Enhancing LLM Capabilities Beyond Scaling Up

Wenpeng Yin[†], **Muhao Chen**[♠‡], **Rui Zhang**[†], **Ben Zhou**[⋆], **Fei Wang**[‡], **Dan Roth**[⋄♯]
[†]Penn State; [♠]UC Davis; [⋆]ASU; [‡]USC; [⋄]Oracle; [♯]UPenn
{wenpeng,rmz5227}@psu.edu; muhchen@ucdavis.edu
benzhou@asu.edu; fwang598@usc.edu; danroth@seas.upenn.edu

Abstract

General-purpose large language models (LLMs) are progressively expanding both in scale and access to unpublic training data. This has led to notable progress in a variety of AI problems. Nevertheless, two questions exist: i) Is scaling up the sole avenue of extending the capabilities of LLMs? ii) Instead of developing general-purpose LLMs, how to endow LLMs with specific knowledge? This tutorial targets researchers and practitioners who are interested in capability extension of LLMs that go beyond scaling up. To this end, we will discuss several lines of research that follow that direction, including: (i) optimizing input prompts to fully exploit LLM potential, (ii) enabling LLMs to self-improve responses through various feedback signals, (iii) updating or editing the internal knowledge of LLMs when necessary, (iv) leveraging incidental structural supervision from target tasks, and (v) defending against potential attacks and threats from malicious users. At last, we will conclude the tutorial by outlining directions for further investigation.[1]

1 Introduction

The advancement of AI can be broadly attributed to two technical trajectories: one involving general-purpose models, and the other centering around task-specific models. In the earlier phases of deep learning and even before its inception, the focal point of research predominantly revolved around the integration of domain-specific and task-specific expertise into model architectures. Nonetheless, the landscape underwent a transformation with the advent of pretrained large language models (LLMs), e.g., BERT (Devlin et al., 2019) and GPT series (OpenAI, 2022, 2023). Recent years have witnessed substantial achievements of those general-purpose models in a variety of AI problems. However, the advancements facilitated by LLMs are primarily rooted in larger scales of model parameters and confidential training data. These factors make LLMs increasingly costly, uninterpretable, unreproducible, uncontrollable, and unmanageable for most users.

Consequently, while acknowledging the substantial benefits offered by LLMs, it becomes crucial to address several pertinent inquiries. Firstly, *does the path to enhancing LLMs' capabilities solely involve scaling up?* The resource-intensive nature of training large-scale LLMs prompts the exploration of potential bottlenecks and the feasibility of further expansion. Secondly, *despite LLMs' versatility, challenges persist in their application to specific disciplines, tasks, and even users*. Thus, strategies to augment LLMs' capabilities for these distinctive challenges warrant consideration.

This tutorial delves into some research lines that extend the capabilities of LLMs beyond mere scale amplification. Specifically, it presents a comprehensive analysis of this objective, identifying challenges across five key dimensions: *optimizing LLM inputs*, *enhancing LLM responses*, *updating LLMs' internal knowledge*, *maximizing supervision from the target task*, and *improving LLM trustworthiness*. In line with these dimensions, the tutorial will address recent advancements in: (i) prompt optimization (§2.2), (ii) LLM self-improvement and inter-LLM collaboration (§2.3), (iii) adapting pre-existing knowledge to integrate new, potentially conflicting information (§2.4), (iv) aligning LLM performance with the constraints and structures of target problems (§2.5), and (v) defending against adversarial threats and malicious attacks (§2.6).

We believe it is necessary to present a timely tutorial to comprehensively summarize the new frontiers in LLM capability extension research and point out the emerging challenges that deserve further investigation. Participants will learn about

[1]Materials available at www.wenpengyin.org/publications/beyond-llm-scaling-emnlp24

recent trends, emerging challenges, and representative tools in this topic, and how related technologies benefit end-user NLP applications.

2 Outline of Tutorial Content

This **half-day** tutorial presents a systematic overview of recent advancements in extending LLMs' capabilities without scaling up. The detailed contents are outlined below.

