PETER (HAO) YU

Montreal, Quebec, Canada

✉ hao.yu2@mail.mcgill.ca   in haoy   ⓞ YHPeter   k yupeter (**Competition Expert**)

Homepage (yhpeter.github.io)   Google Scholar

Education

McGill University	*Montreal, Quebec, Canada*
Software Engineering B.S. 3.82/4.00	**Sep 2021 – May 2024**
Computer Science M.S. (Thesis) Supervised by Prof. David Adelani	**Sep 2024 – Exp. Jan 2026**

Publications

[1] **Hao Yu**, et al. *Evaluation of **Retrieval-Augmented Generation**: A Survey*, **CCF BigData 2024**, arXiv:2405.07437.

[2] Jacob-Junqi Tian, **Hao Yu**, et al. *Web Retrieval **Agents** for Evidence-Based Misinformation Detection*, **COLM 2024**.

[3] Javin Liu*, **Hao Yu***, Vidya Sujaya*, et al. *SWEET: **Weakly Supervised** Person Name Extraction for Fighting Human Trafficking*, Student Research Workshop Poster in ACL, Acceptance in Findings of the Association for Computational Linguistics: **EMNLP 2023**.

[4] **Hao Yu***, Zachary Yang*, et al. *Open, Closed, or Small Language Models for Text Classification?* In: arXiv:2308.10092 (2023).

Work Experience

McGill University Teaching Assistant	*Montreal* Contract **Sep 2024 – Currently**

TA for COMP 202 (Foundations of Programming, Python) and COMP 206 (Introduction to Software Systems, Linux/Bash)

SmartPrep.ai Full-Stack Developer	*Remote* Part-time **Sep 2023 – Currently**

Develop **ThetaWave.ai**, the React-based learning platform hosted on Vercel, enabling student note-taking, lecture transcribing and interactive podcast engagement. Designed and implemented comprehensive Retrieval-Augmented Generation (**RAG**) systems integrating user-uploaded content, AI-generated insights, and web search results using AWS Lambda. Developed backend services with FastAPI, optimized deployment with a CI/CD pipeline, and ensured system performance monitoring through **Prometheus and Grafana**.

Mila - Quebec AI Institute Researcher	*Montreal* Part-time, Affiliated **Oct 2022 – Currently**

From 2024 Fall, supervised by Prof. David Ifeoluwa Adelani on **multilingual** language processing and low-resource languages. From 2022 Fall, supervised by Prof. Reihaneh Rabbany in detecting **misinformation detection with RAG**.

Tencent Research Intern	*Shanghai* Full-Time **May 2024 – Aug 2024**

Research in the Retrieval-Augmented Generation (RAG) with LLMs, focusing on in-game interaction with players.

Tencent Research Intern	*Remote* Part-time **Nov 2023 – Feb 2024**

Explore efficient training and inference methods for large language models (LLMs), including GPTQ, ZO, LoRA, etc..

Morgan Stanley Summer Analyst	*Montreal* Full-time Contract **May 2023 – Aug 2023**
Drakkar Capital Hedge Fund Programmer	*Montreal* Full-time Contract **May 2022 – Aug 2022**

Awards and Competitions

15,000 CAD, **Research Funding from Quebec Gov for Misinfor Detection with RAG**	**Dec 2024**
29,000 CAD plus Tuition, **McGill Master of Computer Science (Thesis) Full Scholarship**	**Mar 2024**
7000 CAD, **Science Undergraduate Research Awards (2023) by McGill University**	**Apr 2023**
4th place, **NeurIPS 2022, OGB-LSC PCQM4Mv2 Predict Quantum Property of Molecular Graphs**	**Oct 2022**
Top 2%, rank 71/3204, **Wechat Big Data Challenge, Multimodal Video Classification**	**Jul 2022**
Silver, rank 46/1505, Team Leader, **Kaggle Sartorius Cell Instance Segmentation**	**Dec 2021**
First prize, top 0.7%, rank 20/2849, Team Leader, **AI-Earth AI Innovation Challenge**	**Feb 2021**
First place, rank 1/1619, **Global AI Technology Innovation Competition Warmup Track**	**Jan 2021**

Skills

Programming Languages: Python (proficient), **SQL**/C++/C#/Java/JavaScript (intermediate), C/OCaml (beginner)

Machine Learning Related: Pandas, Numpy, PyTorch, TF2, HuggingFace, LangChain, LlamaIndex, vLLM

Frameworks: Docker, Git, FastAPI, MongoDB, DB2, PostgreSQL, AWS, Github Action, Prometheus, Grafana, SLURM