Towards Effective and Efficient Multi-Agent Language Model Systems: Foundations, Prospects, and Applications

Xuan Wang, Shuxiang Cao, Yuchen Zhuang, Wenqi Shi

https://xuanwang91.github.io/2026-07-02-acl26-tutorial

Multi-agent systems powered by large language models (LLMs) offer a promising paradigm for tackling complex reasoning, decision-making, and problem-solving tasks. However, achieving both effectiveness and efficiency in such systems remains a critical challenge. This tutorial introduces recent advances in building effective and efficient multi-agent LLM systems, focusing on three core components. First, we discuss the design of individual LLM agents. We present state-of-the-art techniques for enabling capable agents using efficient and compact LLMs, including model distillation, dynamic routing, and memory- and compute efficient serving, providing a foundation for scalable and responsive agent design under resource constraints. Second, we cover coordination and communication among agents, crucial for collective performance, highlighting methods for improving multi-agent reasoning and decision-making through prompt and graph optimization, sycophancy mitigation, and structured LLM-based frameworks. Last, we explore real-world applications of LLM agents in areas such as industry, healthcare, quantum computing, and various scientific domains.

Xuan Wang

email: xuanw@vt.edu

website: https://xuanwang91.github.io/

Dr. Xuan Wang is an Assistant Professor in the Department of Computer Science at Virginia Tech. Her research interests are in natural language processing, data mining, AI for sciences, and AI for healthcare. Xuan was a recipient of the NSF CAEER Award 2025, Cisco Research Award 2025, NSF NAIRR Pilot Award 2024-2025, and NAACL Best Demo Paper Award 2021. Xuan has served as a Program Chair of the SouthNLP Symposium 2024 (>150 participants), from more than 20 universities across the USA. She has also served as a Program Chair for the Undergraduate and High School Symposium at IEEEBigData 2024 and IEEE-ICDM 2025. She has also served as a Senior Area Chair, Area Chair, and Program Committee in major AI conferences (e.g., ARR, ACL, EMNLP, NAACL, NeurIPS, ICLR, KDD). Xuan has delivered tutorials in AAAI 2025, EMNLP 2024, KDD 2022, TheWebConf 2022, and IEEE-BigData 2019.

Shuxiang Cao

email: shuxiangc@nvidia.com

website: https://www.physics.ox.ac.uk/our-people/caos

Dr. Shuxiang Cao is a Senior Research Scientist at NVIDIA and a Long Term Visitor in the Department of Physics at the University of Oxford. His research is at the intersection of quantum technology, artificial intelligence, and high-performance computing. He obtained his PhD in Physics (Quantum Technology) from the University of Oxford, where he worked on superconducting quantum computers. At NVIDIA, his work focuses on advancing artificial intelligence for science, including the development of large language model applications for scientific research and autonomous laboratories that accelerate scientific discovery. Dr. Cao has published in high profile journals in physics, including Physical Review Letters, Physics Reports, and Science Advances, as well as at computer science conferences, including NeurIPS, ACL, and EMNLP.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts), pages 5–6

Yuchen Zhuang

email: yczhuang@google.com

website: https://night-chen.github.io/

Dr. Yuchen Zhuang is a Research Scientist at Google DeepMind. His research focuses on language intelligence, with the goal of developing LLM-based agents that can show human-like reasoning and planning on challenging real-world tasks. His experience covers model pre-training, instruction fine-tuning, reinforcement learning from human feedback (RLHF), and evaluation. He has published more than 30 papers in top AI venues such as NeurIPS, ICLR, ICML, ACL, EMNLP, NAACL, UAI, and KDD. He has received the 2024 J.P. Morgan PhD Fellowship Award, the NeurIPS 2023 Scholar Award, and the Best SIGBio Paper Award at ACM BCB. He serves as an Area Chair and Program Committee member in major AI conferences (e.g., ARR, ACL, EMNLP, NAACL, NeurIPS, ICLR, ICML).

Wenqi Shi

email: wenqi.shi@utsouthwestern.edu

website: https://wshi83.github.io/

Dr. Wenqi Shi is an Assistant Professor in the Department of Health Data Science and Biostatistics at the University of Texas Southwestern Medical Center. Her research focuses on the intersection of AI and healthcare, advancing both foundational algorithms and agentic systems for precision and personalized medicine. Her recent work includes developing LLMs for translational medicine, advancing agentic AI for biomedical discovery, and promoting responsible AI practices to improve clinical research and care. She has received the NVIDIA Academic Grant Program Award (2025), the Texas Advanced Computing Center (TACC) Pilot Award (2025), the Best SIGBio Paper Award at ACM BCB (2023), and recognition as a Rising Star in EECS (2023). She has served as Finance Co-Chair of ACM BCB 2025, Publication Co-Chair of IEEE BHI 2025 (both >300 participants), and as an Area Chair or Program Committee member in major AI and healthcare conferences (e.g., ACL, EMNLP, NAACL, NeurIPS, ICLR, ICML, BCB, BIBM).