2.1 Background and Motivation

We will begin motivating this topic with a selection of real-world applications and emerging challenges of general-purpose LLMs.

2.2 Prompt Optimization for LLMs

Large Language Models (LLMs) have shown remarkable performance across a wide range of tasks. However, they are known to be sensitive to prompt variations, where even slight changes in input can cause substantial differences in output quality (Lu et al., 2021). As a result, effective prompt design has become essential for maximizing LLM performance. Despite this, finding the optimal prompts still often involves manual trial and error, which demands considerable human effort and can yield suboptimal results (Wei et al., 2022; Kojima et al., 2022). In this section, we will introduce several emerging techniques of prompt optimization for LLMs, which aim to systematically search for prompts that improve target task performance. We organize our discussion into several categories including search-based prompt optimization (Prasad et al., 2022; Guo et al., 2023; Schnabel and Neville, 2024), text gradient–based prompt optimization (Pryzant et al., 2023; Ye et al., 2023; Yuksekgonul et al., 2024), and gradient-based prompt optimization (Wen et al., 2024). We will conclude this section by presenting several promising future directions such as prompt optimization for multiagent LLMs, optimization for long and complex prompts, prompt optimization by retrieving and augmenting domain knowledge, human-in-the-loop interactive prompt optimization, and theoretical analysis of prompt optimization.

2.3 LLM Self-improvement & LLM-LLM Collaboration

In this subsection, we provide a detailed discussion on how LLMs can harness their own capabilities for self-improvement or collaborate with peer LLMs to address more complex problems.

The concept of LLM self-improvement has garnered increasing attention in recent literature (Kamoi et al., 2024; Pan et al., 2023b). On one hand, a growing body of work has demonstrated the potential of self-improvement strategies (Kumar et al., 2024; Kim et al., 2023; Huang et al., 2023b; Patel et al., 2024; Jiang et al., 2024a), including techniques like self-feedback (Madaan et al., 2023) and self-discriminative abilities (Ahn et al., 2024). On the other hand, some studies have questioned the effectiveness of these self-improvement mechanisms (Stechly et al., 2023; Huang et al., 2024; Jiang et al., 2024b; Valmeekam et al., 2023).

In addition to exploring the limits of individual LLM capabilities, we also examine recent advancements in combining multiple LLMs. These include: i) LLM-LLM collaboration, such as detecting factual errors through cross-examination (Cohen et al., 2023), multi-agent cooperation (Du et al., 2024; Talebirad and Nadiri, 2023), and LLM control of other AI agents (Shen et al., 2023); ii) LLM-LLM merging, which aims to produce a new, singular "super" LLM (Tam et al., 2024; Tam et al.; Liu et al., 2024a; Goddard et al., 2024; Perin et al., 2024).

2.4 Knowledge Update of LLMs

LLMs encapsulate vast world knowledge acquired during pre-training, yet the ever-evolving nature of information often results in *outdated or biased knowledge*, potentially leading to the dissemination of misinformation. In this section, we first examine the issues caused by unreliable knowledge, such as hallucinations (Xu et al., 2024c; Longpre et al., 2021; Li et al., 2023a; Wang et al., 2023c). Next, we explore approaches to remedy knowledge gaps in LLMs' internal knowledge by integrating external information in a training-free manner. We begin by enforcing LLMs' reliance on external context when the external knowledge is verified as reliable (Wang et al., 2023a; Zhou et al., 2023). We then address more general and realistic scenarios where both internal and external knowledge may be noisy, discussing effective strategies for combining these sources (Zhang et al.; Zhao et al., 2024). Finally, we introduce techniques for knowledge editing in LLMs with lightweight tuning (Lin et al., 2024; Wang et al., 2024c; Huang et al., 2023a).

