Research Statement

My name is Ye Yuan, and I'm self-motivated third year PhD candidate at McGill University. My PhD research focuses on **score-based generative models**, with a particular emphasis on enhancing **guidance integration** and exploring the comparative strengths and limitations of generative models and traditional supervised learning models in real-world applications. My current work centers on applying score-based generative models in two primary real-world domains: **black-box optimization** and **knowledge-centric natural language processing tasks**.

In the domain of offline black-box optimization, my research includes three main contributions. First, my **NeurIPS 2023** paper on **Importance-aware Co-Teaching for Offline Model-Based Optimization** introduces a novel co-teaching approach that uses Gaussian Kernel-based data augmentation to enhance the robustness of surrogate models against out-of-distribution data. Second, in my paper published on **Transactions of Machine Research (TMLR)**, titled **Design Editing for Offline Model-Based Optimization**, I propose a diffusion prior-based approach to edit overly optimized unconstrained candidates back into the valid design space. This involves adding noise to generated designs and using a diffusion model to denoise them, aligning them with the original distribution. Third, my paper on **ICLR 2025**, **ParetoFlow: Guided Flows in Multi-Objective Optimization**, I extend score-based models to offline multi-objective optimization by developing a novel guidance module that incorporates multiple objectives into the sampling process. This work theoretically grounds and empirically validates the integration of flow-matching models with evolutionary algorithms for efficient trade-off exploration across multiple objectives. These optimization techniques can also be extended to be used during the post-training phases of large language models or other domains. Motivated by this, I **closely collaborate with Professor Yoshua Bengio** and other researchers from Mila to write a comprehensive survey of the offline black-box optimization problems, titled **Offline Model-Based Optimization: Comprehensive Review.**

The second domain focuses on leveraging score-based generative models for structured knowledge base construction from unstructured text. In my **EMNLP 2024 Oral Presentation** paper, **Learning to Extract Structured Entities Using Language Models**, I redefine information extraction by introducing structured entities beyond traditional triplets and propose the **Approximate Entity Set OverlaP (AESOP)** metric, which better captures model performance in this task. Currently, in collaboration with Microsoft Research, I am investigating the comparative performance of generative versus discriminative models for entity linking tasks across different data types, with respect to their overall accuracy for capturing the empirical distribution of training data, their generalizability to unseen data, as well as their learning and inference efficiency. This work explores novel methods to derive linking scores for generative models and examines whether these models adhere to fundamental linking axioms.

Additionally, I'm also generally interested in retrieval-augmented generation (RAG) techniques, large language models (LLMs) and their applications in real world tasks. Motivated by these interests, I have taken two internships at Noah's Ark Lab Canada and Royal Bank of Canada (RBC) Borealis, respectively. Moreover, I closely collaborate with other researchers and contribute two surveys about RAG and LLMs in telecommunications, respectively, as well as several works on applying to optimization problems in telecommunications, which are published in IEEE Communications Surveys & Tutorials, Wireless Communications Letters, and Wireless Communications Magazines.

Looking ahead, I plan to further explore score-based generative models in two key areas: 1) improving their performance on discrete data such as text, where auto-regressive models currently outperform them, and 2) accelerating their sampling process through techniques such as distillation and jump-step sampling and combines the advantages of score-based generative models and autoregressive language models into a single framework.