William Kim

williamcjk11@gmail.com | (773) 415-4606 | williamckim.com | linkedin.com/in/william-c-kim

EDUCATION

University of Pittsburgh — **Pittsburgh, PA**
Bachelor of Science in Computer Science — April 2024

PROFESSIONAL SUMMARY

Applied AI Engineer specializing in fine-tuning, deploying, and optimizing LLM systems across web, mobile, and cloud platforms. Experienced in RAG, OCR, and ONNX/quantization with OpenAI, Azure, and Ollama. Proven ability to ship production-grade AI products that improve accuracy, latency, and reliability.

RELEVANT EXPERIENCE

Independent AI Engineer — Production Deployments — **Barrington, IL**
Independent AI Projects — May 2023 – Present

- **Finsight — Domain-Specific Fine-Tuned LLM (Finance)**
  - Fine-tuned Mistral-7B and Qwen-3B with LoRA/PEFT + quantization, improving finance QA accuracy by **+25–30%**.
  - **Deployed production models** via **FastAPI + Docker** on **Cloud Run and Render**, cutting inference latency **2×** with ONNX export.
  - Designed an **evaluation pipeline (BLEU, ROUGE, F1)** for continuous benchmarking and optimization.
- **AI Research Copilot (Enterprise Doc-Chat)**
  - Built **provider-agnostic RAG pipeline** (FastAPI + pgvector + OpenAI/Ollama) enabling semantic retrieval and synthesis.
  - Implemented **failover and load-balancing** for **99.9% uptime** across OpenAI and local inference clusters.
  - Launched **production backend (Render)** and **Vercel frontend** with secure API proxying and modular architecture.
- **SplitChamp AI (Cross-Platform Bill Splitting App)**
  - Engineered **mobile app (Expo/React Native + FastAPI)** automating restaurant receipt parsing and bill splitting.
  - Integrated **Azure Document Intelligence OCR + GPT-4/5**, achieving **95% parsing accuracy** and **20% F1 gain**.
  - Automated **CI/CD with Expo EAS**, cutting release cycles from **2 days to 2 hours** and boosting deploy consistency.

SKILLS & CERTIFICATIONS

Languages: Python, SQL, C++, Java, JavaScript/TypeScript, HTML/CSS
AI & ML: PyTorch, Hugging Face, LoRA/PEFT, Quantization, ONNX, GPT-4/5, Claude, Ollama, Scikit-learn, OCR (Azure, OpenCV)
Cloud & MLOps: Azure, AWS, GCP, Docker, Cloud Run, Render, Vercel, Supabase, CI/CD, Git/GitHub, Linux/Unix
Frameworks: FastAPI, React Native, Expo, Next.js, TailwindCSS
Certifications: Azure AI Engineer (Microsoft, Exam Oct 2025); Applied AI Developer(IBM); Generative AI for Software Developers (IBM); Generative AI Fundamentals (Databricks)

INTERESTS

Competitive gymnast — active competitor; daily 4 AM training demonstrates discipline and resilience.