2.5 Aligning with Structures of Target Problems

Aligning models with pre-defined structures is an efficient method of improving model performances without scaling up. During this process, models adapt to structures that are beneficial to solving target problems and produce outputs that are more consistent with expectations. We discuss three types of such structures in this section. The first type uses symbolic constraints as structures, which include human-written constraints (Wang et al., 2024b), mathematical constraints (Feng et al., 2024), and compiler constraints (Chen et al., 2023; Zhu et al., 2024). The second type finds structures from decomposing the target problem (Sun et al., 2023; Chen et al., 2024b; Zhou et al., 2024b; Wu and Xie, 2024). The last type of structures are procedural structures that come from cognitive or problem-solving processes, such as DSP (Khattab et al., 2022), ReAct (Yao et al., 2022), and RAP (Hao et al., 2023). These procedural structures can also be combined with symbolic constraints (Pan et al., 2023a), task decompositions (Hu et al., 2023; LYU et al., 2023), or both (Zhou et al., 2024a).

2.6 Safety Enhancement for LLMs

Despite the desire to align LLM responses with users' preferences, malicious data may exist in the training corpora, task instructions, and human feedback. These data are likely to cause threats to LLMs before they are deployed as services (Wan et al., 2023; Xu et al., 2024a; Greshake et al., 2023). Due to the limited accessibility of model components in these services, mitigating such threats needs to be addressed through inference-time defense rather than training-time safety enhancement (Wang et al., 2024a). In this part of the tutorial, we will first introduce **inference-time threats** to LLMs through prompt injection, malicious task instructions, jailbreaking attacks, adversarial demonstrations, and training-free backdoor attacks (Liu et al., 2023b; Xu et al., 2024a; Li et al., 2023b; Wang et al., 2023b; Huang et al., 2023c; Greshake et al., 2023; Xu et al., 2024b). We will then provide insights on mitigating some of those threats based on **defense techniques** including prompt robustness estimation, demonstration-based defense and ensemble debiasing (Liu et al., 2023a, 2024b; Graf et al., 2024; Wu et al., 2023), defensive demonstrations (Mo et al., 2023), or detection techniques where defenders can detect and eliminate poisoned data given the compromised model (Kurita et al., 2020; Chen and Dai, 2021; Qi et al., 2021; Li et al., 2021, 2023c). While many issues with inference-time threats remain unaddressed (Chen et al., 2024a). We will also provide a discussion about how the community should develop to combat those issues.

2.7 Future Research Directions

Enhancing general-purpose large language models (LLMs) with specialized capabilities tailored to specific datasets, problems, and user requirements is essential for their effective deployment in real-world applications. We conclude this tutorial by discussing several ongoing challenges and promising avenues for future research, including: (i) adapting LLMs to different scientific disciplines to model complex processes (Jadhav et al., 2024; Thirunavukarasu et al., 2023), (ii) employing Mixture of Experts architectures (Sukhbaatar et al., 2024; Xue et al., 2024), (iii) exploring novel approaches for constructing foundational models that transcend Transformer-based generative AI, such as Liquid Foundation Models [2], and (iv) advancing autonomous systems for goal planning, action execution, and self-evolution through continuous learning (Crowder et al., 2020).

3 Specification of the Tutorial

The proposed tutorial is considered a **cutting-edge** tutorial that introduces new frontiers in LLM capability extension beyond scaling up its size and data. The presented topic has not been covered by any *CL tutorials in the past 4 years.

Audience and Prerequisites Based on the level of interest in this topic, we expect around 250 participants. While no specific background knowledge is assumed of the audience, it would be best for the attendees to know about basic deep learning technologies, pre-trained language models (e.g. encoder-based LLMs and decoder-based LLMs). A reading list that could help provide background knowledge to the audience before attending this tutorial is given in Appx. §A.2.

Breadth We estimate that at least 60% of the work covered in this tutorial is from researchers other than the instructors of the tutorial.

Diversity Considerations This tutorial will explore cutting-edge research on updating and adapting LLMs with new knowledge, user preferences, constraints, defense techniques, task capabilities,

[2] https://www.liquid.ai/liquid-foundation-models

and external tools/models. The team includes a senior Ph.D. student and several assistant and distinguished professors, and will promote the tutorial on social media to broaden audience participation.

