Chia Hui Yen

+1(412) 579-0806 | huiyenc@andrew.cmu.edu | hychia88.github.io

Professional Summary

Software Development Engineer specializing in ML/AI systems, building reliable, scalable services that put models into production. Experienced in data pipelines, retrieval/feature layers, and low-latency inference APIs, with focus on correctness, performance, and operability (testing, async serving, monitoring, safe rollouts). Skilled in deploying LLM/RAG and multimodal backends.

Education

Carnegie Mellon University Pittsburgh, PA
Master of Science in Computational Design (Focus: Applied / Production AI Systems) *May 2026*
- **Relevant Coursework:** ML in Production, Principles of Software Construction, Data Structures & Algorithms, Intro to Deep Learning, WebApp Development.
- **Award:** CMU Architecture Merit Scholarship.

Tsinghua University Beijing, China
Bachelor of Architecture *June 2024*
- **Award:** Tsinghua University Malaysian Outstanding Undergraduate Scholarship.

Technical Skills

Languages: Python, Java, JavaScript/TypeScript, SQL, HTML/CSS.
Backend & Systems: FastAPI, AsyncIO, Docker, PostgreSQL, AWS (EC2, S3, RDS), Kafka, Neo4j.
AI/LLM Systems: LangChain, LangGraph, RAG, FAISS, Hugging Face, LoRA/PEFT, Pydantic/JSON Schema.
Testing/DevOps/Observability: PyTest, GitHub Actions, Prometheus, Grafana, Git.

Work Experience

Lennar Corporation San Francisco, CA
Software Engineer Intern (AI/ML) *June 2025 – August 2025*
- **Automated Construction Content Processing at Scale:** Built FastAPI services for multimodal processing (image + text) using an AsyncIO concurrency model; enforced semaphore-based rate limiting + async batching for external LLM API calls, achieving **1,000+ images in 240s** with stable throughput, enabling faster on-site documentation/search and reducing manual triage effort
- **Search/Retrieval Service Engineering:** Implemented a retrieval layer (LangChain + FAISS) with hybrid ranking (dense + sparse + reranking) to deliver **sub-second** semantic search latency for downstream workflows.
- **Pipeline Throughput & Robustness:** Productionized an NLP processing pipeline (few-shot classification + clustering), generating **20,000+ embeddings in <300s**; improved reliability under noisy data via strict Pydantic schema validation and exponential-backoff retries for transient failures.

Projects

Real-Time Recommendation System | *Kafka, MLOps, Kubernetes (k3s)* October 2025
- Built a real-time user-event streaming pipeline on **Apache Kafka** with data quality gates (schema/type/range checks) and bad-record quarantine; monitored drift using **Population Stability Index (PSI)** on feature snapshots, improving model reliability and operability under live traffic
- Productionized retraining + deployment with **k3s CronJob** and **zero-downtime rollout** via atomic model hot-swap (versioned artifacts + symlink switch), enforced an online **latency eligibility gate** before promoting models.
- Built a hybrid recommender (CBF + UCF) with **RRF fusion** and **MMR re-ranking** for diversity; evaluated with leave-one-out Hit Rate plus service metrics (latency, artifact size).

UniNest: Scalable Search Backend | *AWS, Docker, PostgreSQL* | GitHub Repo April 2025
- **Hybrid Retrieval + Ranking:** Combined Postgres **BM25** with semantic vector retrieval and **RRF** fusion, improving Precision@10 by **16%** while maintaining **p95 latency ∼329ms**, **enabling faster, more accurate housing discovery for students**.
- **Multimodal Query Feature:** Enabled image-based search by generating structured tags/captions from user uploads and retrieving listings via text embeddings (image → tags → embedding → search).
- **Deployment + Observability:** Deployed JWT-secured services to AWS EC2 via Docker; instrumented request latency, sustaining ∼18 req/s throughput and ETL ingest of ∼410 properties/day.

AEC Interpreter System | *LoRA, LangGraph, Neo4j, FastMCP* | GitHub Repo February 2026 – April 2026
- **Solved Low-Data Training:** Engineered a procedural BIM-to-training pipeline, generating **933** augmented synthetic cases to overcome AEC domain data scarcity; designed a **9-condition modality masking** ablation and validated improvements with an **LLM-as-a-Judge** evaluation, enabling reliable visual grounding under noisy field inputs
- **Compound Agent Architecture:** Architected a dual-path LangGraph system that contrasts a ReAct agent with a deterministic Constraints-to-Query pipeline for auditability; isolated tools as stateless FastMCP microservices for independent deployment and hot-swapping of constraint models, ensuring hallucination-free retrieval queries.
- **Compound Agent Architecture:** Architected a dual-path LangGraph system contrasting a ReAct agent against a deterministic Constraints-to-Query pipeline for auditability with isolating tools as stateless MCP microservices via FastMCP, enabling independent deployment and hot-swapping the constraints models, ensure no-hallucination retrieval.
- **Ontology-Aware Graph Retrieval:** Engineered a Neo4j Graph-RAG system, executing deterministic Cypher queries over hierarchical BIM topologies (building → storey → space). Integrated CLIP-based visual reranking and a custom Search Space Reduction (SSR) metric, collapsing candidate pools by 98.9% and reducing latency from minutes to milliseconds.