

Yufei Wang

kate003223@gmail.com || (+1) 213-675-7753 || www.linkedin.com/in/yufeiwang-usc || YufeiWang.github.io

EDUCATION

University of Southern California | Los Angeles, CA

Dec 2022 - Dec 2024

Master of Science, Applied Data Science

GPA: **3.9/4.0**

Relevant Coursework: Data mining, Machine learning, Deep learning / NLP, Artificial Intelligence

Qingdao University | Qingdao, CN

Sep 2018 - Jun 2022

Bachelor of Engineering, Computer Science

GPA: **3.8/4.0**

Relevant Coursework: Data structures & algorithms, Operating systems, Databases, Statistics, Computer networks

WORK EXPERIENCE

Software Engineer

Feb 2025 – Present

Causal Dynamics Lab

Seattle, WA

Project: Enterprise Agent Memory, Retrieval & Agent Harness | Python, LLM Agents, RAG, FastAPI, PostgreSQL

- Built a concurrent LLM-based memory extraction pipeline over 512k documents, storing embeddings, cue anchors, and entity facets in PostgreSQL; raised **throughput 3x** with buffered bulk writes, embedding pre-warming, and deferred deletes; integrated predictive retrieval cues, improving LoCoMo-Plus accuracy from **46% to 73%**.
- Implemented an offline memory-lifecycle pipeline (union-find clustering, cross-batch reconciliation, temporal fact supersession, access-decay scoring) and reduced memory usage by **82%** through blocked incremental similarity scans.
- Built a plan-based **multi-channel RAG retrieval pipeline** (HyDE/BM25/vector/cue, RRF, shared reranking) with 4 quality-latency tiers and LiteLLM routing; lifted LoCoMo accuracy from **83.8% to 93.3%** and EnterpriseRAG from **70.1% to 83.2%**.
- Replaced in-memory BM25 with Tantivy-based on-disk indexing; reduced index storage by **98%** and accelerated builds by **2.7x**.
- Built a tool-use retrieval agent with candidate adjudication, verification and enrichment, improving EnterpriseRAG-Bench from **83.2% to 90.9%** (**#1** on the [leaderboard](#)) and improved BEAM-10M accuracy from **55% to 71%**.
- Developed an asynchronous **Python/FastAPI + PydanticAI agent harness** for an **Electron desktop app**, with **MCP tool calling**, dynamic skills, human-in-the-loop approval-gated actions, and real-time event streaming for auditable agent workflows.
- Developed “My Day” with Gmail, Slack, Google Workspace, Jira, and Calendar integrations for task triage and reply drafting; built an ontology-driven knowledge system with typed relationships and incremental updates for personalized agent context.

Project: AI-Powered Root Cause Analysis (RCA) Agent | Go, Python, RabbitMQ, Docker, Kubernetes, AWS/GCP

- Designed a multi-stage **Python/ReAct** RCA agent orchestrating Kubernetes configurations and secrets, cloud infrastructure, hierarchical log templating, and metrics analysis for distributed systems; improved evaluation score from **56% to 91%**.
- Built an end-to-end, event-driven incident processing pipeline in **Go** with **RabbitMQ message queues**, PostgreSQL, and PagerDuty webhooks, automating alert ingestion, rule-based filtering, and fault-tolerant investigation workflows.
- Implemented a cloud-agnostic **observability and monitoring data layer** integrating **Elasticsearch**, AWS OpenSearch, and GCP Cloud Logging, automating log-schema discovery, severity-aware filtering, and cloud infrastructure dependency extraction.
- Developed a SOP execution framework using Semantic Kernel converting Markdown runbooks into validated plans with dependencies and conditional branches; integrated time-series anomaly detection using Z-scores and STL decomposition.
- Extended the agent with **skills** for multi-cloud/Kubernetes/Git CLIs, workload dependency analysis, deployment history, Prometheus queries, alert validation, and OpenDeepWiki repository knowledge; improved the mean score from **34% to 90%**.
- Built an evals pipeline on AIOpsLab, GKE, and OpenTelemetry to deploy microservices, inject faults, simulate traffic, and score runs with **LLM-as-a-Judge**; maintained 45 test scenarios covering cloud database, credential, API, and network failures.

Project: Multi-Cloud GraphRAG & AI Agent Platform | Go, Milvus, FalkorDB, FastAPI, WebSocket, GitHub Actions

- Built a Go GraphRAG pipeline mapping AWS/Azure/GCP resources into a knowledge graph with Text-to-Cypher retrieval.
- Automated evaluation framework in Python/FastAPI with RAGAS and Arize Phoenix, wired into **GitHub Actions CI/CD**.

SKILLS

Programming Languages: Python, Go, C/C++, SQL, TypeScript/JavaScript, Bash

Backend: FastAPI, Node.js, Spring Boot, REST APIs, gRPC, WebSocket, MySQL, Redis, MongoDB, RabbitMQ

Cloud & Infra: AWS, GCP, Azure, Kubernetes, Docker, Linux, Git, CI/CD (GitHub Actions), microservices

ML & Data: PyTorch, TensorFlow, Transformers, XGBoost, Spark/PySpark, LangChain/LangGraph, OpenAI API, Elasticsearch

LLM Systems: tool calling, MCP, RAG, vector DBs, embeddings, evals, guardrails, prompt engineering, fine-tuning (LoRA)