On the Shoulders of LLMs: From LLM Optimization to LLM Agents

Name(s) and Affiliation(s) of the Lecturer(s):

Zuchao LI	Zhuosheng ZHANG	Yao YAO
WUHAN UNIVERSITY	**Shanghai Jiao Tong University**	**Shanghai Jiao Tong University**
Wuhan University, 299# Bayi Road	Shanghai Jiao Tong University, 800# Dong Chuan Road	Shanghai Jiao Tong University, 800# Dong Chuan Road
Wuhan, Hubei, China	Shanghai, China	Shanghai, China
zcli-charlie@whu.edu.cn	zhangzs@sjtu.edu.cn	yaoya027@sjtu.edu.cn
+86-18073461585	+86-18916892387	+86-18715311890

Tutorial Background and Objectives:

In the following paragraph, please clearly specify the background and objectives of your tutorial. What is the topic? Why is it necessary and timely? What are the specific learning outcomes / skills that participants can expect to take away?

Background and Objectives:

The rapid advancement of Large Language Models (LLMs) has propelled significant progress. Both academia and industries ranging from healthcare to finance are seeking to harness Artifial Intelligence (AI) for innovation and efficiency. Therefore, understanding and leveraging LLMs is crucial. As LLMs becomes increasingly embedded in critical applications, the need for well-rounded LLM professionals who can navigate both the technical and ethical landscapes of LLM deployment is more pressing than ever. This tutorial builds on the foundational knowledge of LLMs to address these urgent needs by equipping participants with cutting-edge skills and deepen their understanding of these powerful tools.

The tutorial will start with the basic architecture of LLMs and move on to advanced optimization techniques. Participants will learn how to enhance the performance of LLMs during the inference stage by optimizing KV cache for quicker response times, extended memory capabilities, and improved answer quality. Furthermore, the tutorial will cover diverse LLM reasoning method including, Chain-of-Thought (CoT) reasoning, to enhance the interpretability, controllability, and flexibility of LLMs.

With a robust understanding of LLM enhancements, we turn to the emerging field of language agents powered by LLMs. The emergence of LLMs has significantly accelerated the evolution of AI agents, pushing closer to the long-standing goal of building intelligent, autonomous agents that can learn and act in distinct environments. This session will guide attendees through the concepts of agents, how LLMs empower these agents, and the challenges these agents might face in the future. Special focus will be given to the safety of LLM agents. We will explore threats and risks confronting LLM agents in diverse application environments aiming to equip participants with the knowledge to effectively implement and manage LLMs in a responsible and efficient manner. (to be paraphrased by Yuan)

We hope that this tutorial will enable participants to stand on the shoulders of Large Language Models and look further into the future of Artificial Intelligence.

Learning Outcomes:

Participants can expect to leave the tutorial with a deeper understanding of LLM structures and optimization strategies, skills in enhancing LLM capabilities through advanced techniques like KV cache optimization, CoT reasoning, insights into the development and empowerment of AI agents using LLMs, and best practices for ensuring the safety and efficiency of these agents. This knowledge is essential for advancing AI applications and developing robust, safe AI systems in various industries.

Tutorial Resources:

List a webpage where participants may be able to access materials for your tutorial.

[1] Igniting Language Intelligence: The Hitchhiker's Guide From Chain-of-Thought Reasoning to Language Agents https://arxiv.org/pdf/2311.11797.pdf

[2] R-Judge: Benchmarking Safety Risk Awareness for LLM Agentshttps://arxiv.org/pdf/2401.10019.pdf

[3] Prioritizing Safeguarding Over Autonomy: Risks of LLM Agents for Science https://arxiv.org/abs/2402.04247

[4] You Only Look at Screens: Multimodal Chain-of-Action Agentshttps://arxiv.org/abs/2309.11436

