THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Department of Computing
電子計算學系

RESEARCH SEMINAR

Insights and Innovations: Building Large Language Models and Path Forward to Surpass OpenAI

Dr Hongxia YANG
ByteDance
USA

Date : 22 March 2024 (Fri)
Time : 11:00 am - 12:00 nn
Venue : FJ302

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.

About the Speaker

Dr Hongxia Yang, PhD from Duke University, has published over 100 papers in top-tier conferences and journals, and holds more than 50 patents in the USA and China. Her contributions to the field have been recognized through numerous prestigious awards, including the coveted Super AI Leader (SAIL) Award at the 2019 World Artificial Intelligence Conference, the Second-Class National Science and Technology Progress Award in 2020, which is one of China's highest technological honors, the First-Class Science and Technology Progress Award from the Electronics Society in 2021 and the First-Class Science and Technology Progress Award from the Ministry of Education in 2022. Forbes China lauded her as one of the Top 50 Women in Tech in 2022, a testament to her trailblazing role in the tech industry. Currently, she serves as the Head of Large Models at ByteDance US. Previously, she worked as a research staff member at IBM T.J. Watson Research Center, principal scientist at Yahoo!, an AI scientist and director at Alibaba DAMO Academy, and an adjunct professor at Zhejiang University's Shanghai Advanced Research Institute.