



Meta-Model: Making Many Models Act as One

SPEAKER Bowei He

Postdoctoral Fellow
Mohamed bin Zayed University of
Artificial Intelligence (MBZUAI)

DATE 23 Oct, 2026 (Fri)

TIME 10:00 AM - 11:00 AM

VENUE CS Seminar Room, Y6405, 6th Floor,
Yellow Zone, Yeung Kin Man Academic
Building, City University of Hong Kong, 83
Tat Chee Avenue, Kowloon Tong

ABSTRACT

Foundation models differ in expertise, quality, latency, and cost, yet users typically commit to a single model that may not meet their needs. This talk introduces the Meta-Model: a single interface that routes requests to the model best suited to the task and user preferences, or composes a tailored Mixture-of-Models when no standalone model suffices. Such compositions can achieve trade-offs unattainable by individual models and, with suitable architectures and collaboration protocols, outperform the strongest frontier model using individually weaker models. I will present vLLM Semantic Router, our open-source framework for realizing this vision, and Vela, its family of open routing models. The project has surpassed 1.3 million downloads in six months, with adoption by AMD, UBS, Intel, and Dell; Vela alone received over 150,000 downloads in the past month. I will then discuss two complementary studies: optimizing which models to compose, and predicting cross-model KV cache migration effectiveness to guide scheduling decisions between migration and recomputation. I will conclude with a research agenda toward Meta-Models trained, evaluated, and served as a single model.

BIOGRAPHY

Bowei He is a Postdoctoral Fellow at Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), working with Prof. Xue (Steve) Liu. His research focuses on AI agents, mixtures of models, and recursive self-improvement. He has published at top-tier AI conferences, with multiple highlight and oral presentations, and collaborated with industry research teams including Microsoft, AMD, NVIDIA, Meta, and Tencent, contributing to technologies deployed on platforms serving over 100 million users. He is the research lead for vLLM Semantic Router, a project with over 1.3 million downloads, and has served as a senior program committee member or area chair for ICLR, AAAI, and COLM. His honors include the WAIC Academic 2026 Best Paper Award Runner-Up, the ACM MM WIPAI Distinguished Paper Award, and selection as a MobiSys 2026 Rising Star and Heidelberg Laureate Forum Young Researcher.

All are welcome!



In case of questions, please contact Prof MA Chen at chenma@cityu.edu.hk, or visit the CS Departmental Seminar Web at <https://www.cs.cityu.edu.hk/events/cs-seminars/recent-cs-colloquiums>.