

Javin Ahuja

College Park, MD, USA | javinahuja18805@gmail.com | 808-469-9684
javinahuja.com | github.com/jav359003 | linkedin.com/in/javin-ahuja-560100237

Education

University of Maryland, College Park, MD
B.S. Computer Science (Machine Learning Concentration)

Aug. 2023 – May 2027
GPA: 3.5/4.0

Skills

Languages & Frameworks: TypeScript, Python, JavaScript, Java, SQL, C, React, React Native, Next.js, FastAPI, Flask, Node.js, Spring Boot, PyTorch, NumPy
AI, Cloud & Data: RAG (hybrid, hierarchical), Agentic Systems, LLM Fine-Tuning (LoRA/SFT, Unsloth), Evaluation Frameworks, LangChain, Anthropic (Claude) API, OpenAI API, llama.cpp, YOLOv8, AWS, Supabase (Edge Functions/Deno, RLS), Docker, CI/CD, PostgreSQL, pgvector, FAISS

Professional Experience

Peptide AI, AI / Founding Engineer

May 2026 – Present

- Architected the production RAG system (**TypeScript**, **Supabase Edge Functions/Deno**) behind AI Chat: hybrid dense + **BM25** retrieval, **Voyage cross-encoder reranking**, and streamed **Claude (Anthropic API)** responses grounded per user.
- Built an agentic framework (**TypeScript**) reasoning over user data, wearable signals, and retrieved evidence, evaluated on agent trajectories rather than final output alone.
- Designed a privacy-safe semantic cache (**TypeScript**, **pgvector**) cutting response latency **~70% on cache hits (9s to under 3s)** at **\$0 marginal cost**, guarded by a **Deno** regression suite.
- Shipped cross-platform apps (**React**, **React Native**, **Capacitor**) weekly, driving **10,000+ downloads (+66% MoM)** and **85K+ daily App Store impressions (+850%)**; leading a **Swift/SwiftUI** rewrite.

Booz Allen Hamilton, Backend Engineer, App Dev Club

Jan. 2026 – May 2026

- Designed a two-pass detection algorithm (**DCT perceptual hashing**, **adaptive SSIM**) cutting redundant frame captures **~3x**; built a **FastAPI** extraction service with an **OpenCV OCR pipeline** for structured, resize-resilient video processing.

UMD Smith Financial Wellness Team, Software Engineer

Sep. 2025 – Dec. 2025

- Shipped a production app to **5,000+ students** end to end: **Flask REST APIs**, a **React Native** frontend, and **Python** pipelines running **10,000+ Monte Carlo simulations**.

Garden For Wildlife, Machine Learning Intern

Aug. 2024 – Aug. 2025

- Built and shipped a production RAG pipeline (**LangChain**, **FAISS**, **OpenAI API**) evaluated across **5,000+ real-world queries**, achieving a **34.5% accuracy improvement** and **15% latency reduction**.

DocuGuardian, Full Stack Engineer Intern

Dec. 2024 – Jul. 2025

- Built full-stack features (**Spring Boot APIs**, **PostgreSQL**, **TypeScript/React**) for secure document workflows; improved **release stability 30%** through **CI/CD** hardening and automated testing.
-

Projects

SynthDrive | Next.js, TypeScript, FastAPI, YOLOv8, GPT-4o, Supabase

Mar. 2026

- Built and deployed an AV perception test suite at **BigThink (UMD) x World Labs**, chaining **GPT-4o** scenario generation, **World Labs Marble 3D** worlds, **YOLOv8** detection, and **GPT-4o** safety analysis into a **0–100 Perception Difficulty Score**.
- Generated **3 navigable 3D worlds in parallel (~45s)** from a one-line prompt; deployed detection as a standalone **FastAPI** microservice on **Render**, with **Supabase** Auth and Postgres row-level security per user.

LegalLease | Python, PyTorch, Unsloth, FAISS, FastAPI, llama.cpp

Apr. 2026 – Present

- Fine-tuned **Qwen2.5-3B (LoRA r=16, Unsloth)** on **961 lease Q&A pairs** built via **GPT-4o** distillation, quantized to **Q4_K_M GGUF** for local inference, **lifting answer accuracy ~35% → ~55%**.
- Built **hierarchical FAISS retrieval** (doc → section → chunk) with **BM25** hybrid recall and cross-encoder reranking, scored by a custom harness over **112 questions on 12 lease documents: 99% retrieval recall (111/112), ~54% answer accuracy**.
- Built a lease **red-flag scanner** (YAML checks mapped to curated Maryland statutes) gating every flag behind **deterministic verbatim-quote grounding**, so hallucinated legal findings are impossible.

Tetherboard | React, TypeScript, Firebase, Gemini API

Apr. 2025

Built the **Gemini API** assistant for a collaborative knowledge platform at **Bitcamp 2025**, passing the active board plus all workspace boards as context so questions span the whole workspace.