Ye Yuan

Ph.D. Candidate, Computer Science, McGill University, Mila - Quebec AI Institute
Visiting Scholar, Machine Learning, MBZUAI
DAAD AINeT Fellowship, BMO Responsible AI Fellowship

Last updated on Jan. 2026.

Room 305, McConnell Engineering Building
School of Computer Science, McGill University
✆ (+1) 438-351-1806
✉ ye.yuan3@mail.mcgill.ca
✉ ye.yuan@mbzuai.ac.ae
 GitHub, *X (Twitter)*, *LinkedIn*
 Google Scholar, *Personal Home Page*

Research Interests

Areas: Large Language Models [C2, C4, J1], Score-Based Generative Models [C3, J4, S8], Offline Black-Box Optimization [C1, J5, S5], Retrieval-Augmented Generation [S2, S4, S6], Autonomous Agents [S7].

My research centers on developing *large language models* and *score-based generative models* to advance intelligent systems, with applications in (i) addressing challenges in offline black-box optimization, (ii) enabling knowledge-centric NLP tasks such as automatic knowledge base construction, and (iii) improving the efficiency and efficacy of AI systems for real-world applications. My work is supported by multiple fellowships and has been published at top venues including NeurIPS, ICLR, EMNLP, and TMLR. Representative contributions include authoring a comprehensive survey on offline black-box optimization with Turing Award Winner Prof. Yoshua Bengio, which is accepted by TMLR with the Survey Certification, and presenting an EMNLP 2024 oral paper (top 3%), complemented by industrial research experiences or close collaborations at Microsoft Research, RBC Borealis, and Samsung Research America.

Academic Experience

Oct. 2025 - Present — **MBZUAI - Mohamed bin Zayed University of Artificial Intelligence**, *Abu Dhabi, United Arab Emirates.*
Visiting Scholar, Machine Learning.
Faculty Host: Associate Vice President for Research, Prof. Steve Liu.

Jan. 2023 - Present — **McGill University & Mila - Quebec AI Institute**, *Montreal, Canada.*
Doctor of Philosophy (Ph.D.), Computer Science.
CGPA: 3.89 / 4.00.
Supervisor: Prof. Steve Liu.
Supervisory Committee: Prof. Adriana Romero-Soriano; Prof. Gintare Karolina Dziugaite.

Sep. 2019 - Dec. 2022 — **McGill University & Mila - Quebec AI Institute**, *Montreal, Canada.*
Bachelor of Science (B.Sc.), Honours Computer Science.
CGPA: 3.95 / 4.00.
First Class Honours; Dean's Honour List (Top 10%); Recipient of Faculty of Science Scholarship.

Industrial Experience

May. 2025 - Oct. 2025 — **RBC Borealis**, *Toronto, Canada.*
Research Intern.
Manager: Dr. Eric He; Project Mentors: Dr. Mohammad Amin Shabani; Dr. Siqi Liu.

- Led two concurrent research projects that resulted in two conference paper submissions and one accepted NeurIPS workshop paper within a single internship term.
- Proposed an embedding-based context-aware reranking algorithm leveraging structural information and hybrid attention, improving both accuracy and efficiency.
- Developed an LLM-based agentic framework for query-focused table summarization, achieving higher summarization quality while ensuring privacy compliance by avoiding raw data exposure.

Jan. 2025 -	**Samsung Research America**, *Mountain View, United States.*
Present	**Student Researcher** (Remote).

Manager: Dr. Charlie Zhang.

- Designed entity extraction strategies to automatically construct domain-specific knowledge bases, enabling scalable knowledge graph creation for telecommunications data.

Apr. 2023 -	**Noah's Ark Lab**, *Montreal, Canada.*
Mar. 2025	**Research Intern**.

Manager: Dr. Alex Chen.

- Recognized with multiple awards, including Outstanding Contribution Award (R&D Peripheral Fields), Overseas Business Contribution Award, and President's Spot Award (Canada Research Institute).
- Adapted soft prompting techniques to address distribution shift, enabling a single model to generalize across multiple tasks.
- Explored model compression methods (e.g., LoRA-pruning, Mixture-of-Depth) to reduce computation cost for edge deployment.
- Investigated state-space language models and hybrid architectures with transformers for improved efficiency and scalability.

