

Xing Gao

Backend & Applied AI Engineer

xinggao6@illinois.edu | 302-765-7930 | xinggao.dev | github.com/happyaaa | linkedin.com/in/xing-gao

EDUCATION

University of Illinois Urbana-Champaign

Master of Computer Science

Champaign, IL

Aug 2024 – Dec 2025

University of Delaware

Bachelor of Science in Computer Science

Newark, DE

Aug 2017 – May 2021

EXPERIENCE

Vicino AI

Software Engineer, Applied AI

Seattle, WA

Jan 2026 – Present

- Shipped end-to-end **multimodal creation workflows** across React, FastAPI, async execution, billing, and asset persistence; onboarded 52 model/endpoint configurations across 8 providers.
- Hardened long-running AI jobs with **SQL-backed state**, Redis Streams, idempotent retries, and billing checkpoints; validated recovery after a mid-flight worker kill with no stuck tasks or duplicate billing.
- Load-tested graceful overload handling across 45K+ alpha requests with **zero server-side 5xx** and sub-100 ms p99 for served collaboration traffic under 429 load shedding.
- Introduced a shared **media-asset manifest** and contract tests to keep generated video, thumbnail, subtitle, and last-frame paths consistent across publish, clone, and snapshot workflows.

PayPal

Software Engineer (via Beaconfire Solutions)

Austin, TX

Oct 2021 – Dec 2023

- Developed **Java/Spring Boot backend components** for PayPal compliance workflows, supporting transaction/account screening, review-state transitions, policy-result mapping, and downstream integrations.
- Instrumented 5+ core components with **OpenTelemetry and SignalFx**, correlating traces, metrics, structured logs, and **Splunk** queries to isolate failures across workflow stages and downstream dependencies.
- Expanded regression, integration, and E2E coverage for compliance screening workflows, validating status transitions, policy-result mappings, downstream responses, and error paths.
- Triaged cross-service compliance failures by correlating request IDs with dependency latency, traces, logs, metrics, and downstream responses; coordinated remediation with engineering and compliance teams.

ADDITIONAL EXPERIENCE

University of Illinois Urbana-Champaign

Research Assistant – Backend Software Engineer

Champaign, IL

Aug 2025 – Dec 2025

- Extended a **FastAPI/Neo4j knowledge-graph backend** with material ingestion, human-reviewed publishing, and concept-mastery analytics over 125 course artifacts across 28 instructional units.
- Parallelized 3 LLM relationship-extraction agents, reducing the critical path from 6 sequential calls to 2 concurrent rounds; added retrieval from up to 8 context sources, entity resolution, and human validation.

Bili

Software Engineer Intern

Seattle, WA

Jul 2025 – Aug 2025

- Built a **LangGraph**-orchestrated Google Docs/Sheets pipeline with section-level diff/patch, JSON Schema validation, OAuth2 token refresh, idempotent retries, and audit logs.
- Reduced E2E latency by 30% through async I/O, fan-out/fan-in, API batching, and schema caching; added contract validation, adaptive rate limiting, state tracking, structured logs, and trace spans.

NaviData AI (Supstat Inc.)

Software Engineer Intern, Backend

New York, NY

May 2025 – Jun 2025

- Built a **FastAPI, Azure SQL, and Redis** recommendation engine with semantic search, retrieval/ranking, collaborative filtering, and query optimization; reduced P95 latency from 800ms to 200ms.
- Implemented **Pydantic schemas and SQLAlchemy data access**, Redis caching, audit logs, structured logging, and health checks for production recommendation APIs.

TECHNICAL SKILLS

Languages: Java, Python, Go, C++, JavaScript, TypeScript

Backend: Spring Boot, FastAPI, Next.js, React, Node.js, Express.js, Microservices

gRPC, Hibernate, REST APIs, Pydantic, SQLAlchemy

AI/Systems: Generative AI, LLM Applications, LangGraph, RAG, Multimodal AI, Agent Harnesses, Tool Use, Tool Orchestration, Context Engineering, Agent Evals, Evaluation Loops, Deterministic Guardrails, Selective Regeneration, Durable Execution, Model Routing, Embeddings, Vector Search, Distributed Systems

Data: PostgreSQL, MySQL, MongoDB, Snowflake, Azure SQL, Neo4j, pgvector, Redis, Redis Streams, Supabase

Cloud/Observability: AWS, Amazon SQS, GCP Cloud Run, Azure, Docker, AWS CDK, CI/CD, CloudWatch, OpenTelemetry, SignalFx, Splunk, Structured Logging, Distributed Tracing, Pytest, Playwright, Vitest