# Jeevan Bhatta

jeevanbhattacs@gmail.com | (628) 222-8526 | San Francisco, CA | linkedin.com/in/jeevancs | github.com/jeevanbhatta

## EDUCATION

**MINERVA UNIVERSITY, San Francisco, GPA: 3.94/4.0** May 2026
Bachelor of Science in Applied Computer Science & AI; Second Major: Economics & Society

## SKILLS

**Languages:** Python, SQL, JavaScript/TypeScript, Java
**Evals & ML:** LLM evaluation, rubric and metric design with domain experts, paired A/B experiments, data pipelines, RAG, retrieval and reranking, fine-tuning, agents and tool calling, embeddings, vector search, Neo4j
**Frameworks:** Flask, FastAPI, Django, React, Next.js, Node.js
**DevOps & Cloud:** AWS, GCP, Azure, Docker, CI/CD, PostgreSQL, Datadog

## WORK EXPERIENCE

**Software Engineer,** *TaxGPT, San Francisco, CA* *August 2026 – Present*

- Own the evaluation system for our AI agents: defined 4 to 5 metrics per product with tax experts and grade answers against 1,000+ questions they wrote, offline and in production
- Run paired A/B trials against a control arm with a measured noise floor; the results set the product's cost basis and showed that an apparent accuracy gain came from a model change
- Traced a production log error to a retrieval fault and fixed it, bringing large gains in answer accuracy and citation quality as measured on live traffic
- Built ingestion pipelines for AI agents used by tens of thousands of tax professionals: crawled, chunked and embedded 11 state tax jurisdictions, with double-digit gains in citation quality and accuracy
- Unified per-state pipelines into one parameterized framework in 11 days: a new jurisdiction is now a module and a config block, covering more states in 14% fewer lines
- Added reranking inside oversized source pages and a cheaper listwise reranker, since adopted by a second product
- Run production: cut p95 latency by about 70%, added retries that clear most upstream model failures, and launched an MCP server approved in the Claude directory in 31 days

**Co-Founder and CEO,** *InstantApply, San Francisco, CA* *March 2025 – December 2025*

- Built a full-stack job platform (React, Flask, PostgreSQL) with LLM-driven Playwright agents that filled multi-step applications on data from 30+ sources; paid beta of 30+ users averaging 6 interviews each

**Software Engineering Intern,** *NAIC, Kansas City, MO* *June 2025 – August 2025*

- Fine-tuned SaulLM-7B, a legal-domain LLM, on 100,000+ insurance filings, increasing automated testing accuracy by 30% relative to generic LLM inferencing
- Engineered an end-to-end and unit testing infrastructure from the ground up using PyTest and Playwright for a platform used by 20,000+ regulators, reducing post-release regressions

**Quantum SWE Intern,** *Inter-American Open University, Argentina* *October 2024 – April 2025*

- Architected quantum risk models, improving algorithm efficiency by up to 30%, and constructed ancilla circuits simulating non-unitary operations with 90%+ accuracy versus classical methods

**Software Engineer Intern,** *Expatrio, Berlin, Germany* *January 2024 – April 2024*

- Led a team of 5 to build an LLM-powered recommendation engine with 20+ parameters, using Retrieval-Augmented Generation over 1,700 data sources

## HONORS & ACTIVITIES

- Winner, MCP/A2A Hackathon @ AWS (1st / 156): Co-created VoyAgent, an end-to-end trip planner
- Silverstream AI Award, Useful Agents Hackathon w/ NEAR AI (1st / 100+ teams): Developed InstantApply
- Recipient, NAIC Foundation Scholarship ($5,000); Certified SAFe 6 Practitioner, Scaled Agile