Feb. 2023 -	**Microsoft Research**, *Cambridge, United Kingdom.*
Present	**Student Researcher** (Remote).

Manager: Yordan Zaykov; Project Mentors: Dr. Bhaskar Mitra; Dr. Tom Minka.

- Introduced and formalized the structured entity extraction as a new paradigm within information extraction.
- Proposed a new evaluation metric (AESOP) and its variants for structured entity extraction, better aligning evaluation with human judgment.
- Developed a multi-stage decoding strategy that improved extraction effectiveness and efficiency.
- Analyzed the learning dynamics of score-based generative models and LLM-based discriminative models in structured entity linking tasks through a series of controlled experiments.

Publications

Within the first three years of my Ph.D., I produced around 10 peer-reviewed publications, 7 as (co-)first author, across top-tier venues such as NeurIPS, ICLR, EMNLP, and TMLR. By January 2026, my work had attracted 550+ citations (h-index 8) on Google Scholar, and my open-source contributions were recognized with 150+ stars on GitHub. **Detailed publication links are accessible via the item numbers listed below when available.**

Submissions / Preprints

S12 Weixu Zhang, **Ye Yuan**, Changjiang Han, Yuxing Tian, Zipeng Sun, Linfeng Du, Jikun Kang, Hong Kang, Xue Liu, Haolun Wu. "*Preference Heads in Large Language Models: A Mechanistic Framework for Interpretable Personalization*". Under Review.

S11 Difan Jiao, Yilun Liu, **Ye Yuan**, Zhenwei Tang, Linfeng Du, Haolun Wu, Ashton Anderson. "*Leveraging Internal Representations for LLM Safeguard*". Under Review.

S10 Zhiwei Liu, Yupeng Cao, Yuechen Jiang, Mohsinul Kabir, Polydoros Giannouris, Chen Xu, Ziyang Xu, Tianlei Zhu, Md. Tariquzzaman, Triantafillos Papadopoulos, Yan Wang, Lingfei Qian, Xueqing Peng, Zhuohan Xie, **Ye Yuan**, Saeed Almheiri, Abdulrazzaq Alnajjar, Ming-Bin Chen, Harry Stuart, Paul Thompson, Prayag Tiwari, Alejandro Lopez-Lira, Xue Liu, Jimin Huang, Sophia Ananiadou. "*Same Claim, Different Judgment: Benchmarking Scenario-Induced Bias in Multilingual Financial Misinformation Detection*". Under Review.

S9 Jiaxuan Chen, Rajat Ghosh, **Ye Yuan**, Karan Gupta, Xue Liu, Oana Balmau. "*DeltaServe: SLO-Budgeted Co-Serving of LLM Inference and Adaptation*". Under Review.

S8 Sajad Ebrahimi, Bhaskar Mitra, Negar Arabzadeh, Haolun Wu, **Ye Yuan**, Fattane Zarrinkalam, Ebrahim Bagheri. "*From Noise to Order: Learning to Rank via Denoising Diffusion Models*". Under Review.

S7 **Ye Yuan**, Mohammad Amin Shabani, Siqi Liu. "*FACTS: Table Summarization via Offline Template Generation with Agentic Workflows*". arXiv preprint arXiv:2510.13920.

S6 **Ye Yuan**, Mohammad Amin Shabani, Siqi Liu. "*Embedding-Based Context-Aware Reranker*". arXiv preprint arXiv:2510.13329.

S5 **Ye Yuan***, Can Chen*, Zipeng Sun, Dinghuai Zhang, Christopher Pal, Xue Liu. "*Diffusion Large Language Models for Black-Box Optimization*". Under Review.

S4 Linfeng Du, **Ye Yuan**, Zichen Zhao, Fuyuan Lyu, Emiliano Penaloza, Xiuying Chen, Zipeng Sun, Jikun Kang, Laurent Charlin, Xue Liu, Haolun Wu. "*Optimizing User Profiles via Contextual Bandits for Retrieval-Augmented LLM Personalization*". Under Review.

