Mateo Alado

Bay Area, CA | (510) 543-1444 | aladomateo@gmail.com | http://linkedin.com/in/mateoalado | http://github.com/matesuu

SKILLS

- **Languages:** Python, TypeScript, C/C++, SQL, Java
- **AI/ML:** LLMs, PyTorch, NumPy, RAG, GraphRAG, Ollama, MCP, Agentic Engineering
- **Backend:** FastAPI, REST APIs, Node.js, Cloudflare Workers, Distributed Systems
- **Database:** PostgreSQL, MongoDB, Neo4j, Cloudflare D1
- **Tools:** Docker, AWS, Git

PROFESSIONAL EXPERIENCE

Machine Learning Researcher **Hayward, CA**
California State University, East Bay January 2026 – June 2026

- Benchmarked 6 RAG and GraphRAG frameworks on token usage, latency, win rate, F1 score and validation accuracy
- Engineered and deployed a full-stack RAG research assistant using FastAPI and Supabase for use among a team of 10 faculty researchers uploading 20+ academic papers and datasets.
- Developed a backend inference engine using PolyG, reducing average token usage by 79.7% compared to similar open source RAG frameworks

Open Source Contributor **Remote**
Independent August 2024 – January 2026

- Optimized async I/O, request handling, and database query patterns across 5 Python and TypeScript repositories, improving backend performance and reliability
- Reduced development cycle time by approximately 10% by improving CI workflows and code integration processes

PROJECTS

Multi-Model Coding Agent (TypeScript, Bun, Perplexity API, Cloudflare Workers, Cloudflare D1)

- Developed a terminal-native AI coding agent with multi-model orchestration (Codex, Claude Haiku/Sonnet/Opus), allowing for dynamic routing requests based on task complexity and token constraints
- Built a session-sharing infrastructure using Cloudflare Workers and D1 enabling collaborative coding via shared URLs
- Achieved an 88/100 on SWE benchmarks in multi-file debugging, async concurrency, and SQLAlchemy migrations, outperforming Codex (72) and Qwen 397B (58)

GraphRAG Research QA Platform (RAG, Python, TypeScript, React, OpenAI API, FastAPI, Neo4j, Supabase)

- Architected a RAG-based inference engine that leverages external knowledge graphs for adaptive graph traversal, optimizing long-context prompting and reducing model hallucinations
- Built an interactive chat interface with React and Typescript supporting context-aware multi-turn conversations with real-time streaming responses and user authentication
- Benchmarked against GraphRAG and reasoning-based baselines, achieving highest F1 score (0.71) and 89% hit rate while reducing token usage by 75.5% compared to Microsoft GraphRAG and 93.3% compared to Graph-CoT

Visual Speech Recognition Engine (Python, PyTorch, NumPy, CNN, MediaPipe, OpenCV)

- Engineered a real-time lip-reading engine using MediaPipe Face Mesh and OpenCV webcam capture for classifying speech patterns for audibly impaired individuals
- Implemented a complete data collection and training workflow with CUDA/MPS/CPU device selection, supporting FPS tracking, confidence display, and webcam overlay feedback.
- Achieved a 94% validation accuracy across 24 speech classes through extensive deep learning reinforcement

EDUCATION

California State University, East Bay **December 2026**
Bachelor of Science, Computer Science

- **Distinctions and Honors:** HackHayward 2026 Winner, Dean's Honor List
- **Relevant Coursework:** Data Structures and Algorithms, Operating Systems, Computer Networks, Machine Learning, Natural Language Processing, Software Engineering, Web Development, Linear Algebra