Research Interest. Large Language Model (LLM) powered agents have demonstrated immense promise in automating many complex workflows, such as coding, literature review, and data analysis. However, their applicability to real-world tasks remains limited, with a core bottleneck being their lack of adaptability to new environments, diverse data formats, and extensive toolsets. Broadly, my research interest is to address this bottleneck by **designing agentic frameworks that generalize across tasks**, focusing on **efficient self-improvement** to minimize manual engineering and **robust task-scale scaling** to handle challenging, large-scale reasoning tasks.

Efficient Self-Improvement. During my internship on ServiceNow's CoreLLM research team, I observed that the development of state-of-the-art agentic systems relied heavily on manual, labor-intensive engineering efforts. Inspired by this observation, I wondered if agentic system architectures could be autonomously optimized. In **Agentic Reasoning Module** [2], I proposed a framework where an evolutionary tree search-guided meta-agent iteratively proposes and reviews novel architectures. We demonstrate that these automatically designed agentic systems are powerful and generalize effectively, achieving state-of-the-art performance across math, science, and commonsense reasoning benchmarks, and across three different backbone LLMs. Notably, these architectures generalized without requiring domain or model specific re-optimization. This work resulted in a first-author publication which I presented at the **NeurIPS 2025 Math-AI Workshop** and is currently under review at **ICLR 2026**.

With the promise of automatic agentic system design demonstrated on textual reasoning tasks, my next target was to extend this paradigm to automate the design of heterogeneous agentic systems capable of reasoning over multiple data sources of various modalities, including text, charts, tables, long documents, and web data. In this problem setting, the agentic system designer must learn to design architectures that are capable of dynamic interactions with a vast array of tools, such as multimodal models, OCR tools, web search, etc. I am currently working on this problem as a part time research scientist at ServiceNow.

Robust Task-Scale Scaling. With the capabilities of agents rapidly increasing, I am passionate about developing agentic systems that effectively utilize compute and data to address problems of larger scale. While a visiting researcher at ServiceNow, I found that current agentic frameworks for math reasoning failed to scale effectively to competition-level problems. I identified that the key limitation was the reliance on a single tool invocation per reasoning step, which frequently led to tool invocation mistakes and erroneous tool selection. To address this limitation, I proposed **Multi-TAG** [1], an agentic framework that guides the backbone LLM to

invoke multiple tools at each reasoning step and aggregate their outputs to produce accurate, cross-tool verified reasoning paths. Multi-TAG significantly outperformed state-of-the-art frameworks with parallel compute scaling and achieved the largest improvements on the most difficult subset of questions. These results highlight the importance of effectively scaling agent compute for solving complex reasoning tasks. This work resulted in a first-author publication at **Findings of EMNLP 2025** and the **NeurIPS 2025 Math-AI Workshop**.

Motivated by the success of this project, I moved on to tackle a project of even greater scale: automatically documenting linguistic feature differences between non-standard English dialects and Standard American English [3]. Collaborating with Prof. Noah Smith's *Noah's ARK* lab and Prof. Yulia Tsvetkov's *Tsvetshop* lab, I designed an agentic system that analyzes a massive corpus of dialectal texts in parallel to propose and verify linguistic feature differences. We found that our approach significantly outperforms non-agentic baselines, such as simple prompting of frontier models, on LLM-as-a-judge evaluations, demonstrating the potential for agentic systems to conduct novel scientific research through large-scale data analysis. We are currently working on conducting the final human evaluation experiments and preparing the work for publication in January 2026.

Future Goals. During my PhD, I intend to continue unlocking the potential of agentic systems to solve challenging, real world problems. My vision is to design generalist agentic systems capable of efficient self improvement and tackling large-scale tasks requiring effective compute and data scaling. Ultimately, I aim to produce truly powerful autonomous agents that push forward scientific discovery, serve as capable assistants, and meaningfully improve the quality of life for everyone. I believe that a PhD offers the ideal environment for me to develop my research skills, collaborate with peers with a similar vision, and contribute my ideas to the academic community.

References

[1] **B. Yao**, and V. Yadav, "A Toolbox, Not a Hammer – Multi-TAG: Scaling Math Reasoning with Multi-Tool Aggregation," Findings of Conference on Empirical Methods in Natural Language Processing (EMNLP), 2025.

[2] **B. Yao**, S. K. R. Malay, and V. Yadav, "ARM: Discovering Agentic Reasoning Modules for Generalizable Multi-Agent Systems," The 5th Workshop on Mathematical Reasoning and AI (NeurIPS), 2025.

[3] **B. Yao**, O. Ahia, K. Ahuja, N. Smith, Y. Tsvetkov, "Documenting Dialectal Differences Using Language Models," In preparation.