Ye Yuan

ye.yuan3@mail.mcgill.ca | (438) 351-1806 | GitHub | LinkedIn | Google Scholar

EDUCATION

McGill University & Mila – Quebec AI Institute Montreal, Quebec, Canada
*Doctor of Philosophy in **Computer Science** supervised by **Xue (Steve) Liu*** January 2023 - January 2028 (Expected)
Cumulative GPA: 3.90/4.00
Awards: BMO Responsible AI Senior Scholar | BMO Responsible AI Fellowship | NeurIPS 2023 Scholar Award | Faculty of Science Graduate Scholarship | Graduate Excellence Awards

McGill University Montreal, Quebec, Canada
*Bachelor of Science in **Honours Computer Science*** September 2019 - December 2022
Cumulative GPA: 3.95/4.00 | **Graduated as First Class Honours and Distinction**
Awards: Tomlinson Undergraduate Award | Faculty of Science Scholarship | Dean's Honour List

PUBLICATIONS/PREPRINTS

Importance-Aware Co-Teaching for Offline Model-Based Optimization **NeurIPS 2023**
Authors: Ye Yuan, Can Chen*, Zixuan Liu, Willie Neiswanger, Xue Liu*

Learning to Extract Structured Entities Using Language Models **EMNLP 2024 Main**
Authors: Ye Yuan, Haolun Wu*, Liana Mikaelyan, Alexander Meulemans, Xue Liu, James Hensman, Bhaskar Mitra*
- ***Selected as an oral presentation, 168 out of 2271 accepted papers (7%)***

Large Language Model (LLM) for Telecommunications: A Comprehensive Survey on Principles, Key Techniques, and Opportunities **IEEE COMST 2024**
Authors: Hao Zhou, Chengming Hu, Ye Yuan, Yufei Cui, Yili Jin, Can Chen, Haolun Wu, Dun Yuan, Li Jiang, Di Wu, Xue Liu, Charlie Zhang, Xianbin Wang, Jiangchuan Liu

Design Editing for Offline Model-based Optimization **Preprint**
Authors: Ye Yuan, Youyuan Zhang*, Can Chen, Haolun Wu, Zixuan Li, Jianmo Li, James J. Clark, Xue Liu*

Large Language Model (LLM)-enabled In-context Learning for Wireless Network Optimization: A Case Study of Power Control **Preprint**
Authors: Hao Zhou, Chengming Hu, Dun Yuan, Ye Yuan, Di Wu, Xue Liu, Charlie Zhang

Generative AI as a Service in 6G Edge-Cloud: Generation Task Offloading by In-context Learning **Preprint**
Authors: Hao Zhou, Chengming Hu, Dun Yuan, Ye Yuan, Di Wu, Xue Liu, Zhu Han, Charlie Zhang

Retrieval-Augmented Generation for Natural Language Processing: A Survey **Preprint**
Authors: Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue Liu, Tei-Wei Kuo, Nan Guan, Chun Jason Xue

WORK EXPERIENCES

Noah's Ark Lab Canada **April 2023 - Present**
Associate Researcher Intern at Reasoning-Decision-Making Team and NLP Team Montreal, Quebec, Canada
- Implemented the ideas of Parameter Sharing, Vision Transformer with variable input length, Conditional Batch-Norm Layers, Prompting with Conditional Embedding layers; Tested these methods in the simulated environment.
- Explored effective model compression techniques, including structured pruning, quantization, and distillation. Innovated and crafted new techniques, such as LoRA-pruning and Mixture of Depth model, to save calculations.
- Took strategy insights into cutting-edge LLM techniques, especially for efficient model training, knowledge injections and editing, reinforcement learning with AI feedback, and potential techniques for next-generation model architecture.
- Researched on novel states-based model architecture to enhance the in-context learning and retrieval abilities.

Mila – Quebec AI Institute **May 2024 - Present**
Mentor for Professional MSc Student's Internship Montreal, Quebec, Canada
- Supervised a professional Master of Science student for their industrial internship's project at Deep River.
- Provided guidance and suggestions for entity-linking, fine-tuning, knowledge distillation, and RAG for LLMs.

OTHERS

Services: Reviewer of ICLR 2025, AISTATS 2025, NeurIPS 2024, NeurIPS 2024 SafeGenAi, ICML 2024 FM-Wild; Program Committee Member of AAAI 2025 Undergraduate Consortium; Mentor and Judge for McGill McHacks 9, 10, 11.
Tutorships: Teaching Assistants (Comp 202 in F2020, F2023, W2024 and Comp250 in W2021, F2023, W2024) at McGill.