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** with hands-on experience **fine-tuning, deploying, and scaling LLM-powered systems** across web, mobile, and cloud. Skilled in **RAG, OCR, ONNX/quantization, and multi-provider inference** with OpenAI, Azure, and Ollama. Strong foundation in **Python, SQL, and cloud-native microservices** with a track record of shipping **production-grade AI pipelines** that deliver measurable improvements in accuracy, efficiency, and reliability.

RELEVANT EXPERIENCE

Founder & Lead AI Engineer — **Barrington, IL**
Self-Launched AI Applications — May 2023 – Present

- **Domain-Specific Fine-Tuned LLM (Finance)**
  - Fine-tuned open-source LLMs (Mistral-7B, Qwen-3B) using **LoRA/PEFT, quantization, and ONNX export** for optimized inference on finance-specific QA tasks.
  - Implemented an **evaluation framework (BLEU, ROUGE, F1)** to benchmark domain adaptation, improving task accuracy against baselines.
  - Deployed production-ready models with **FastAPI + Docker on Cloud Run/Render**, supported by a **Next.js interactive demo**.
- **AI Research Copilot (Enterprise Doc-Chat)**
  - Developed a **provider-agnostic RAG pipeline** (FastAPI + pgvector + OpenAI/Ollama) enabling semantic retrieval and multi-provider synthesis.
  - Deployed backend on **Render** with **Vercel frontend**, secured via **API proxying and provider failover** for enterprise reliability.
  - Delivered an **enterprise-extensible architecture** with modular embeddings, scalable search, and orchestration patterns.
- **SplitChamp AI (Cross-Platform Bill Splitting App)**
  - Built a **cross-platform mobile app** (Expo/React Native + FastAPI) for **automated restaurant receipt splitting**.
  - Integrated **Azure Document Intelligence OCR + GPT-4/5**, achieving **95%+ parsing accuracy** and reducing manual input for users.
  - Applied **post-processing (deduplication, tax/tip reconciliation)** to boost parsing F1 score by **20%**, and implemented **CI/CD with Expo EAS** to cut release cycles from 2 days → 2 hours.

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, Tesseract/OpenCV)
Cloud & MLOps: AWS, GCP, Docker, Cloud Run, Render, Vercel, Supabase, CI/CD pipelines, Git/GitHub, Linux/Unix
Frameworks: FastAPI, React Native, Expo, Next.js, TailwindCSS
Certifications: Azure Fundamentals (expected 2025) | AWS Solutions Architect – Associate (target 2025)

INTERESTS

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