Ali Benrami

972-266-3247 | abenrami06@gmail.com | github.com/AliBenrami | Alibenrami.com

Education

The University of Texas at Dallas Richardson, TX
Bachelor of Computer Science GPA: 3.98 *May 2027*

Experience

AIM Fundthesis Co-Mentor — Artificial Intelligence Mentorship Program August 2025 - Present
UT Dallas AIS *UT Dallas*

- Mentored a team of 6 students in the AIM program, guiding them from AI project ideation to deployment.
- Led weekly check-ins, code reviews, and technical support in ML and full-stack dev, resulting in a functional MVP demoed to judges and improved team communication skills.

AIS Inno Labs — AskTemoc August 2025 - Present
UT Dallas AIS *UT Dallas*

- Collaborating with a 6-member team under AIS Inno Labs to design a Retrieval-Augmented Generation (RAG) platform for UT Dallas students.
- Planned backend architecture with FastAPI, Pinecone, and LangChain, including document parsing, embeddings, and hybrid search (BM25 + vector).
- Defined evaluation strategy using DeepEval/RAGAS to guide iterative refinements to chunking, embeddings, and reranking.

Projects

RagTimeAPI | *FastAPI, Python, ChromaDB, Gemini API, Hugging Face, LangChain* July 2025 - Present

- An AI-powered RAG service that delivers 10× more relevant responses by combining LLM reasoning with semantic search.
- Built with FastAPI for 1,000+ req/s scalability and millions of embeddings in ChromaDB.
- Supports pluggable LLMs (e.g., Gemini) and customizable prompts, enabling 95%+ accuracy in context-heavy queries.
- Developers can integrate context-rich AI features in under 5 minutes.

YFin Dashboard | *Next.js, TypeScript, TanStack Query, Yahoo Finance API, Vercel* August 2025 - Present

- Developed custom 60fps charts with interactive tooltips, zoom, and multi-timeframe analysis (1W–20Y).
- Integrated Yahoo Finance API for <1s latency updates and 20+ years of historical data.
- Optimized with server-side sampling (80% data reduction) and TanStack Query caching (5-min stale window).
- Delivered a responsive, mobile-first UI with dark/light theme support.
- Deployed to Vercel, achieving 99.9% uptime and scalable serverless performance.

Pocket Secretary AI Scheduling Assistant | *Flutter, Python, Supabase, Gemini API* January - May 2025

- Built an AI-powered scheduling assistant with Flutter (cross-platform) and a Python backend.
- Integrated Google Calendar & Gemini API to convert natural language into structured events and reminders.
- Implemented real-time updates with Supabase/Firestore and support for collaborative scheduling.

Market Research Toolkit & API | *Python, FastAPI, Hugging Face Transformers, Prophet* August 2025 - Present

- Integrated news sentiment analysis and summarization using Hugging Face transformers for unstructured market data.
- Implemented time-series forecasting with Prophet to generate stock trend predictions.
- Designed clustering-based investor personas with visualizations for small equity universes.
- Delivered a streamlined API surfacing actionable insights, improving decision-making efficiency for retail investors.

Skills

Languages: TypeScript, Python, C/C++, SQL, Java
Frameworks: Next.js, PyTorch, scikit-learn, Flask, FastAPI, Flutter
Libraries: NumPy, Pandas, Hugging Face, LangChain
Databases: MongoDB, ChromaDB, Supabase, Firestore
Developer Tools: GitHub, Vercel