

HAO YU

✉ hao.yu2@mail.mcgill.ca  haoy  YHPeter  yupeter

🏠 yhpeter.github.io  (700+ citation, h-index 8)

Globally Consider, Canada (Valid Work Permit), US (Visa Sponsor), UK (HPI Visa), China

TL;DR

Work Style: Figure out requirements & Abstract correctly & Just solve it.

Research: 2 ACL first-author Oral papers (top 1%); 700+ Google Scholar citations

Modeling: SOTA African LLM (5K+ downloads Huggingface downloads), Enterprise Model Finetuning

Education: Mila & McGill Master's (4.0/4.0) Computer Science, Research-based

Agents: Lead **multimodal evaluation** framework and **Agentic Memory** for realtime audio chat with self-evolving

Education

McGill University

Montreal, Quebec, Canada

B.S. Software Engineering, with Prof. Reihaneh Rabbany 3.82/4.00

Sep 2021 – May 2024

Retrieval Augmented Generation (RAG), Evaluation, Rerank, Misinformation

M.S. Computer Science (Thesis Based) with Prof. David Adelani 4.00/4.00

Sep 2024 – May 2026

Multilingual, Machine Translation, Data Mixture, Continue-Pretrain, SFT, DPO, GRPO, RL

Work Experience

▷ Boson AI Member of Technical Staff

Toronto Permanent May 2025 – Currently

Build **multi-modal evaluation framework** for text and **audio** models (ASR, Understanding, Full-Duplex) across 60+ benchmarks with official score calibration. Build the agentic workflow for adding benchmarks, calibration and debugging. Lead **Memory Layer/Connector** for Agentic Audio/Video Chat, including full-stack multi-layer knowledge management.

▷ Mila - Quebec AI Institute Researcher

Montreal Part-time, Affiliated Oct 2022 – Currently

From 2022 Fall, sponsored by Prof. Reihaneh Rabbany in detecting **misinformation detection with RAG**.

From 2024 Fall, sponsored by Prof. David Adelani on **machine translation (MT)**, **multilingual math reasoning (MMR)**.

Training base model with multiple stages: data curation, data mixture, continue pretrain (CPT), SFT, DPO.

▷ SmartPrep.ai Machine Learning Engineer

Remote Part-time Sep 2023 – May 2025

Develop *ThetaWave.ai*, the all-in-one note-taking and study platform powered with AI and RAG. Designed and built **RAG** Chat and quality evaluation systems **from scratch** based on user-uploaded content and history dialogue with **NextJS** and **PostgreSQL**. Build the releasing CI/CD pipeline and the monitoring system through Prometheus and Grafana.

▷ Tencent Research Intern

Shanghai Full-Time May 2024 – Aug 2024

Research in **Retrieval-Augmented Generation evaluation** and **agents** for analyzing players in-game interactions.

Developed a **diffusion transformer (DiT)** based time series forecasting model for game sales pricing prediction that easily integrates human domain expertise to adjust automated predictions without training into data agent workframe.

▷ 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

Publications 📖 (700+ citation, h-index 6)

- [1]Hao Yu, ... “*AfriqueLLM: How Data Mixing and Model Architecture Impact CPT for African Languages*”, ACL 2026 Oral.
- [2]Hao Yu*, ... “*Search-on-Graph: Iterative Informed Navigation for LLM Reasoning on Knowledge Graphs*”, KDD 2026.
- [3]Hao Yu, ... “*TRUTH: Teaching LLMs to Rerank for Truth in Misinformation Detection*”, COLM, Workshop, 2025.09.
- [4]Aoran Gan, Hao Yu, ... “*Retrieval Augmented Generation Evaluation ... Comprehensive Survey*”, 2025.04.
- [5]Hao Yu, ... “*INJONGO: A Multicultural Intent Detection ... Dataset for African Languages*”, ACL 2025 Oral, 2025.02.
- [6]Hao Yu, ... *How to Unlock Time Series Editing? A Diffusion-Driven Multi-Grained Control*”, Arxiv, 2025.01.
- [7]Hao Yu, ... “*Evaluation of Retrieval-Augmented Generation: A Survey*”, CCF BigData 2024, Springer Published.
- [8]Jacob-Junqi Tian, Hao Yu, ... “*Web Retrieval Agents for Evidence-Based Misinformation Detection*”, COLM 2024.
- [9]Hao Yu*, ... “*Weakly Supervised Person Name Extraction for Fighting Human Trafficking*”, Findings, EMNLP 2023.

Skills

Most Useful: Claude Code, Copilot, Planing Mode+Ecution Mode, SKILLS, Agents, daily token consumption 10MT, exponentially increasing, System Design, Toward Endless Optimization in System Level to Drain any possible speedup

AI: PyTorch, HuggingFace, vLLM, SGLang, LangChain, LangGraph, pd, np, TF2, LlamaIndex, Megatron, NeMo-RL
FSDP 2, Deepspeed, Llamafactory, AReal, Miles, TRL

Tools: SLURM (up to 16 nodes H100s), PostgreSQL, Elastic Search, MongoDB, Docker, Git, AWS, Prometheus

Languages: Python/JavaScript/React/NextJS/SQL/C++/C#/Java/C/OCaml

Awards and Competitions

\$15,000, Research Funding from Canada First Research Excellence Fund	Dec 2024
Tuition plus \$29,000, Full Scholarship, McGill Master of Computer Science (Thesis)	Mar 2024
750 CAD, Retail GenAI Hackathon Award by Bensadoun School of Retail Management, McGill Uni.	Oct 2023
\$7,000 CAD, Science Undergraduate Research Awards (2023) by McGill University	Apr 2023
4 th 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
Top 1.5% in Kaggle competitions, highest rank 2660/185,584, Kaggle Expert	Apr 2022
Bronze, rank 86/1588, Kaggle Happywhale - Whale and Dolphin Identification	Apr 2022
Bronze, rank 159/2058, Team Leader, Kaggle Feedback Prize - Evaluating Student Writing	Mar 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
Global AI Technology Innovation Competition Track I, Top 5%, rank 198/4337	Jan 2021
First place, rank 1/1619, Global AI Technology Innovation Competition Warmup Track	Jan 2021