Yixuan Xie

xie39@illinois.edu | 447-902-1956 | LinkedIn | Homepage
201 N Goodwin Ave, Urbana, IL 61801

Research Focus

Large Language Models • Reinforcement Learning • Retrieval-Augmented Generation • Multimodality

My research focuses on optimizing LLM systems through retrieval-augmented generation (RAG) and reinforcement learning, with recent work on multimodal RAG.

Education

University of Illinois Urbana-Champaign *08/2022 – Expected 05/2026*
B.S. in Computer Science & Statistics – Advisor: Prof. Jiawei Han *GPA: 3.9/4.0*

Courseworks: Text Mining with LLM, Machine Learning System, Deep Learning for Computer Vision, Database System

Experience

UIUC Data Mining Group | *Undergraduate Researcher* *08/2025 - Present*
Retrieval-Augmented Generation & RL & Agent Memory – Advised by Prof. Jiawei Han *Champaign, IL*

- Developed TaSR-RAG, a taxonomy-guided structured reasoning framework using relational triples for multi-hop evidence selection, outperforming strong RAG and structured-RAG baselines by up to 14%.
- Co-first author; led engineering and scaling of the TaSR-RAG pipeline, conducting evaluation on 7 multi-hop QA benchmarks and systematic ablations against RAG baselines (e.g.HippoRAG, GraphRAG, StructRAG).

Matsuo-Iwasawa Lab | *Research Assistant* *07/2025 - Present*
LLM Post-Training – Mentored by Dr. Qian Niu *Remote*

- Conducted two-stage post-training of lightweight expert models (Qwen3-4B, 1.7B), conducting SFT with both Full and LoRA-based updates, followed by reinforcement learning (GRPO) via LlamaFactory and Verl.
- Designed medical-domain reward functions and conducted single-node multi-GPU training (Slurm, ZeRO2), tuning hyperparameters to improve stability and convergence.

IBM-Illinois Discovery Accelerator Institute | *Research Assistant* *05/2025 - 08/2025*
Open-Schema Information Extraction – Advised by Prof. Jiawei Han *Champaign, IL*

- Conducted comparative analysis of training-free information extraction methods (RolePred, Code4Struct, ZOES, ODIE) versus LLM in-context learning across DocEE, MAVE, and NERRE datasets.
- Demonstrated that modern LLMs rival structured IE pipelines on standard benchmarks, revealing limitations of legacy evaluation and motivating more robust assessment methods.

DataLynn, Inc. | *Machine Learning Engineer Intern* *05/2024 - 08/2024*
Retrieval-Augmented Generation & Production AI Systems *Long Island City, NY*

- Designed and deployed a RAG pipeline using Milvus and Zilliz Cloud.
- Built and launched production AI systems, including an interview preparation agent (500+ sessions, 78% positive feedback) and a customer support chatbot (2,000+ inquiries).

Publications

† indicates equal contribution.

TaSR-RAG: Taxonomy-guided Structured Reasoning for Retrieval-Augmented Generation
Jiashuo Sun†, **Yixuan Xie**†, Jimeng Shi, Shaowen Wang, Jiawei Han
KDD 2026 (Under Review)

Selected Projects

Interactive Spider Solitaire Agent Environment | *Personal Project* *02/2026 - Ongoing*
A Casual Gameplay Environment for LLM-Driven AI Agents

- Built a Python Spider Solitaire environment with RESTful API interfaces to support sequential LLM agent interaction, enabling modular agent integration and automated gameplay evaluation across models.
- Established baseline evaluation protocol and leaderboard design to compare model performance, laying groundwork for trajectory collection, RL and memory-augmented agent research.

Token-Perplexity Signals for Stable RL Post-Training | *Team Lead* *09/2025 - 12/2025*

Machine Learning System Course Project – Mentored by Prof. Fan Lai [Tech Report]

- Investigated gradient-based and token-level diagnostic signals for RLHF dynamics, implementing pipeline instrumentation to capture objective-specific gradients and prefix perplexity during GRPO training.
- Systematically analyzed gradient similarity and perplexity distributions across model scales, revealing limited sensitivity of gradient conflict and substantial overlap in perplexity between correct and incorrect rollouts.

Taxonomy-Enhanced Scientific Retrieval | *Group Lead* *01/2025 - 05/2025*

CS497 Team Project – Advised by Prof. Jiawei Han [Tech Report]

- Investigated taxonomy-enhanced retrieval methods for scientific papers using LitSearch dataset (64K docs).
- Developed and evaluated multiple retrieval strategies—including taxonomy alignment, fusion embedding, and a fine-tuned encoder—demonstrating an improvement in Recall@5 from 39.0% to 47.5% (+8.5%).

Technical Skills

Languages: Python | C/C++

Core Frameworks & Tools: PyTorch | Verl | LlamaFactory | Slurm

Systems & Training: ZeRO/DeepSpeed | Retrieval-Augmented Generation | RLHF/RLVR

Others: Docker | MySQL | MongoDB | Kubernetes | FastAPI | Milvus | Flask | Neo4j

Activities & Honors

- **Hackathons** – (2024 & 2025)
- **Datathons** – (2024 & 2025)
- **Course Assistant** – (08/2023 - 12/2023)
- **Dean's List** – (Spring 2025, Fall 2024, Spring 2023)
- **Selected Talks:**
  - A Survey on RL for Large Reasoning Models (Community Talk, 02/2026)
  - Intro to Video RAG (Presentation, 11/2025)

Last updated: February 15, 2026