4 Tutorial Instructors

The following are biographies of the speakers. Past tutorials given by us are listed in Appx. §A.1.

Wenpeng Yin is an Assistant Professor in the Department of Computer Science and Engineering at Penn State University. Prior to joining Penn State, he was a tenure-track faculty member at Temple University (1/2022-12/2022), Senior Research Scientist at Salesforce Research (8/2019-12/2021), a postdoctoral researcher at UPenn (10/2017-7/2019), and got his Ph.D. degree from the Ludwig Maximilian University of Munich, Germany, in 2017. Dr. Yin's research focuses on natural language processing with three sub-areas: (i) NLP/LLM for scientific research, (ii) human-centered AI, and (iii) multimodal learning. Additional information is available at www.wenpengyin.org.

Muhao Chen is an assistant professor in the Department of Computer Science at UC Davis, where he directs the Language Understanding and Knowledge Acquisition (LUKA) Group. His research focuses on data-driven machine learning approaches for natural language understanding and knowledge acquisition. His work has been recognized with an NSF CRII Award, two Amazon Research Awards, a Cisco Faculty Research Award, an EMNLP Outstanding Paper Award, and an ACM SIGBio Best Student Paper Award. Muhao obtained his PhD degree from UCLA Department of Computer Science in 2019, was a postdoctoral researcher at UPenn, and worked as as Assistant Research Professor of Computer Science at USC prior to joining UC Davis. Additional information is available at http://luka-group.github.io.

Rui Zhang is an Assistant Professor in the Computer Science and Engineering Department of Penn State University and a co-director of the PSU NLP Lab. His overarching research goal is to build natural language interfaces for efficient information access and knowledge sharing including summarization for unstructured documents, question answering for semi-structured web tables and pages, and semantic parsing for structured knowledge. He has led a tutorial on contrastive data and learning for natural language processing at NAACL 2022. He is the co-organizer of several workshops including SUKI at NAACL 2022, MIA at NAACL 2022, and IntEx-SemPar at EMNLP 2020. Additional information is available at https://ryanzhumich.github.io/.

Ben Zhou is an Assistant Professor in the School of Computing and Augmented Intelligence at Arizona State University. Ben's research uses data and symbolic cognitive processes to improve model reasoning, controllability, and trustworthiness from learning/inference schemes and architectural perspectives. He has more than 10 recent papers on related topics. Ben obtained his Ph.D. degree from the University of Pennsylvania. He is a recipient of the ENIAC fellowship from the University of Pennsylvania and a finalist for the CRA Outstanding Undergraduate Researcher Award. Additional information is available at http://xuanyu.me/.

Fei Wang is a Ph.D. student in the Department of Computer Science at University of Southern California. His research interests lie in natural language processing and machine learning. His recent work focuses on enhancing the trustworthiness of LLMs with dynamic knowledge integration and robust alignment. Fei is a recipient of an Amazon ML Fellowship and an Annenberg Fellowship. Additional information is available at https://feiwang96.github.io/.

Dan Roth is the Eduardo D. Glandt Distinguished Professor at the Department of Computer and Information Science, UPenn, the Chief AI Scientist at Oracle, and a Fellow of the AAAS, ACM, AAAI, and ACL. In 2017, Roth was awarded the John McCarthy Award, the highest award the AI community gives to mid-career AI researchers. Roth was recognized "for major conceptual and theoretical advances in the modeling of natural language understanding, machine learning, and reasoning." Roth has published broadly in machine learning, NLP, KRR, and learning theory, and has given keynote talks and tutorials in all ACL and AAAI major conferences. Roth was the Editor-in-Chief of JAIR until 2017, and was the program chair of AAAI'11, ACL'03 and CoNLL'02; he serves regularly as an area chair and senior program committee member in the major conferences in his research areas. Additional information is available at www.cis.upenn.edu/~danroth.

