Hao Yan

Fairfax, VA • +1-571-354-3431 • yanhao821@gmail.com • LinkedIn • Google Scholar

Education

George Mason University **Fairfax, VA**
Ph.D. in Computer Science (advised by Dr. Ziyu Yao) *Aug. 2021 – Expected Dec. 2026*
M.S. in Computer Science *Aug. 2018 – Dec. 2020*

Fujian Normal University **Fujian, China**
B.E. in Electronic and Information Engineering *Aug. 2013 – Jun. 2017*

Experience

Applied Scientist Intern, Amazon *(Seattle, WA)* **May 2026 – Aug. 2026**

- Researched evolutionary strategies (ES) for backpropagation-free LLM post-training at production scale.
- Designed an improved evolutionary strategy with higher task performance and better training efficiency.
- Evaluated ES post-trained models against GRPO-trained baselines and analyzed performance–efficiency trade-offs.

Applied Scientist Intern, Amazon *(Seattle, WA)* **May 2025 – Aug. 2025**

- Built a reward model that incorporates personalized user preferences into its evaluation of smart-home actions (API calls) and serves as both an automatic evaluator and the reward signal for GRPO-based post-training.
- Designed a distillation-based training strategy for the reward model and implemented the end-to-end data-processing and model-training pipeline.
- Evaluated state-of-the-art reward-modeling methods and led error analysis to identify future improvements.

Machine Learning Engineer Intern, Salesteq *(Remote)* **Jan. 2025 – Apr. 2025**

- Participated in the design of an agentic LLM framework for car sales and service workflows, using AutoGen.
- Built conversation memory and RAG-style car-inventory retrieval to ground recommendations in dealership data.
- Integrated external tools (email, scheduling APIs) via structured prompting and adaptive control logic.

Graduate Research & Teaching Assistant, George Mason University *(Fairfax, VA)* **Aug. 2021 – Present**

- Conduct research on LLM for Code Generation, LLM Safety and Trustworthiness, Agentic AI, Mechanistic Interpretability of LLMs, LLM Explainability, Human–AI Interaction, Persona Simulation, and Efficient LLM Reasoning.

Front-end Developer, Bright Sky Tech Inc. *(Annandale, VA)* **Jan. 2021 – Aug. 2021**

- Built and maintained a web-based online grocery platform using Angular, Firebase, TypeScript, and Ionic.

Publications *(* denotes equal contribution)*

Hao Yan, Ziyu Yao. "Interpreting and Steering for Safe and Correct Code Generation". **EMNLP 2026**, Long Paper.

Hao Yan, Swapneel Suhas Vaidya, Xiaokuan Zhang, Ziyu Yao. "Guiding AI to Fix Its Own Flaws: An Empirical Study on LLM-Driven Secure Code Generation". **ACM TOSEM 2026**. [paper]

Hao Yan, Saurabh Srivastava, Yintao Tai, Sida I. Wang, Wen-tau Yih, Ziyu Yao. "Learning to Simulate Natural Language Feedback for Interactive Semantic Parsing". **ACL 2023**, Long Paper. [paper]

Saaketh Anand Sai Kumar*, **Hao Yan***, Sai Charan Perepa, Murong Yue, Ziyu Yao. "Can LLMs Simulate Personas with Reversed Performance? A Systematic Investigation for Counterfactual Instruction Following in Math Reasoning Context". **COLM 2025 Workshop** on Social Simulation with LLMs. [paper]

Hao Yan, Thomas D. LaToza, Ziyu Yao. "IntelliExplain: Enhancing Conversational Code Generation for Non-Professional Programmers". **arXiv preprint**, 2024. [paper]

Saurabh Srivastava, Janit Bidhan, **Hao Yan**, et al. "Reasoning Under Constraint: How Batch Prompting Suppresses Overthinking in Reasoning Models". **ICLR 2026 Workshop** on Logical Reasoning of LLMs. [paper]

Professional Service

- **Program Committee**: IJCAI 2026, 2025; ECAI 2025
- **Reviewer**: NeurIPS 2026; ICLR 2026, 2025; ACL ARR 2026, 2025; IEEE CI-NLPSoMe 2025; WLLFM 2024
- **Invited Talk**: "LLM Agents: Build Your Own Personal Assistant," EPIC SPICE Intern Program, George Mason University, 2024
- **Volunteer**: Student Volunteer, 61st Annual Meeting of the Association for Computational Linguistics (ACL), 2023

Technical Skills

Programming: Python, SQL, Java, JavaScript, TypeScript

ML & LLM Frameworks: PyTorch, Hugging Face (Transformers, Datasets, PEFT), vLLM, AutoGen, scikit-learn, NumPy, Pandas

LLM Post-Training: Supervised Fine-Tuning & LoRA, RLHF, GRPO, Evolutionary Strategies, Reward Modeling, Knowledge Distillation

LLM Systems & Agents: Agentic Frameworks, Tool Use, Retrieval-Augmented Generation, Prompt Engineering

Evaluation & Research Methods: Benchmark & Dataset Construction, Model Evaluation, Error Analysis, Human & LLM-Based User Studies, Model Interpretability & Explainability

Infrastructure: Linux, Git, AWS, Docker, GPU/Cluster Training, Weights & Biases