[5] Multimodal Chain-of-Thought Reasoning in Language Models

https://arxiv.org/abs/2302.00923

[6] Automatic Chain of Thought Prompting in Large Language Models

https://arxiv.org/abs/2210.03493

[7] Identifying the Risks of LM Agents with an LM-Emulated Sandbox

https://arxiv.org/abs/2309.15817

[8] A Trembling House of Cards? Mapping Adversarial Attacks against Language Agentshttps://arxiv.org/abs/2402.10196

[9] Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations

https://arxiv.org/abs/2312.06674

[10] ShieldLM: Empowering LLMs as Aligned, Customizable and Explainable Safety Detectors https://arxiv.org/abs/2402.16444

[11] Efficient Memory Management for Large Language Model Serving with PagedAttention https://dl.acm.org/doi/abs/10.1145/3600006.3613165

[12] GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints https://aclanthology.org/2023.emnlp-main.298/

Target Audience:

In the following paragraph, please describe the target audience for your tutorial. If participants need to have specific prerequisites, please outline them clearly here.

This tutorial is designed for a broad audience across various domains within artificial intelligence, especially those interested in Natural Language Processing (NLP) and its practical applications. It is particularly targeted at researchers, developers, and students from both academia and industry who are eager to leverage, explore, and optimize large language models (LLMs).

Participants are expected to have a foundational understanding of NLP and machine learning, which should include:

- Machine Learning: Basic knowledge of probability theory, supervised learning, and transformer models.
- NLP: Familiarity with LLMs, including techniques for prompt tuning and generative NLP.

Tutorial Outline:

Introduction: (Estimated time: 5 minutes)

- Introduction, part 1: Welcome and opening remarks
  - Briefly introduce the topic and its significance
  - Highlight the growing importance of Language Models (LLMs) in various fields
  - Engage the audience by expressing the potential of LLMs
- Introduction, part 2: Understanding Language Models (LLMs)
  - Provide a brief explanation of what LLMs are and how they work

- Discuss the advancements in LLMs, such as GPT-3.5/GPT-4 architecture
- Explain the role of LLMs in natural language processing and generation

Overview: (Estimated time: 5 minutes)

- Overview, part 1: History and evolution of language models
  - Explore the timeline of language models, from early approaches to LLMs
  - Discuss key milestones and breakthroughs in the field
  - Highlight the impact of LMs on natural language understanding and generation
- Overview, part 2: Key components and architecture of LLMs
  - Explain the fundamental components of LLMs, such as attention mechanisms
  - Describe the GPT-3.5 architecture and its improvements over previous versions
  - Discuss the underlying principles that make LLMs powerful language processors

Tutorial Main Part, Theme 1: LLMs Optimization and Inference (Estimated time: 30 minutes)

- LLM Inference Part 1: Auto-regressive Decoding
  - Introduce the concept of Auto-regressive Decoding.
  - Examine the key property that a language model must have to performe text generation task efficiently.
- LLM Inference Part 2: Efficency of Pre-trained Language Model
  - Define metrics to measure the efficeny of a generative language model.
  - Introduce the GQA as an efficient Transformer-Decoder modification.
  - Explore different open-source LLMs and compare their effiecency.
- LLM Inference Part 3: Inference System with Paged Attention
  - Discuss the challenge when deploying KV-Cache.
  - Highlight the high performance of Paged Attention.
- LLM Inference Part 4: Instant Modification to KV-Cache
  - Introduce the eviction and quantization paradigms for KV-Cache.
  - Discuss the trade-offs of utilizing instant modifications.
- LLM Inference Part 5: Conclusion and Vision
  - Summerize key methods to build a efficient inference system.
  - Vision the future directions to improve LLM inference efficiency.

