

Guangzhi Xiong

guangzhi@virginia.edu | [Website](#) | [Google Scholar](#) | [GitHub](#) | [LinkedIn](#)

Research Interests

My research focuses on building **LLM and multimodal agents** that search and reason over long horizons. I optimize how these agents acquire and use information through both post-training (e.g., reinforcement learning) and training-free methods (e.g., in-context learning).

I also investigate how these models work internally and how to evaluate them, with the goal of making them **faithful, interpretable, and controllable**. My work has broad applications, from healthcare and scientific discovery to wearable assistants and personalized shopping.

Education

University of Virginia	United States
Ph.D. in <i>Computer Science</i> (GPA 4.00/4.00)	08/2022 – Present
Advisor: Aidong Zhang	
Tsinghua University	China
B.A. in <i>English Language</i> (GPA 3.81/4.00)	08/2018 – 06/2022

Experience

Research Scientist Intern , Meta – New York, NY	05/2026 – 11/2026
Mentors: Xinyuan Zhang , Raffay Hamid , Xin Luna Dong	

- **Long-term memory**: Built MemLife^[1], which writes egocentric videos into time- and entity-anchored episodes for wearable assistants; it outperforms the best training-free baseline by 4.6–12.0% on four benchmarks.
- **RL post-training**: Proposed MemOpt to train Qwen3.5-9B and Qwen3.6-27B memory writers with GRPO and multi-granular rewards, improving MemLife by 2.7–5.0%; gains transfer across readers and memory systems.

Applied Scientist Intern , Amazon – Seattle, WA	05/2025 – 08/2025
Mentors: Hakan Karagul , Rutu Mulkar	

- **Product interest classification**: Fine-tuned Qwen3 classifiers that tag sports and outdoor products (ASINs) from their titles, descriptions, and other attributes, to improve annotation quality.
- **Customer interest reasoning**: Demonstrated that reasoning over histories (e.g., searches, clicks, purchases) beats direct classification; a GRPO-tuned Qwen3-0.6B surpasses Claude Opus 4 on verified internal data.

Graduate Research Assistant , University of Virginia – Charlottesville, VA	08/2022 – Present
---	-------------------

- **Agents & RAG**: Built RAG-Gym^[2] to post-train agents with process supervision (SFT, DPO, and PPO with a process reward model); the agents generalize better out of domain than Search-R1 and R1-Searcher. Also built the MIRAGE benchmark and MedRAG toolkit for medical RAG^[5].
- **Faithfulness & evaluation**: Built RAGLens^[8] to detect RAG hallucinations from sparse autoencoder features, and co-led MedCalc-Bench^[10], adopted by [Anthropic](#), [Fastino & NVIDIA](#), and [Stanford's MedHELM](#).
- **Interpretability & control**: Represented in-context demonstrations as an analytically derived weight update (Task Operator^[14]), and designed additive models with controllable feature interactions (NAE^[15]).

Publications

As of 09/2026: 1,780+ [Google Scholar](#) citations (h-index 12); 1,000+ [GitHub](#) stars and 290K+ [Hugging Face](#) downloads for released code, models, and data. * denotes equal contribution.

Agents, RAG & Post-training

- [1] **MemLife: Curating and Reasoning over Long-Term Egocentric Video Memories** [\[paper\]](#)
Guangzhi Xiong, Xinyuan Zhang, Xiao Yang, Hyokun Yun, Kai Zhang, Shiun-Zu Kuo, Hyeonjeong Ha, Xilun Chen, Kai Sun, Lucas Liang, Guangqiang Dong, Ejaz Ahmed, Ahmed A Aly, Anuj Kumar, Raffay Hamid, Aidong Zhang, Xin Luna Dong
Under Review, 2026

