香港浸會大學 HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE 計算機科學系

DEPARTMENT OF COMPUTER SCIENCE

SEMINAR

2024 SERIES

InfiAgent: A Multi-Tool Agent for AI Operating Systems

DATE & TIME

20 MAR 2024 (WED) 10:00 – 11:00 AM

VENUE

Dr. Wu Yee Sun Lecture Theatre, WLB109, Shaw Campus

DR. HONGXIA YANG

Head of Large Language Model
ByteDance, USA

ABSTRACT

Following the launch of GPT4-Agent, GPT4 has demonstrated its flexibility in utilizing tools like Advanced Data Analytics (ADA, previously known as code interpreter) and DALL-E3, although the details of GPT4-Agent have not been fully disclosed. Over the past years, we have intensively studied the core functionalities of GPT4, progressively developing a system comparable to GPT4-Agent, named InfiAgent. Initially, we replicated Codex and discovered that while existing models such as CodeLlama, StarCoder, and WizardCoder excel in programming capabilities, they fall short in handling FreeformQA problems for coding. To address this, we created InfiCoder—the first open-source model capable of handling text-to-code, code-to-code, and freeform code-related QA tasks simultaneously. Building on this, we developed InfiCoder-Eval (FreeformQA benchmark), which includes 270 high-quality automated test questions. Our findings indicate that even GPT4 has room for improvement in this area (achieving a score rate of only 59.13%). Based on InfiCoder, we launched the InfiAgent framework, focusing on the field of data analysis. This framework first defines the problem framework and evaluation objectives for data analysis. Then, in line with the data analysis scenarios, we developed a specialized Agent system based on the React format and LLM, effectively addressing data analysis challenges. This system integrates an LLM with programming capabilities and a sandbox environment for executing Python code, generating solutions and corresponding code through multiple rounds of dialogues. It is the industry's first Agent framework closest to the capabilities of ADA. Additionally, we expanded the application scenarios of InfiAgent, multimodal LLM (MLLM) reasoning tool InfiMM, achieving excellent results. Among the open-source models, InfiMM performs the best on the MMMU leaderboard with the smallest size of only 7B. Particularly in MLLM reasoning, we found that there is significant room for improvement in the current GPT4V (achieving a score rate of only 74.44%). These achievements not only reveal the tremendous potential of InfiAgent but also showcase our possible directions in surpassing the capabilities of GPT4.

SPEAKER'S BIOGRAPHY

REGISTER NOW

Enquiries: 3411-2385 **Email:** comp@comp.hkbu.edu.hk **Website:** https://bit.ly/bucs-events