Ethical Considerations

We do not anticipate any ethical issues particularly to the topics of the tutorial. Nevertheless, some work presented in this tutorial extensively uses large-scale pretrained models with self-attention, which may lead to substantial financial and environmental costs.

Acknowledgment

Muhao Chen was supported by the DARPA FoundSci Grant HR00112490370, the NSF of the United States Grant ITE 2333736 and an Amazon Research Award. Fei Wang was supported by the Amazon ML Fellowship.

References

Jihyun Janice Ahn, Ryo Kamoi, Lu Cheng, Rui Zhang, and Wenpeng Yin. 2024. Direct-inverse prompting: Analyzing llms' discriminative capacity in self-improving generation. *CoRR*, abs/2407.11017.

Chuanshuai Chen and Jiazhu Dai. 2021. Mitigating backdoor attacks in LSTM-based text classification systems by backdoor keyword identification. *Neurocomputing*, 452:253–262.

Muhao Chen, Chaowei Xiao, Huan Sun, Lei Li, Leon Derczynski, Anima Anandkumar, and Fei Wang. 2024a. Combating security and privacy issues in the era of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 8–18, Mexico City, Mexico. Association for Computational Linguistics.

Sihao Chen, Hongming Zhang, Tong Chen, Ben Zhou, Wenhao Yu, Dian Yu, Baolin Peng, Hongwei Wang, Dan Roth, and Dong Yu. 2024b. Sub-sentence encoder: Contrastive learning of propositional semantic representations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1596–1609.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023. Teaching large language models to self-debug. *ArXiv*, abs/2304.05128.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. LM vs LM: detecting factual errors via cross examination. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12621–12640. Association for Computational Linguistics.

James A Crowder, John Carbone, Shelli Friess, James A Crowder, John Carbone, and Shelli Friess. 2020. Artificial creativity and self-evolution: Abductive reasoning in artificial life forms. *Artificial Psychology: Psychological Modeling and Testing of AI Systems*, pages 65–74.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yu Feng, Ben Zhou, Weidong Lin, and Dan Roth. 2024. Bird: A trustworthy bayesian inference framework for large language models. *arXiv preprint arXiv:2404.12494*.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's mergekit: A toolkit for merging large language models. *CoRR*, abs/2403.13257.

Victoria Graf, Qin Liu, and Muhao Chen. 2024. Two heads are better than one: Nested poe for robust defense against multi-backdoors. In *NAACL*.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. More than you've asked for: A comprehensive analysis of novel prompt injection threats to application-integrated large language models. *arXiv preprint arXiv:2302.12173*.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2023. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. *arXiv preprint arXiv:2309.08532*.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173.

Y. Hu, Haotong Yang, Zhouchen Lin, and Muhan Zhang. 2023. Code prompting: a neural symbolic method for complex reasoning in large language models. *ArXiv*, abs/2305.18507.

James Y. Huang, Wenxuan Zhou, Fei Wang, Fred Morstatter, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023a. Offset unlearning for large language models. *arXiv preprint arXiv:2311.09763*.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023b. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1051–1068. Association for Computational Linguistics.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yujin Huang, Terry Yue Zhuo, Qiongkai Xu, Han Hu, Xingliang Yuan, and Chunyang Chen. 2023c. Training-free lexical backdoor attacks on language models. In *Proceedings of the ACM Web Conference 2023*, pages 2198–2208.

Yayati Jadhav, Peter Pak, and Amir Barati Farimani. 2024. Llm-3d print: Large language models to monitor and control 3d printing. *arXiv preprint arXiv:2408.14307*.

Chunyang Jiang, Chi-Min Chan, Wei Xue, Qifeng Liu, and Yike Guo. 2024a. Importance weighting can help large language models self-improve. *CoRR*, abs/2408.09849.

Dongwei Jiang, Jingyu Zhang, Orion Weller, Nathaniel Weir, Benjamin Van Durme, and Daniel Khashabi. 2024b. Self-[in]correct: Llms struggle with discriminating self-generated responses.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? A critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, abs/2406.01297.