S3 Hao Zhou, Chengming Hu, Dun Yuan, **Ye Yuan**, Di Wu, Xue Liu, and Charlie Zhang. "*Large Language Model (LLM)-Enabled In-Context Learning for Wireless Network Optimization: A Case Study of Power Control*". arXiv preprint arXiv:2408.00214.

S2 Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, **Ye Yuan**, Lianming Huang, Xue Liu, Tei-Wei Kuo, Nan Guan, Chun Jason Xue. "*Retrieval-Augmented Generation for Natural Language Processing: A Survey*". arXiv preprint arXiv:2407.13193.

S1 Shuhao Zheng, Yanxi Lin, Yang Yu, **Ye Yuan**, Yongzheng Jia, Xue Liu. "*FIAT: Fine-grained Information Audit for Trustless Transborder Data Flow*". arXiv preprint arXiv:2209.11451.

Journal Articles

J5
TMLR
Survey
Certification
Minsu Kim*, Jiayao Gu*, **Ye Yuan**, Taeyoung Yun, Zixuan Liu, Yoshua Bengio, and Can Chen. "*Offline Model-Based Optimization: Comprehensive Review*". Transactions on Machine Learning Research, 2026.

J4
TMLR
Ye Yuan*, Youyuan Zhang*, Can Chen, Haolun Wu, Zixuan Li, Jianmo Li, James J. Clark, and Xue Liu. "*Design Editing for Offline Model-based Optimization*". Transactions on Machine Learning Research, 2025.

J3
IEEE WCM
IF: 11.5
Hao Zhou, Chengming Hu, Dun Yuan, **Ye Yuan**, Di Wu, Xi Chen, Hina Tabassum, and Xue Liu. "*Large Language Models for Wireless Networks: An Overview from the Prompt Engineering Perspective*". IEEE Wireless Communications, 2024.

J2
IEEE WCL
IF: 5.5
Hao Zhou, Chengming Hu, Dun Yuan, **Ye Yuan**, Di Wu, Xue Liu, Zhu Han, and Jianzhong Zhang. "*Generative AI as a Service in 6G Edge-Cloud: Generation Task Offloading by In-Context Learning*". IEEE Wireless Communications Letters, 2024.

J1
IEEE COMST
IF: 46.7
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, and Jiangchuan Liu. "*Large Language Model (LLM) for Telecommunications: A Comprehensive Survey on Principles, Key Techniques, and Opportunities*". IEEE Communications Surveys and Tutorials, 2024.

Conference Proceedings

C4
GlobeCom
Oral, 39.6%
Ye Yuan, Haolun Wu, Hao Zhou, Xue Liu, Hao Chen, Yan Xin, Jianzhong (Charlie) Zhang. "*Understanding 6G through Language Models: A Case Study on LLM-aided Structured Entity Extraction in Telecom Domain*". In the 2025 IEEE Global Communications Conference. Taipei, Taiwan. Dec. 2025.

C3
ICLR
Poster, 32.1%
Ye Yuan*, Can Chen*, Christopher Pal, Xue Liu. "*ParetoFlow: Guided Flows in Multi-Objective Optimization*". In the 13th International Conference on Learning Representations. Singapore EXPO, Singapore. Apr. 2025.

C2 EMNLP *Oral, Top 3%*	**Ye Yuan***, Haolun Wu*, Liana Mikaelyan, Alexander Meulemans, Xue Liu, James Hensman, and Bhaskar Mitra. *"Learning to Extract Structured Entities Using Language Models"*. In the 19th Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA. Nov. 2024.
C1 NeurIPS *Poster, 26.1%*	**Ye Yuan***, Can Chen*, Zixuan Liu, Willie Neiswanger, Xue Liu. *"Importance-aware Co-teaching for Offline Model-based Optimization"*. In the Advances in Neural Information Processing Systems 36. New Orleans, Louisiana, USA. Dec. 2023.

