

FELIPE VELASQUEZ

Foster City, CA • +1 (650) 430-6272 • fevelasquin@gmail.com • felipevelasquez.dev • linkedin.com/in/felipe-velasquez-703828300

Software engineer who builds complete products end to end — backend, frontend, and the AI that runs inside them. Founder of Munveo, a private on-premises AI search appliance, with other tools running in production today.

EDUCATION

Greeneville, TNTusculum UniversityMay 2025

- **Major:** B.S. Computer Science — GPA 3.85, Magna Cum Laude
- **Coursework:** Data Structures & Algorithms, Machine Learning, Operating Systems, Software Engineering, Databases, Cybersecurity
- **Athletics:** NCAA Division II Men's Tennis

PROFESSIONAL EXPERIENCE

Founder & Software EngineerMunveo (startup)Aug 2025 – Present

munveo.com: private AI document search that runs entirely on-premises · pilot program in preparation

- Founded Munveo, an on-premises AI search appliance for firms that can't send confidential files to the cloud: plain-language questions, page-level citations, designed for zero outbound traffic.
- Found that semantic search alone missed exact reference numbers while keyword search missed synonyms, so moved retrieval to a hybrid design: FAISS embeddings and SQLite BM25 run in parallel and merge through RRF, so both kinds of query land.
- Exposed a gap document-level recall had hidden — the right document reached the model 93% of the time, the passage with the answer only 66% — by building a chunk-level metric and an 80-question benchmark over a 500+ document corpus.
- Cut query latency ~49% (12.7s → 6.5s) by profiling every pipeline stage, removing redundant LLM calls, and splitting work across models by task: Qwen 2.5 handles routing and classification, Llama 3.1 only writes the final answer.
- Cut silent OCR corruption from 83% to 7% across real scanned records by benchmarking OCR engines page-by-page against source images, after finding the system scored corrupted and clean extractions identically.
- *Utilized:* Python, FastAPI, SQLite, FAISS, BM25, Ollama, Tesseract, Alembic, Docker, Next.js, TypeScript

Software Engineer (Contract)ProyesingJun 2026 – Sep 2026

Remote · Bogotá, Colombia · CUBICA: construction drawing analysis system

- Automated a quantity takeoff that took ~6 hours by hand, now in use at Proyesing: DWG-to-DXF conversion, layer and geometry parsing, element classification, and Excel export.
- Kept every quantity reproducible by limiting rules and ML to classification and computing quantities in a deterministic engine, since errors cost money on site.
- Built human-in-the-loop review into the workflow — engineers confirm classifications and check computed quantities — so every correction becomes training data with tracked provenance and the system keeps sharpening with use.
- *Utilized:* Python, DXF/DWG, ODA, NumPy, JSONL, pytest

IT Systems & AI Automation SpecialistTechPro BuildersSep 2025 – Jun 2026

Part-time · San Carlos, CA

- Saved ~10 hours/week and \$1,500+/month by automating expense tracking, inventory, and reporting workflows with Airtable and Zapier.
- Generated \$20K/year in recurring revenue by creating an annual maintenance plan covering every unit TechPro had installed, then selling it and scheduling the visits.
- Built a private document search app (FastAPI, Next.js, SQLite) that staff used to digitize an archive of ~700–800 job sites going back to 2017, finding project files locally through OCR and a local LLM — the prototype Munveo grew out of.
- Kept 15–18 office and field staff on shared project files by setting up the company's network and storage: an encrypted network, a NAS serving as private cloud, and network and physical security.

Technology & Digital OperationsEvolution BurlingameMay 2025 – Present

Intern May 2024 – Aug 2024, then Staff · Peninsula Tennis Club · Burlingame, CA

- Manage the club's customer databases in MySQL, keeping member data organized and reliable for staff.
- Maintain a 5,000+ customer network shared with partner tennis shop Tennis Station, keeping records clean and usable for outreach.

SOFTWARE PROJECTS

LLM Performance Benchmarking Study — Academic Symposium, Tusculum UniversitySpring 2025

- Found that no single model led overall — each won on different task types — by benchmarking ChatGPT, Claude, DeepSeek, and Gemini on standardized tasks for latency, accuracy, and suitability.

Boone Lake Water Quality Platform — Full-Stack DeveloperSpring 2025

- Built a PHP/MySQL platform, with a 10–15 person team, where researchers publish water-quality findings across 8 lake zones to residents, with threshold alerts, dashboards, and login roles (viewer, editor, admin).
- *Utilized:* PHP, MySQL, JavaScript, HTML/CSS

SKILLS

AI / LLM: Retrieval-Augmented Generation (RAG), LLM evaluation (evals, benchmarks), embeddings, vector & hybrid search (FAISS, BM25, RRF), local LLM inference (Ollama), OCR, prompt engineering

Languages: Python, JavaScript, TypeScript, SQL, PHP, HTML/CSS

Frameworks & Tools: FastAPI, REST APIs, Next.js, React, SQLite, MySQL, Alembic, Docker, pytest, Playwright, Git, Linux