O. Khattab, Keshav Santhanam, Xiang Lisa Li, David Leo Wright Hall, Percy Liang, Christopher Potts, and Matei A. Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *ArXiv*, abs/2212.14024.

Taehyeon Kim, Joonkee Kim, Gihun Lee, and Se-Young Yun. 2023. Distort, distract, decode: Instruction-tuned model can refine its response from noisy instructions. *CoRR*, abs/2311.00233.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. 2024. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*.

Keita Kurita, Paul Michel, and Graham Neubig. 2020. Weight poisoning attacks on pretrained models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2793–2806, Online. Association for Computational Linguistics.

Bangzheng Li, Ben Zhou, Fei Wang, Xingyu Fu, Dan Roth, and Muhao Chen. 2023a. Deceiving semantic shortcuts on reasoning chains: How far can models go without hallucination? In *Proceedings of NAACL 2023*.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, and Yangqiu Song. 2023b. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*.

Jiazhao Li, Zhuofeng Wu, Wei Ping, Chaowei Xiao, and VG Vydiswaran. 2023c. Defending against insertion-based textual backdoor attacks via attribution. *arXiv preprint arXiv:2305.02394*.

Zichao Li, Dheeraj Mekala, Chengyu Dong, and Jingbo Shang. 2021. BFClass: A backdoor-free text classification framework. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 444–453, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zihao Lin, Mohammad Beigi, Hongxuan Li, Yufan Zhou, Yuxiang Zhang, Qifan Wang, Wenpeng Yin, and Lifu Huang. 2024. Navigating the dual facets: A comprehensive evaluation of sequential memory editing in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13755–13772, Bangkok, Thailand. Association for Computational Linguistics.

Deyuan Liu, Zecheng Wang, Bingning Wang, Weipeng Chen, Chunshan Li, Zhiying Tu, Dianhui Chu, Bo Li, and Dianbo Sui. 2024a. Checkpoint merging via bayesian optimization in LLM pretraining. *CoRR*, abs/2403.19390.

Qin Liu, Fei Wang, Chaowei Xiao, and Muhao Chen. 2024b. From shortcuts to triggers: Backdoor defense with denoised poe. In *NAACL*.

Xiaogeng Liu Liu, Shengshan Hu Hu, Muhao Chen, and Chaowei Xiao. 2023a. Pred: Label-only test-time textual trigger detection. In *EMNLP (in submission)*.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. 2023b. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

QING LYU, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. *ArXiv*, abs/2301.13379.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Wenjie Mo, Jiashu Xu, Qin Liu, Jiongxiao Wang, Jun Yan, Chaowei Xiao, and Muhao Chen. 2023. Test-time backdoor mitigation for black-box large language models with defensive demonstrations. *arXiv preprint arXiv:2311.09763*.

OpenAI. 2022. OpenAI: Introducing chatgpt.

OpenAI. 2023. Gpt-4 technical report.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023a. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *ArXiv*, abs/2305.12295.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023b. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*.

Ajay Patel, Markus Hofmarcher, Claudiu Leoveanu-Condrei, Marius-Constantin Dinu, Chris Callison-Burch, and Sepp Hochreiter. 2024. Large language models can self-improve at web agent tasks. *CoRR*, abs/2405.20309.

Gabriel Perin, Xuxi Chen, Shusen Liu, Bhavya Kailkhura, Zhangyang Wang, and Brian Gallagher. 2024. Rankmean: Module-level importance score for merging fine-tuned LLM models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 1776–1782. Association for Computational Linguistics.

Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. 2022. Grips: Gradient-free, edit-based instruction search for prompting large language models. *arXiv preprint arXiv:2203.07281*.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with" gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*.

Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2021. ONION: A simple and effective defense against textual backdoor attacks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9558–9566, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tobias Schnabel and Jennifer Neville. 2024. Symbolic prompt program search: A structure-aware approach to efficient compile-time prompt optimization.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving AI tasks with chatgpt and its friends in hugging face. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. 2023. GPT-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems. *CoRR*, abs/2310.12397.

