Ye Yuan

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

What truly captivates me is the potential of intelligent systems and generative modelling to assist humans. How can artificial intelligence accurately and flawlessly complete tasks assigned by humans? How can generative models be applied for specific tasks? More specifically, my research is concentrated on (i) addressing out-of-distribution challenges in Offline Black Box Optimization, (ii) developing foundational knowledge models using generative models and language models, as well as (iii) enhancing language model efficiency through compression and distillation. My greatest assets are my eagerness to learn new things, my curiosity to delve into uncharted territories, my drive to stay abreast of the latest developments, and my persistence to work hard. Beyond academia, I collaborate closely with Microsoft Research Cambridge and Noah's Ark Lab Montreal.

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 | Faculty of Science Graduate Scholarship | Graduate Excellence Awards
Relevant Courses: Computer Networks, Matrix Computations, Computational Biology Methods, ML for Biomedical Data, Natural Language Understanding with Deep Learning, Deep Learning, Introduction to Ethics of Intelligent System

McGill University Montreal, Quebec, Canada
*Bachelor of Science in **Honours Computer Science*** September 2019 - December 2022
Cumulative GPA: 3.95/4.00 | Dean's Honour List | **Graduated as First Class Honours and Distinction**
Awards: Tomlinson Undergraduate Award | Faculty of Science Scholarship
Relevant Courses: Cryptography and Data Security, Intelligent Software System, Applied Machine Learning, Reinforcement Learning, Operating Systems, Natural Language Processing, Software Design, Modern Computer Games, Data Structures, Probability, Statistics, Numerical Analysis, Linear Algebra, Discrete Mathematics, Calculus

PUBLICATION/PREPRINT

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

- Introduced Importance-aware Co-Teaching(ICT) for offline MBO. ICT consists of two steps. In the pseudo-label-driven co-teaching step, a proxy is iteratively chosen as the pseudo-labeler, initiating a co-teaching process that facilitates knowledge exchange between the other two proxies.
- Utilized meta-learning-based sample reweighting to alleviate potential inaccuracies in pseudo-labels. In this step, pseudo-labeled samples are assigned importance weights, which are then optimized through meta-learning.

Learning to Extract Structured Entities Using Language Models **EMNLP 2024 Main (Oral Presentation)**
***Authors:* Ye Yuan*,** *Haolun Wu*, Liana Mikaelyan, Alexander Meulemans, Xue Liu, James Hensman, Bhaskar Mitra*

- Introduced and formalized the task of structured entity extraction within the realm of strict information extraction.
- Proposed an evaluation metric AESOP with numerous variants tailored for assessing structured entity extraction.
- Proposed a new model leveraging the capabilities of large language models (LLMs), improving the effectiveness and efficiency of structured entity extraction.

Large Language Model (LLM) for Telecommunications: A Comprehensive Survey on Principles, Key Techniques, and Opportunities **IEEE Communications Surveys and Tutorials**
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*

- Provided a comprehensive survey of the principles, key techniques, and applications for LLM-enabled telecom networks, ranging from LLM fundamentals to novel LLM-inspired generation, classification, optimization and prediction techniques along with telecom applications.
- Covered nearly 20 telecom application scenarios and LLM-inspired novel techniques, aiming to be a roadmap for researchers to use LLMs to solve various telecom tasks.

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

- Introduced Design Editing for Offline MBO (DEMO), which operates in two main phases: the first, pseudo-target distribution generation, involves employing a surrogate model to create a synthetic dataset and training a conditional diffusion model on this synthetic dataset to serve as the pseudo-target distribution.

- Refined existing top designs by introducing random noise to them and using the trained conditional diffusion model to denoise, resulting in designs which not only inherit high-scoring features from existing top designs but also achieve higher scores by leveraging information from the pseudo-target distribution.

Large Language Model (LLM)-enabled In-context Learning for Wireless Network Optimization: A Case Study of Power Control **Preprint (Under Review)**

Authors: *Hao Zhou, Chengming Hu, Dun Yuan,* ***Ye Yuan****, Di Wu, Xue Liu, Charlie Zhang*

- Utilized natural language to control complicated systems, demonstrating the potential of leveraging human language for optimizing wireless networks via pre-trained large language models.
- Explored a fundamental power control problem using LLM-enabled in-context learning and uncovered several fascinating insights: LLM agents can explore and learn from language-based demonstrations with the right prompt designs; LLM-enabled in-context learning can bypass the complexity of tedious model training and fine-tuning; LLMs can provide reasonable explanations for their power control decisions, aiding understanding of complex systems.

Generative AI as a Service in 6G Edge-Cloud: Generation Task Offloading by In-context Learning **Preprint (Under Review)**

Authors: *Hao Zhou, Chengming Hu, Dun Yuan,* ***Ye Yuan****, Di Wu, Xue Liu, Zhu Han, Charlie Zhang*

- Modeled the service delay of foundation GAI models in wireless networks.
- Provided a specified metric to evaluate the delay experienced by mobile users.
- Proposed an in-context learning method for tasks offloading, using natural language for network management.

Retrieval-Augmented Generation for Natural Language Processing: A Survey **Preprint (Under Review)**

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*

- Introduced the retriever from building to querying and techniques of the retrieval fusions with tutorial codes.
- Exhibited different RAG training strategies, including RAG with/without datastore update.
- Discussed the applications of RAG on downstream NLP tasks and practical NLP scenarios.
- Identified promising future directions for exploring and main challenges for addressing.

WORK EXPERIENCES

Noah's Ark Lab Canada **April 2023 - Present**

Associate Researcher Intern 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.

Other Research Projects

Existence of Anchors Can Help Information Compression **February – May 2024**

- Identified the existence of anchors in the textual documents with saliency scores and proposed a prompt compression technique based on this finding, which can be used for long context modelling.
- Leveraging the existence of anchors, we only prepend all activations of anchor words in each layer, avoiding the use of the entire demonstration, which reduces the time and memory complexity for calculating the attention of about 8 times.

Naïve Implementation of Operating System **September – December 2021**

- Developed a command-line interface to support basic functions, I/O redirection, and pipelines by multi-processes.
- Implemented one CPU, two CPUs, and I/O threads schedulers based on the First Come First Serve Algorithm.
- Realized the fundamental functions of the files system with the idea of i-nodes and files table in the Linux system.

ADDITIONAL

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.