Workshops

W3 NeurIPS	**Ye Yuan**, Mohammad Amin Shabani, Siqi Liu. *"FACTS: Fast, Accurate, and Privacy-Compliant Table Summarization via Offline Template Generation"*. In NeurIPS 2025 Workshop on Generative AI in Finance.
W2 ICML	Hao Zhou, Chengming Hu, Dun Yuan, **Ye Yuan**, Di Wu, Xue Liu, Jianzhong Charlie Zhang. *"Prompting Wireless Networks: Reinforced In-Context Learning for Power Control"*. In ICML 2025 Workshop on Machine Learning for Wireless Communication and Networks.
W1 ICLR	**Ye Yuan***, Youyuan Zhang*, Can Chen, Haolun Wu, Melody Zixuan Li, Jianmo Li, James J. Clark, Xue Liu. *"Design Editing for Offline Model-based Optimization"*. In ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy.

Grants & Fellowships

Total funding: CAD $339,000. Total share of funding: CAD $148,500.

I led or contributed to the work of the following grant and fellowship proposals.

2025	Cohere Labs Catalyst Grant *(CAD $1,000, share 100%)*
	Funding Body: Cohere Lab
2025	Tinker Research Grant *(CAD $7,000, share 100%)*
	Funding Body: Thinking Machines Lab
2025	DAAD AINeT Fellowship *(CAD $5,000, share 100%)* (1 of 18 AINet Fellows in NLP)
	Funding Body: DAAD
2024	Samsung Global Research Outreach Program *(CAD $210,000, share 25%)* (1 of 11 Recipients Globally)
	Lead PI: Prof. Steve Liu
	Title: LLM-Driven Efficient and Flexible Network Management for NextG Cellular Communications
	Funding Body: Samsung Research
2023	MSR-Mila Collaboration Grant *(CAD $66,000, share 50%)*
	Lead PI: Prof. Steve Liu; Industrial Mentor: Dr. Bhaskar Mitra
	Title: SNAKE: Structured Neural Attention over Knowledge base Entities
	Funding Body: Microsoft Research
2023	BMO Responsible AI Fellowship *(CAD $50,000, share 100%)*
	Funding Body: Bank of Montreal (BMO)

Honors & Awards

Total award amount: CAD $27,000.

2025	Graduate Research Enhancement and Travel Awards (GREAT Travel Award), *McGill University*
2025	Travel Grant, *NeurIPS 2025 Workshop GenAI in Finance*
2025	AINet Fellows in Natural Language Processing, *DAAD*
2025	Graduate Excellence Award, *McGill University*

2025	Student Travel Award, *International Conference on Learning Representations (ICLR)*
2023-2025	Responsible AI Senior Scholar, *Bank of Montreal*
2023-2025	Faculty of Science Graduate Scholarship, *McGill University*
2024	Outstanding Contribution Award in R&D Peripheral Fields, *Noah's Ark Lab Canada*
2024	Overseas Business Contribution Award, *Noah's Ark Lab Canada*
2024	Canada Research Institute President's Spot Award, *Noah's Ark Lab Canada*
2023	Scholar Award, *Neural Information Processing Systems (NeurIPS)*
2022	Graduate Excellence Award, *McGill University*
2022	Graduated as First Class Honours, *McGill University*
2019-2022	Dean's Honour List, *McGill University*
2020-2021	Tomlinson Undergraduate Award, *McGill University*

Media

I actively share and communicate research on social media, with 1,800+ followers across LinkedIn and X.

2025	I represent Mila at its NeurIPS 2025 booth to present my research about "From Surrogates to Diffusion and Language Models: Generative Paradigms for Offline Black-Box Optimization". Mila NeurIPS 2025 Booth.
2024	Oral Presentation of Learning to Extract Structured Entities Using Language Models. EMNLP Recording.
2024	Large Language Model (LLM) for Telecommunications: A Comprehensive Survey on Principles, Key Techniques, and Opportunities. IEEE COMST Post.
2023	Importance-Aware Co-Teaching for Offline Model-Based Optimization. Mila Blog.