- [2] **Supervising the search process produces reliable and generalizable information-seeking agents** [[homepage](#)] [[paper](#)] [[code](#)] [[models](#)]
Guangzhi Xiong*, Qiao Jin*, Xiao Wang, Yin Fang, Haolin Liu, Yifan Yang, Fangyuan Chen, Zhixing Song, Dengyu Wang, Minjia Zhang, Zhiyong Lu, Aidong Zhang
Nature Communications (accepted in principle), 2026
120+ GitHub stars; 36 released checkpoints with 75K+ Hugging Face downloads
- [3] **Retrieving Counterfactuals Improves Visual In-Context Learning** [[homepage](#)] [[paper](#)] [[code](#)]
Guangzhi Xiong, Sanchit Sinha, Zhenghao He, Aidong Zhang
CVPR, 2026
- [4] **Improving Retrieval-Augmented Generation in Medicine with Iterative Follow-up Questions.** [[paper](#)]
Guangzhi Xiong*, Qiao Jin*, Xiao Wang, Minjia Zhang, Zhiyong Lu, Aidong Zhang
PSB, 2025
150+ citations
- [5] **Benchmarking retrieval-augmented generation for medicine.** [[homepage](#)] [[paper](#)] [[code](#)] [[corpus](#)]
Guangzhi Xiong*, Qiao Jin*, Zhiyong Lu, Aidong Zhang
ACL Findings, 2024
890+ citations; 800+ GitHub stars; 210K+ Hugging Face downloads of the retrieval corpora
- [6] **Cell-o1: training LLMs to solve single-cell reasoning puzzles with reinforcement learning** [[paper](#)] [[code](#)]
Yin Fang, Qiao Jin, Guangzhi Xiong, Bowen Jin, Xianrui Zhong, Siru Ouyang, Yifan Yang, Aidong Zhang, Jiawei Han, Zhiyong Lu
Bioinformatics, 2026

Faithfulness & Evaluation

- [7] **Rethinking Visual Attribution for Chest X-ray Reasoning in Large Vision Language Models** [[paper](#)] [[code](#)]
Guangzhi Xiong, Qiao Jin, Sanchit Sinha, Zhiyong Lu, Aidong Zhang
NeurIPS, 2026
- [8] **Toward Faithful Retrieval-Augmented Generation with Sparse Autoencoders** [[homepage](#)] [[paper](#)] [[code](#)]
Guangzhi Xiong, Zhenghao He, Bohan Liu, Sanchit Sinha, Aidong Zhang
ICLR, 2026
- [9] **Toward Reliable Scientific Hypothesis Generation: Evaluating Truthfulness and Hallucination in Large Language Models** [[paper](#)] [[code](#)]
Guangzhi Xiong, Eric Xie, Corey Williams, Myles Kim, Amir Hassan Shariatmadari, Sikun Guo, Stefan Bekiranov, Aidong Zhang
IJCAI, 2025
- [10] **MedCalc-Bench: Evaluating Large Language Models for Medical Calculations.** [[paper](#)] [[code](#)] [[dataset](#)]
Nikhil Khandekar*, Qiao Jin*, Guangzhi Xiong*, Soren Dunn, Serina S Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad W Safranek, Abid A Anwar, Andrew Zhang, Aidan Gilson, Maxwell B Singer, Amisha Dave, Andrew Taylor, Aidong Zhang, Qingyu Chen, Zhiyong Lu
NeurIPS D&B Track (Oral), 2024
Adopted by [Anthropic](#), [Fastino & NVIDIA](#), and [Stanford's MedHELM](#)
- [11] **IdeaBench: Benchmarking Large Language Models for Research Idea Generation** [[paper](#)]
Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Myles Kim, Corey M. Williams, Stefan Bekiranov, Aidong Zhang
KDD D&B Track, 2025
- [12] **MedCite: Can Language Models Generate Verifiable Text for Medicine?** [[paper](#)]
Xiao Wang, Mengjue Tan, Qiao Jin, Guangzhi Xiong, Yu Hu, Aidong Zhang, Zhiyong Lu, Minjia Zhang
ACL Findings, 2025

- [13] **Biomedical question answering: a survey of approaches and challenges.** [\[paper\]](#) [\[summary\]](#)
Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, Sheng Yu
ACM Computing Surveys, 2022