Sainbayar Sukhbaatar, Olga Golovneva, Vasu Sharma, Hu Xu, Xi Victoria Lin, Baptiste Rozière, Jacob Kahn, Daniel Li, Wen-tau Yih, Jason Weston, et al. 2024. Branch-train-mix: Mixing expert llms into a mixture-of-experts llm. *arXiv preprint arXiv:2403.07816*.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Daniel Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-agent collaboration: Harnessing the power of intelligent LLM agents. *CoRR*, abs/2306.03314.

Derek Tam, Mohit Bansal, and Colin Raffel. 2024. Merging by matching models in task parameter subspaces. *Trans. Mach. Learn. Res.*, 2024.

Derek Tam, Margaret Li, Prateek Yadav, Rickard Brüel Gabrielsson, Jiacheng Zhu, Kristjan Greenewald, Mikhail Yurochkin, Mohit Bansal, Colin Raffel, and Leshem Choshen. Llm merging: Building llms efficiently through merging. In *NeurIPS 2024 Competition Track*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. 2023. Can large language models really improve by self-critiquing their own plans? *CoRR*, abs/2310.08118.

Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. 2023. Poisoning language models during instruction tuning. *arXiv preprint arXiv:2305.00944*.

Fei Wang, Ninareh Mehrabi, Palash Goyal, Rahul Gupta, Kai-Wei Chang, and Aram Galstyan. 2024a. Data advisor: Dynamic data curation for safety alignment of large language models. In *Proceedings of EMNLP*.

Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. 2023a. A causal view of entity bias in (large) language models. *In submission at EMNLP*.

Fei Wang, Chao Shang, Sarthak Jain, Shuai Wang, Qiang Ning, Bonan Min, Vittorio Castelli, Yassine Benajiba, and Dan Roth. 2024b. From instructions to constraints: Language model alignment with automatic constraint verification. *arXiv preprint arXiv:2403.06326*.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024c. mdpo: Conditional preference optimization for multimodal large language models.

Jiongxiao Wang, Zichen Liu, Keun Hee Park, Muhao Chen, and Chaowei Xiao. 2023b. Adversarial demonstration attacks on large language models. *arXiv preprint arXiv:2305.14950*.

Yiwei Wang, Bryan Hooi, Fei Wang, Yujun Cai, Yuxuan Liang, Wenxuan Zhou, Jing Tang, Manjuan Duan, and Muhao Chen. 2023c. How fragile is relation extraction under entity replacements? In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 414–423.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36.

Fangzhao Wu, Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, and Xing Xie. 2023. Defending chatgpt against jailbreak attack via self-reminder.

Penghao Wu and Saining Xie. 2024. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094.

Jiashu Xu, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. 2024a. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. In *NAACL*.

Nan Xu, Fei Wang, Ben Zhou, Bangzheng Li, Chaowei Xiao, and Muhao Chen. 2024b. Cognitive overload: Jailbreaking large language models with overloaded logical thinking. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3526–3548.

Rongwu Xu, Zehan Qi, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024c. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

Fuzhao Xue, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, and Yang You. 2024. Openmoe: An early effort on open mixture-of-experts language models. *arXiv preprint arXiv:2402.01739*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *ArXiv*, abs/2210.03629.

Qinyuan Ye, Maxamed Axmed, Reid Pryzant, and Fereshte Khani. 2023. Prompt engineering a prompt engineer. *arXiv preprint arXiv:2311.05661*.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. Merging generated and retrieved knowledge for open-domain qa. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Zheng Zhao, Emilio Monti, Jens Lehmann, and Haytham Assem. 2024. Enhancing contextual understanding in large language models through contrastive decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4225–4237.

