Murong Yue

⌂ 2251 Pimmit Dr, Falls Church, VA 22043 ✉ myue@gmu.edu ☎ (703) 975-8878 🌐 personal webpage

RESEARCH INTEREST

Natural Language Processing, Large Language Model, Interactive AI

EDUCATION

George Mason University 2022 — present
Ph.D.(Advised by Prof. Ziyu Yao) in Computer Science

University of Southern California 2018 — 2020
M.S. in Electrical and Computing Engineering

Beijing Jiaotong University 2014 — 2018
B.E. in Electrical and Computing Engineering

PUBLICATION

Murong Yue, Jie Zhao, Min Zhang, Liang Du, Ziyu Yao, ”Large Language Model Cascades with Mixture of Thoughts Representations for Cost-efficient Reasoning”, **Preprint**[paper](Featured in Hugging Face Daily Papers).

Binfeng Xu, Xukun Liu, Hua Shen, Zeyu Han, Yuhan Li, **Murong Yue**, Zhiyuan Peng, Yuchen Liu, Ziyu Yao, Dongkuan Xu, “Gentopia: A Collaborative Platform for Tool-Augmented LLMs”, **EMNLP 2023 Demo**.[paper]

EXPERIENCE

George Mason University Fairfax, VA
Graduate Research Assistance 2023-present

- Explore the practical, task-oriented interactive large language model systems.
- Develop the cost-efficient LLM cascades to reduce the cost. Our cascade considers the “answer consistency” of the weaker LLM as a signal of the question difficulty to dynamically select the LLM version. We leveraged a mixture of two thought representations in sampling and achieved a comparable result with fully using the GPT-4 but only 40% cost in multiple reasoning tasks.
- Participated in the construction of the Gentopia, which enabled us to create a context/distribution shift specialized to some target tasks easily. We focus on furnishing essential components for the construction, experimentation, and evaluation of LLM agents.

Graduate Teaching Assistance 2022-2023

- Teaching Assistant for CS110 Essentials of Computer Science and CS112 Introduction to Computer Programming

Alibaba Group, TmallGenie AI labs Beijing, China
Machine Learning Engineer 2020 - 2022

- Responsible for vocal search in voice assistant. Designed a cross-modal retrieval method based on contrastive learning connecting the text and music. Our innovation approach achieved zero-shot capabilities in tagging the music and was applied in the music scene with greater than 1M daily active users.
- Responsible for dialect classification in voice assistant. Designed a dialect classification system with an early classification module based on curriculum learning.

Pingan Technology, AI Labs Beijing, China
Research Intern 2019

- Trained the TDNN neural network to extract the features of speakers and designed a speaker verification method based on the extracted speaker features. The error rate decreased from 13.7% to 8.4%.

PROFESSIONAL SERVICE

Organizing Committee

MASC-SLL 2023 - 10th Mid-Atlantic Student Colloquium on Speech, Language and Learning

Reviewer

Served as a reviewer for NeurIPS’23

Invited Talks

2023 Large Language Model Cascades with Mixture of Thoughts Representations for Cost-efficient Reasoning, Invited talk at **Microsoft Semantic Machines**. Host: Hao Fang.

HONOR AND AWARD

2016 BJTU Study Scholarship
2014 First Prize in BJTU Mathematical Modeling Competition