Invited Talk, Teaching & Mentorship

○ **Invited Talk**
  - *"From Surrogates to Diffusion and Language Models: Generative Paradigms for Offline Black-Box Optimization."*, Mila.
    Host: Mila NeurIPS 2025 Booth. December. 2025.
  - *"Generative VS. Discriminative Approaches for Entity Linking."*, Microsoft Research.
    Host: Yordan Zaykov. May. 2025.
  - *"Offline Model-Based Optimization"*, Beijing Academy of Artificial Intelligence.
    Host: Dr. HongJiang Zhang. Jul. 2024.
  - *"Offline Model-Based Optimization"*, ByteDance AI Lab.
    Host: Dr. Hang Li. Jun. 2024.

○ **Teaching**
  - *"Global AI Leadership Program"*, Teaching Assistant, MBZUAI.
    Worked as the Teaching Assistant for MBZUAI's Global AI Leadership Program equips UAE leaders for the AI era. The program brought together 46 participants from across the UAE's public and private sectors, including senior executives, policymakers, entrepreneurs, and thought leaders.
  - *"COMP 202: Foundations of Programming"*, Teaching Assistant, McGill University.
    Worked as a Teaching Assistant for two semesters at McGill for COMP 202: Foundations of Programming. This course is an introduction to computer programming in a high level language. I worked with Prof. Faten M'hiri for this course.
  - *"COMP 250: Introduction to Computer Science."*, Teaching Assistant, McGill University.
    Worked as a Teaching Assistant for two semesters at McGill for COMP 250: Introduction to Computer Science. This course is about Mathematical tools, Datastructures, Recursive and non-recursive algorithms, Abstract data types and Object oriented programming in Java. I worked with Prof. Giulia Alberini for this course.

○ **Mentorship**
  - *"Mila Industry Mentorship Program for Machine Learning Internships"*, Mentor, Mila - Quebec AI Institute.
    Mentored Master students for their internship projects.

- · Work with Arvind Nair for mentoring his internship at Bell Canada with Prof. Dhanya Sridhar.
- · Work with Paul Kelendji for mentoring his internship at DeepRiver with Prof. Bang Liu.
- *"Mentor and Collaborate with Junior Students for Research Projects"*
  Mentored and collaborated with several undergraduate and junior graduate students:
  - · Linfeng Du (McGill)
  - · Zipeng Sun (McGill)
  - · Weixu Zhang (McGill)
  - · Jiaxuan Chen (McGill)
  - · Yuyan Lin (McGill)
  - · Changjiang Han (MBZUAI)
  - · Meng Bi (MBZUAI)
  - · Zichen Zhao (MBZUAI)
  - · Zhengxi Li (MBZUAI)
  - · Yonghan Yang (MBZUAI)

Professional & Community Service

- **Conferences Reviewer**
  - ICML 2026
  - WWW 2026
  - ICLR 2026
  - AISTATS 2026
  - NeurIPS 2025
  - ACL Rolling Review 2025 (ACL, EMNLP)
  - ICML 2025
  - ICLR 2025
  - AISTATS 2025
  - NeurIPS 2024

- **Journals Reviewer**
  - TMLR
  - ACM TOIS

- **Workshops Reviewer**
  - ICLR DeLTa 2025
  - AAAI Undergrad Consortium 2025
  - AAAI KnowFM 2025
  - NeurIPS SafeGenAi 2024
  - ICML FM-Wild 2024

- **Community Service**
  - Co-organizing CLEF 2026 FinMMEval Lab (Multilingual and Multimodal Evaluation of Financial AI Systems)
  - Mentor and Judge for McGill McHacks Hackathon 9, 10, 11, 12, 13 (2022 - 2026)

Extracurricular Activities

- One of my biggest hobby is playing **Xiangqi** (Chinese Chess), which is a strategy board game and is the most popular board game in China. The game represents a battle between two armies, with the primary object being to checkmate the enemy's general. It is my fortunate to be mentored by the World Xiangqi Federation international master Bo Peng.
- When I was at junior high and senior high school, I was a member of the school's **basketball varsity**. We won the district's championship once, and the International Schools Basketball Cup Championship once.
- Before studying computer science, I was quite interested in **aerodynamics for plane design**. I received the China National First Tier Project award once, and Beijing Municipal First Tier Project award once.