Yuanhao (Henry) Shen

https://henrysyh2000.github.io/

Education

Queen's University — **Kingston, ON** — **Sep 2024 – Present**

- **Ph.D in Computer Science**
- **Supervisor:** Prof. Xiaodan Zhu
- **Research Interests:** AI for Science, Mechanistic Interpretability, Large Language Models (LLMs), Natural Language Processing (NLP).

New York University — **New York City, NY** — **Sep 2022 – May 2024**

- **M.S. in Data Science (Cum Laude), GPA: 3.82/4.00**
- **Relevant Coursework:** Natural Language Understanding, Reinforcement Learning, Deep Learning, Mathematics in Deep Learning, Big Data.

New York University Shanghai — **Shanghai, China** — **Sep 2018 – May 2022**

- **B.S. in Data Science (Cum Laude), GPA: 3.81/4.00**
- **Relevant Coursework:** Machine Learning, Linear Algebra, Probability and Statistics, Fundamental Algorithms.
- **Scholarships and Awards:** ZhenTong Scholarship (RMB 120, 000), Dean's List of Academic Year.

Skills

- **Programming:** Python, Slurm, MySQL, Latex
- **AI & NLP:** Pytorch, Huggingface
- **Model Interpretability:** Hidden State Geometry Analysis, Steering

Research

Feasibility Probe for Interdisciplinary Research (IDR): Validation Platform — Sep 2026 – Present

Work in Progress

- Scoping a feasibility probe for interdisciplinary research (IDR), defining what can be validated and how to evaluate cross-disciplinary ideas.
- Designing and building a validation platform to test IDR feasibility; currently under active development.

On the Diversity of Analogy Making in Large Language Models — Sep 2025 - June 2026

Work submitted to ICLR 2027. Under review.

- Developed a multi-faceted evaluation framework covering sentence diversity, target-domain clustering, inter-model similarity, and generation quality, with human-in-the-loop verification.
- Revealed pervasive domain homogeneity: despite being prompted to generate ten diverse analogies, models produced only two semantic target-domain clusters on average, with the best-performing model reaching 3.5 clusters.
- Evaluated eight inference-time diversity-enhancement methods and uncovered a model-dependent diversity-quality trade-off; layer-wise causal analysis further linked their instability to mismatches between intervention locations and analogy-sensitive model regions.

IDRBench: Understanding the Capability of Large Language Models on Interdisciplinary Research — Nov 2024 - Sep 2025

ICML 2025 Assessing World Models Workshop, 2026 Main.

- Introduced IDRBench, the first comprehensive closed-book benchmark for evaluating LLMs' interdisciplinary research capabilities, featuring 335 expert-annotated paper triplets across six disciplines.
- Designed three progressive tasks to evaluate LLMs' capabilities in interdisciplinary paper identification, idea integration, and research idea recommendation.
- Benchmarked ten state-of-the-art LLMs and revealed their difficulty in distinguishing genuine interdisciplinary integration from merely feasible combinations, as well as a reasoning-creativity trade-off in reasoning-oriented models.
- Validated LLM-generated ideas with 6 experts and 56 domain researchers, and demonstrated the scalability of IDRBench through a 12.72% average Macro-F1 improvement after LoRA fine-tuning.

SMARTCAL: An Approach to Self-Aware Tool-Use Evaluation and Calibration — May 2023 - July 2024

EMNLP 2024 Industry Track.

- Demonstrated the tool-abuse behavior during tool-augmented reasoning of LLM agents as well as an tendency of agents being overconfident in tool usage.
- Proposed a novel framework SMARTCAL that consists of three modules to mitigate the observed issues in tool-abuse.

- Conducted extensive experiments to demonstrate the effectiveness of SMARTCAL with an improvement of 8.6 percent more in QA performance and 21.6 percent less in Calibration Error

Services and Affiliations

Conference Reviewer

ICML 2025, 2026, NeurIPS 2025, 2026, ARR 2026, ICLR 2027, AAAI 2027

Affiliated Researcher

- Vector Institute for Artificial Intelligence (Sep 2025 - present)
- National Research Council Canada (Sep 2024 - present)

Project Experience

Causal Relationship Extraction from Political News Texts **Jan 2024 - May 2024**

Master Capstone Project

- Implemented an LLM-driven pattern-recognition pipeline to extract causal relations within political news texts.
- Created a synthetic dataset containing the causal relation triplets used for fine-tuning on smaller scale LLMs.
- Visualized the relationship between extracted entities with a directed graph and conducted qualitative analysis.

Transfer Learning Based Fine-Tuning on Figurative Language Detection **Feb 2023 - May 2023**

Coursework Project (Demo Award)

- Linked Figurative speech with valence, arousal, and dominance (VAD) scores based on previous works and explored its potential application in figurative speech detection.
- Used sarcasm and metaphor dataset from twitter and performed single task learning on BERTweet for VAD regression. Applied Max-pooling technique to obtain text representation and add a decoder layer for classification task.
- Conducted HPC-based model pre-training and used adapter as the parameter-efficent fine-tuning strategy.

Sentiment Analysis in High Frequency Trading **Mar 2022 - May 2022**

Undergraduate Capstone Project

- Acquired historical daily stock prices and daily news headlines from Wharton Researh Database (10 million) and constructed a LSTM model for sentiment analysis.
- Incorporated sentiment analysis results into the DDPG algorithm and used GCC for remote training.
- Calculated the overall returns of the model and compared it with the baseline. Achieved around 10% improvement in cumulative return and 20% less in maximum drawback.