

Yuchen Liu

Seattle, WA | yql6113@gmail.com | 814-876-2346 | github.com/Upccccc | linkedin.com/in/liuyuche

Search/AI infrastructure engineer building production-scale indexing, hybrid retrieval, and co-first-author LLM/mech interp research.

EDUCATION

M.S. Computer Science	University of Pennsylvania, Philadelphia	GPA: 3.80/4.00	May 2025
B.S. Applied Economics	Pennsylvania State University, State College	GPA: 3.86/4.00	May 2023

WORK EXPERIENCE

Software Engineer, Search Infrastructure | DoorDash Aug 2025 - Present | Seattle, WA

- Owned core components of a **next-generation search ingestion platform**, re-architecting the indexing write path across **~40 production search stacks**, hundreds of sharded indexer instances, billions of indexed documents, and tens of TB of index storage.
- Designed a **streaming incremental-indexing framework** using **Kafka, Iceberg, Spark**, per-namespace watermarks, and exact consumer cutover offsets, sustaining millions of document updates per cycle while preventing missed or double-applied updates.
- Productionized a **Lucene-native hybrid retrieval layer** that fuses **BM25 keyword retrieval** with open-source dense embeddings into a single query plan, consolidating semantic search previously served through a standalone vector retrieval stack.
- Migrated **30+ production stacks** to modular **Jsonnet** configuration, eliminating **~56K lines** of duplicate config and reducing verification time by **~90%**; added pre-deployment validators and **Kyverno guardrails** to prevent capacity regressions.

Software Engineering Intern | Penn Medicine TissueLab Oct 2024 - May 2025 | Philadelphia, PA

- Built a **Python/FastAPI ML orchestration platform** for **PyTorch-based medical-imaging workflows**, enabling natural-language-driven segmentation and classification tasks.
- Implemented an event-driven **DAG workflow engine** for async task execution, processing **10K+ pathology images daily** and streaming **12M+ imaging events** through WebSockets.

SELECTED AI SYSTEMS PROJECTS

nanoRL

- Led an open-source **RL training framework** inside nanochat for objective design, rollout experiments, reproducible ablations, and **vLLM-backed inference serving**.

nanochat

- Built a hackable pretraining and chat stack for architecture definition, **FLOP-controlled ablations**, and reproducible LLM experiments.

llm-interp

- Authored reproducible interpretability scripts for model circuit research, including **attention-sink analysis** and **LLM decode indeterminism** reproductions.

RESEARCH

Do Value Vectors in Deep Layers Need Context from the Residual Stream? EMNLP 2026, under review

- Co-first author; co-developed the central research hypothesis that late-layer Transformer value vectors may encode **context-free token information**, challenging the assumption that deep-layer attention values require residual-stream context.
- Co-designed controlled architecture experiments and ablations across **135M/780M models** and **21 benchmarks**; proposed **Bank of Values**, a learned value-vector table that removes the **V-cache** while preserving downstream quality.

Vision Language Models Cannot Plan, but Can They Formalize?

ECCV 2026, under review

- Co-designed experiments testing whether **VLMs** can translate visual planning tasks into solver-executable **PDDL problem files**, separating perception, formalization, and planning ability.
- Built evaluation pipelines for visual state extraction, goal/constraint formalization, and solver-based correctness checking.

A co-evolving agentic AI system for medical imaging analysis

Under review

- Co-developed agentic medical-imaging workflows for **tool selection, planning, knowledge updates**, and human-in-the-loop co-evolution.

SKILLS

Languages: Python, Kotlin, Java, C++, JavaScript

ML/AI: PyTorch, vLLM, Transformers, RL, LLM evaluation, mechanistic interpretability

Search/Data Infra: Lucene, Kafka, Spark, Iceberg, Delta Lake, Redis, FastAPI

Systems/Cloud: Docker, Kubernetes, AWS, Spring Boot