.

3.1 Introduction and Background

This part consists of: (1) Background and definition: background of editing LLMs with problem definition; (2) Metrics and datasets: a brief introduction of metrics and benchmark datasets; (3) Editing LLMs as phenomenon: the importance and impact of editing LLMs to other broader AI and NLP tasks with challenges and opportunities.

3.2 Methods for Preserve LLMs' Parameters

This part includes cutting-edge methods for editing LLMs with methods for preserve LLMs' parameters, including: (1) SERAC (Mitchell et al., 2022b); (2) CaliNet (Dong et al., 2022); (3) T-Patcher (Huang et al., 2023).

3.3 Methods for Modify LLMs' Parameters

This part includes cutting-edge methods for editing LLMs with methods for modify LLMs' parameters, including: (1) KE (De Cao et al., 2021); (2) MEND (Mitchell et al., 2022a); (3) KN (Dai et al., 2022); (4) ROME (Meng et al., 2022); (5) MEMIT (Meng et al., 2023);

3.4 Open-sourced Tools

This part includes the open-sourced tools *EasyEdit*[1] (Wang et al., 2023) which is an easy-to-use framework to edit LLMs. We will introduce how to use and develop model editing methods with *EasyEdit* via examples.

3.5 Discussion on Main Issues & Opportunities

This section will present the primary challenges and opportunities, including: (1) Generalization issues such as those encountered while performing edits in language models through multi-hop questions (Zhong et al., 2023); (2) Challenges related to locality, robustness, and safety in the context of model editing (Hoelscher-Obermaier et al., 2023); (3) Other types of model (knowledge) editing, touching on aspects such as commonsense

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

Presentation Topic	Presenter	Time
Introduction and Background	Ningyu Zhang	20min
Methods for Preserve LLMs' Parameters	Shumin Deng	40min
Methods for Modify LLMs' Parameters (I)	Yunzhi Yao	30min
Coffee break	-	30min
Methods for Modify LLMs' Parameters (II)	Yunzhi Yao	30min
Open-sourced Tools	Yunzhi Yao	30min
Discussion on Main Issues & Opportunities	Ningyu Zhang	30min

Table 1: Tutorial Schedule

reasoning and multimodal LLMs (Gupta et al., 2023); (4) Efficient model editing using parameter-efficient tuning strategies; (5) In-context model editing approaches (Zheng et al., 2023).

4 Prerequisites

This tutorial is accessible to anyone with a foundational knowledge in natural language processing. In addition, a rudimentary understanding of neural networks is beneficial, and familiarity with large language models, along with parameter-efficient tuning, would be particularly advantageous.

5 Reading list

- "Editing Large Language Models: Problems, Methods, and Opportunities", (Yao et al., 2023b)
- "Memory-Based Model Editing at Scale", (Mitchell et al., 2022b)
- "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", (De Cao et al., 2021)
- "Fast Model Editing at Scale", (Mitchell et al., 2022a)
- "Knowledge Neurons in Pretrained Transformers", (Dai et al., 2022)
- "Locating and Editing Factual Associations in GPT", (Meng et al., 2022)
- "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)

6 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[2] (1.9K+ 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 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. He has contributed the following tutorials:

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

(1) **IJCAI 2023 Tutorial: Open-Environment Knowledge Graph Construction and Reasoning: Challenges, Approaches, and Opportunities (3-hour tutorial)**;

(2) **AACL 2022 Tutorial: Efficient and Robust Knowledge Graph Construction (3-hour tutorial)**;

(3) **China Conference on Knowledge Graph and Semantic Computing (CCKS 2022) Tutorial: Efficient Knowledge Graph Construction and Reasoning (1.5-hour tutorial)**;

(4) **The 18th Reasoning Web Summer School: Cross-Modal Knowledge Discovery, Inference, and Challenges (3-hour tutorial)**;

(5) **Knowledge Graph Conference (KGC2023 Speakers): Efficient Knowledge Graph Construction with Pre-trained Language Models (to present)**;

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. Her 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**" 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 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 and so on. She has constructed a billion-scale Open Business Knowledge Graph (OpenBG) (Deng et al., 2023), and released a leaderboard[3] which has attracted thousands of teams and researchers. She has contributed the following tutorial:

IJCAI 2023 Tutorial: Open-Environment Knowledge Graph Construction and Reasoning: Challenges, Approaches, and Opportunities (3-hour tutorial)

Email: shumin@nus.edu.sg

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

References

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.

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

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

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.

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

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*, pages 6491–6506, Online and Punta Cana, Dominican Republic. 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.

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*, pages 5937–5947, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Kevin 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*, pages 30–45, Abu Dhabi, United Arab Emirates. 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*, pages 5484–5495, Online and Punta Cana, Dominican Republic. 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*.

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. 2023. 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.

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*.

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.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. 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*.

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

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

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

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.

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

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.

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.

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

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.

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.