Tutorial Main Part, Theme 2: LLM Reasoning (Estimated time: 30 minutes)

- LLM Reasoning Part 1: What is reasoning?
  - Introduce the concept and definition of reasoning.
  - Explore different types of reasoning utilized across various contexts.
- LLM Reasoning Part 2: Reasoning in LLMs

- Discuss how LLMs perform reasoning
- Highlight recent progress in LLM reasoning

- LLM Reasoning Part 3: Chain-of-Thought reasoning in LLMs
  - Define Chain-of-Thought reasoning in LLMs.
  - Examine paradigm shifts of CoT
  - Discuss scenarios where CoT is particularly effective.
  - Explore the underlying mechanisms that make CoT effective.
- LLM Reasoning Part 4: Conclusion and vision for LLM reasoning
  - Summarize key takeaways of LLM reasoning.
  - Share insights and future perspectives on the evolution of reasoning in LLMs.

Tutorial Main Part, Theme 3: LLM Agent and Safety (Estimated time: 45 minutes)

- Part 1: Introduction to LLM Agents
  - Introduce definition, applications and classfication of LLM Agents.
- Part 2: Architecture & Key Technique
  - Introduce the general architecture of LLM agents, including modules and techniques.
  - Explore the key technique CoT reasoning on how it effectively unlock capabilities of LLMs to develop agents.
  - Explore the key technique tool learning on how it facilitate the shortcomings of LLMs to develop agents.
  - Explore how multimodal ability extend perception of LLMs to develop agents.
- Part 3: Auto-UI: Multimodal Chain-of-Action Agents
  - Introduce a First Principle-based method to develop multimodal agents.
  - Discuss effects and give analysis on the method.
- Part 4: Challenge and Prospects
  - Summarize challenges and prospects of research on LLM agents, including environment adaptation, capability enhancement and safety guardrail.

- Part 5: Challenge and Prospects
  - Discuss possible ways to facilitate agent safety: Identify and mitigate agent risks by Red-teaming.
  - Discuss possible ways to strengthen agent safety: Develop techniques and mechanisms for agent alignment and regulation.

Conclusion, Discussion, and Q&A (Estimated time: 5 minutes)

- Recap the key points discussed throughout the tutorial

- Emphasize the importance of LLM optimization for advancing natural language processing
- Highlight the potential of LLM agents in various fields and applications
- Encourage further exploration and research in LLM optimization and LLM agent development
- Q&A

Provisional Schedule of the Tutorial [note, most likely Tutorials will be on July 3rd, Wednesday, however, this is subject to change]:

Schedule:

Introduction / Overview / Theme 1 / Theme 2

Break

Theme 3 / Conclusions / Q&A

About the Lecturers:

	**Zuchao LI** is an associate professor at Wuhan University. He got the PhD degree from Department of Computer Science and Technology, Shanghai Jiao Tong University, advised by Prof. Hai Zhao. He has visited National Institute of Information and Communications Technology (NICT) as Limited Technical Researcher, Japan. His research interests including the language sequence modeling, linguistic structure parsing, and language representation learning from various types of data (unlabeled or noisy data, structured data like tree, graph, etc.). Specifically, he focus on theoretical and algorithmic approaches for language models, self-supervised/weakly supervised learning, structure learning, and related NLP tasks.
	**Zhuosheng ZHANG** is a tenure-track assistant professor at Shanghai Jiao Tong University, China. His research interests include NLP, LLMs, and autonomous agents. He has published over 50 papers in top-tier conferences and journals, including TPAMI, ICLR, ACL, AAAI, EMNLP, TNNLS, TASLP, and COLING. He has won 1st place in various language understanding and reasoning leaderboards, such as HellaSwag, SQuAD2.0, MuTual, RACE, ShARC, and CMRC. He has several tutorials at international conferences, including IJCAI 2021, IJCNLP-AACL 2023, LREC-COLING 2024, and CVPR 2024. Homepage: https://bcmi.sjtu.edu.cn/~zhangzs.
	**Yao YAO** is a Ph.D. student with the Center for Brain-like Computing and Machine Intelligence of Shanghai Jiao Tong University, Shanghai, China, coadvised by Prof. Zuchao Li and Prof. Hai Zhao. She received the bachelor's degree from Southeast University, Nanjing, China, in 2022. Her research interests mainly include natural language processing, especially lie in Large Language Model, Chain-of-Thought Reasoning.