Interpretability & Control

- [14] **Capturing In-Context Learning Dynamics with Task Operators** [\[paper\]](#) [\[code\]](#)
Guangzhi Xiong, Zhenghao He, Bohan Liu, Sanchit Sinha, Wenqian Ye, Aidong Zhang
NeurIPS, 2026
- [15] **Neural Additive Experts: Context-Gated Experts for Controllable Model Additivity** [\[paper\]](#) [\[code\]](#)
Guangzhi Xiong, Sanchit Sinha, Aidong Zhang
AISTATS, 2026
- [16] **ProtoNAM: Prototypical Neural Additive Models for Interpretable Deep Tabular Learning.**
[\[paper\]](#) [\[code\]](#)
Guangzhi Xiong, Sanchit Sinha, Aidong Zhang
TKDD, 2025
- [17] **DeepGSEA: Explainable Deep Gene Set Enrichment Analysis for Single-cell Transcriptomic Data.**
[\[paper\]](#) [\[code\]](#)
Guangzhi Xiong, Nathan John LeRoy, Stefan Bekiranov, Aidong Zhang
Bioinformatics, 2024
- [18] **ProtoCell4P: an explainable prototype-based neural network for patient classification using single-cell RNA-seq.** [\[paper\]](#) [\[code\]](#)
Guangzhi Xiong, Stefan Bekiranov, Aidong Zhang
Bioinformatics, 2023
- [19] **CASL: Concept-Aligned Sparse Latents for Interpreting Diffusion Models** [\[paper\]](#)
Zhenghao He, Guangzhi Xiong, Boyang Wang, Sanchit Sinha, Aidong Zhang
NeurIPS, 2026
- [20] **Triggering Chain-of-Thought via Latent Feature Interventions in Large Language Models** [\[paper\]](#) [\[code\]](#)
Zhenghao He, Guangzhi Xiong, Bohan Liu, Sanchit Sinha, Aidong Zhang
EMNLP, 2026
- [21] **Towards Robustness against Typographic Attack with Training-free Concept Localization** [\[paper\]](#) [\[code\]](#)
Bohan Liu, Wenqian Ye, Guangzhi Xiong, Zhenghao He, Sanchit Sinha, Aidong Zhang
ECCV, 2026
- [22] **Concept-Residual Representation Expansion for Robustness to Spurious Correlations** [\[paper\]](#)
Eric Xie, Guangzhi Xiong, Wenqian Ye, Aidong Zhang
KDD, 2026
- [23] **Concept-RuleNet: Grounded Multi-Agent Neurosymbolic Reasoning in Vision Language Models** [\[paper\]](#)
Sanchit Sinha, Guangzhi Xiong, Zhenghao He, Aidong Zhang
AAAI (Oral), 2026
- [24] **COCO-Tree: Compositional Hierarchical Concept Trees for Enhanced Reasoning in Vision Language Models** [\[paper\]](#) [\[code\]](#)
Sanchit Sinha, Guangzhi Xiong, Aidong Zhang
EMNLP, 2025
- [25] **GCAV: A Global Concept Activation Vector Framework for Cross-Layer Consistency in Interpretability** [\[paper\]](#) [\[code\]](#)
Zhenghao He, Sanchit Sinha, Guangzhi Xiong, Aidong Zhang
ICCV, 2025

- [26] **ASCENT-ViT: Attention-based Scale-aware Concept Learning Framework for Enhanced Alignment in Vision Transformers** [[paper](#)] [[code](#)]
Sanchit Sinha, [Guangzhi Xiong](#), Aidong Zhang
IJCAI, 2025
- [27] **CoLiDR: Concept Learning using Aggregated Disentangled Representations.** [[paper](#)]
Sanchit Sinha, [Guangzhi Xiong](#), Aidong Zhang
KDD, 2024
- [28] **A Self-explaining Neural Architecture for Generalizable Concept Learning.** [[paper](#)]
Sanchit Sinha, [Guangzhi Xiong](#), Aidong Zhang
IJCAI, 2024

Technical Skills

Programming: Python, R, C/C++

Frameworks: PyTorch, Hugging Face Transformers, TRL, verl, OpenRLHF, PEFT (LoRA), LangChain

Teaching & Services

Teaching Assistant

- UVa CS 4774 - Machine Learning (Fall 2025)
- UVa CS 6316 - Machine Learning (Fall 2023, Fall 2024)
- UVa CS 6501-017 - GenAI Foundation and Application Overviews (Spring 2025)

Conference Reviewer

- ICLR, NeurIPS, ICML, AISTATS, IJCAI, AAAI Main & AIA Track, KDD, ACL ARR, ACM BCB

Journal Reviewer

- JAMIA, JBI, Bioinformatics, npj Digital Medicine, Nature Scientific Data, JHIR, JMIR, BMC Bioinformatics