

Hsiang Yu Huang (Anna)

Allston, MA | Tel: +1 617-319-5044 | huanganna1004@gmail.com | github.com/annaandmandy | linkedin.com/in/hsiangyuhuang

SUMMARY

AI Systems Engineer with a strong foundation in LLM orchestration, RAG pipelines, and Multi-Agent workflows. Experienced in building full-stack AI applications using FastAPI, Next.js, and AWS Serverless architecture. Proven track record in designing autonomous agents (LangGraph) and optimizing retrieval systems for semantic search. Winner of multiple hackathons for innovative AI-driven solutions.

SKILLS

- **Programming & Tools:** Python, SQL, R, Git, Linux, FastAPI, REST APIs
- **LLM & AI Engineer:** RAG Pipelines, Multi-Agent Systems (LangGraph), Vector Databases (Embeddings & Retrieval), Semantic Search, Prompt Engineering, LLM Orchestration
- **Machine Learning & Analytics:** PyTorch, TensorFlow, Scikit-Learn, Pandas, NumPy, Feature Engineering, Predictive Modeling, Time Series Analysis, NLP Statistical Analysis
- **Data & Cloud:** Azure, AWS(Lambda, S3, EC2, Step Functions), Supabase, SQL Database Design
- **Visualization & Frontend:** Power BI, Looker Studio, Matplotlib, React, Next.js

EDUCATION

Boston University <i>Master of Science in Data Science GPA: 3.57 / 4.0 Expected Dec. 2025</i> Relevant Courses: Deep Learning, Graduate Databases, Big Data Engineering, Time Series, Artificial Intelligence	Boston, MA
National Taiwan University of Science and Technology <i>BBA in Industrial Management and Bachelor Program of Finance, Minor in Computer Science GPA: 3.85 / 4.3 Jun. 2023</i> Relevant Courses: Algorithms, Machine Learning, Data Analytics, Statistics	Taipei, Taiwan

PROJECTS

LLM Platform – Multi-Agent Orchestration & Retrieval System <i>Research Project at BU BIT Lab Tech Stack: FastAPI, Next.js (React), MongoDB, LangGraph</i>	Boston, MA Sep. 2025 – Present
<ul style="list-style-type: none">◦ Architected a multi-agent system using LangGraph (Coordinator, Writer, Product, Memory, Vision agents) to orchestrate complex contextual retrieval and summarization workflows for Generative Engine Optimization (GEO) research.◦ Engineered a "Product-Enrichment" RAG pipeline that parses generated text to trigger Google Shopping API lookups via SerpAPI, rendering real-time interactive product cards with purchase links.◦ Built a high-performance observability backend with FastAPI and MongoDB to log interactions and 1536-dim embeddings, enabling granular analytics on model intent via semantic search.◦ Developed a model-agnostic console (OpenAI, Anthropic, Gemini, OpenRouter) with real-time event tracking to streamline the comparison of multi-model outputs for empirical marketing science studies.	
Creator – Multi-Agent AI Novel Generator (Personal Project) <i>Personal Project Tech Stack: LangGraph, Python, Supabase, Cloudflare Workers</i>	Boston, MA Dec. 2025 – Present
<ul style="list-style-type: none">◦ Architected a hierarchical multi-agent workflow using LangGraph to generate coherent long-form narratives, coordinating specialized agents (Director, Planner, Writer, Editor).◦ Engineered an automated quality assurance loop where the Editor agent evaluates prose against stylistic guidelines, triggering recursive rewrites to ensure consistency and quality.◦ Solved context management challenges by implementing state-driven interactions, enabling the system to maintain plot continuity across extended generation sessions.◦ View live website at https://dogblood-novel.dogblood-novel.workers.dev/.	
Creator – Boston Weekend Agent (Personal Project) <i>Personal Project Tech Stack: AWS Step Functions, Lambda, S3, Python</i>	Boston, MA Oct. 2025 – Present
<ul style="list-style-type: none">◦ Designed a fully automated serverless pipeline using AWS Step Functions to orchestrate data retrieval, LLM summarization, and report generation without manual intervention.◦ Achieved scalable content delivery by integrating Lambda triggers with S3 storage, ensuring reliable weekly report generation and deployment.◦ View live reports at hsiangyuhuang-anna.vercel.app/weekend_report.	

ADDITIONAL EXPERIENCE

Winner – DS+X Hackathon 2025 (Best Overall, HackBU 1st) – BU Spark! Boston, MA
Oct. 2025

- Developed [RhettSearch](#), an interactive research gamification platform connecting BU students with AI-driven paper discovery.
- Integrated semantic search, user gamification, and AI-generated recommendations using OpenAI API and OpenAlex API.

Efficient Open-Vocabulary Models for Low-Power Computer Vision (LPCV Competition) Boston, MA
Course: Deep Learning *Feb. 2025 – May. 2025*

- Optimized X-Decoder using DyT, SwiGLU, and linear attention to reduce inference cost.
- Evaluated on COCO and RefCOCOg datasets to align segmentation performance with low-power device requirements.
- Achieved a 7.5% GPU usage reduction and improving segmentation accuracy from 17 to 22 mIoU.

WORK EXPERIENCE

Research Assistant – Machine Learning for Sales Forecast in Graphic Card Manufacturing Taipei, Taiwan
NTUST Artificial Intelligence and Decision Analysis Lab *Nov. 2023 – Jul. 2024*

- Built an ARIMA–XGBoost forecasting pipeline, boosting R^2 from 8.3% to 73.4%.
- Developed a conditional rolling window model for adaptive, real-time predictions.
- Supported procurement and inventory planning using data-driven sales insights.

Research Assistant – Smart Vending Machine Shelf Optimization Taipei, Taiwan
NTUST Decision Analysis and Applied Statistics Lab *Apr. 2023 – Sep. 2023*

- Clustered product sales with K-means using metrics like mean, CV, revenue, and unit price.
- Built a classification tree to identify product-shelf performance patterns by price segment.
- Delivered actionable recommendations to improve shelf planning and profit optimization.