Ben Zhou, Hongming Zhang, Sihao Chen, Dian Yu, Hongwei Wang, Baolin Peng, Dan Roth, and Dong Yu. 2024a. Conceptual and unbiased reasoning in language models. *arXiv preprint arXiv:2404.00205*.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024b. Universalner: Targeted distillation from large language models for open named entity recognition. In *The Twelfth International Conference on Learning Representations*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*.

Xuekai Zhu, Biqing Qi, Kaiyan Zhang, Xinwei Long, Zhouhan Lin, and Bowen Zhou. 2024. Pad: Program-aided distillation can teach small models reasoning better than chain-of-thought fine-tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2571–2597.

A Appendix

A.1 Past Tutorials by the Instructors

The presenters of this tutorial have given the following tutorials at leading international conferences in the past.

- Wenpeng Yin:
  - EMNLP'23: Learning from Task Instructions.
  - KONVENS'23: Learning from Task Instructions.
  - ACL'23: Indirectly Supervised Natural Language Processing.
- Muhao Chen:
  - ACL'23: Indirectly Supervised Natural Language Processing.
  - NAACL'22: New Frontiers of Information Extraction.
  - ACL'21: Event-Centric Natural Language Processing.
  - AAAI'21: Event-Centric Natural Language Understanding.
  - KDD'21: From Tables to Knowledge: Recent Advances in Table Understanding.
  - AAAI'20: Recent Advances of Transferable Representation Learning.
- Rui Zhang:
  - NAACL'22: Contrastive Data and Learning for Natural Language Processing
- Ben Zhou:
  - ACL'23: Indirectly Supervised Natural Language Processing.
  - NAACL'22: New Frontiers of Information Extraction
- Dan Roth:
  - ACL'23: Indirectly Supervised Natural Language Processing.
  - NAACL'22: New Frontiers of Information Extraction.
  - ACL'21: Event-Centric Natural Language Processing.
  - AAAI'21: Event-Centric Natural Language Understanding.
  - ACL'20: Commonsense Reasoning for Natural Language Processing.
  - AAAI'20: Recent Advances of Transferable Representation Learning.
  - ACL'18: A tutorial on Multi-lingual Entity Discovery and Linking.
  - EACL'17: A tutorial on Integer Linear Programming Formulations in Natural Language Processing.
  - AAAI'16: A tutorial on Structured Prediction.
  - ACL'14: A tutorial on Wikification and Entity Linking.
  - AAAI'13: Information Trustworthiness.
  - COLING'12: A Tutorial on Temporal Information Extraction and Shallow Temporal Reasoning.
  - NAACL'12: A Tutorial on Constrained Conditional Models: Structured Predictions in NLP.
  - NAACL'10: A Tutorial on Integer Linear Programming Methods in NLP.
  - EACL'09: A Tutorial on Constrained Conditional Models.
  - ACL'07: A Tutorial on Textual Entailment.

A.2 Recommended Paper List

The following is a reading list that could help provide background knowledge to the audience before attending this tutorial:

- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv
- Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023a. Teaching large language models to self-debug. ArXiv
- Zhoujun Cheng, Jungo Kasai, and Tao Yu. 2023. Batch prompting: Efficient inference with large language model apis. CoRR, abs/2301.08721
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT
- Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. CRITIC: large language models can self-correct with tool-interactive critiquing. CoRR, abs/2305.11738

- Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. arXiv preprint arXiv:2301.00303
- Yujin Huang, Terry Yue Zhuo, Qiongkai Xu, Han Hu, Xingliang Yuan, and Chunyang Chen. 2023. Training-free lexical backdoor attacks on language models. In Proceedings of the ACM Web Conference 2023
- Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, and Yangqiu Song. 2023a. Multi-step jailbreak- ing privacy attacks on chatgpt. arXiv
- Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raf- fel, and Mohit Bansal. 2023. Resolving interference when merging models. CoRR, abs/2306.01708
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge conflicts. arXiv preprint arXiv:2305.13300
- Yashar Talebirad and Amirhossein Nadiri. 2023. Multi- agent collaboration: Harnessing the power of intelli- gent LLM agents. CoRR, abs/2306.03314.