

# Bill Zhang

408-585-8267 | [jzhang71@usc.edu](mailto:jzhang71@usc.edu) | [in/bill-zhang1/](https://in/bill-zhang1/) | [github.com/IdkwhatImD0ing/](https://github.com/IdkwhatImD0ing/) | [billzhang.tech](https://billzhang.tech)

## EDUCATION

### University of Southern California

*Master of Science in Computer Science - Artificial Intelligence*

### University of California, Santa Cruz

*Bachelor of Science in Computer Science*

August 2023 – May 2025

*Los Angeles, CA*

September 2020 – March 2023

*Santa Cruz, CA*

## EXPERIENCE

### Applied AI Engineer

June 2025 – Present

*Scale AI*

*San Francisco, CA*

- Shipped the **TIME AI Agent** on TIME.com end-to-end (TIME APIs, OpenAI, Scale tooling), delivering RAG-powered search, summarization, and 13-language translation across TIME's 102-year archive at **~10 req/min**.
- Hardened production safety by implementing policy guardrails (input filtering, refusal logic, prompt-injection defenses), increasing adversarial prompt blocking **30% → 95%**.
- Built an agentic medical claims system now running in production, semi-autonomously resolving **20,000+ claims/week**; improved end-to-end accuracy **30% → 90%** via multi-agent routing, structured tool schemas, and targeted prompt refinement.
- Designed an offline eval harness (**300 golden traces** + judge-LLM rubric) and ran weekly regressions, driving audit accuracy to **~90%** with only **~10% human review**.

### AI Engineer

August 2023 – June 2025

*RingCentral*

*Belmont, CA*

- Owned an end-to-end LLM analytics + eval pipeline (OpenAI, Pandas, GCP) scaling from **1-2% → 100%** coverage of **10,000+ weekly** support chats; cut processing time **8 hours → 1 hour**.
- Improved accuracy **90% → 95%** and cut agent hand-offs **40% → 13%** by adding rubric-based retrieval scoring and weekly monitoring (**20 Looker** dashboards).

### Software Engineer Intern

June 2023 – August 2023

*RingCentral*

*Belmont, CA*

- Shipped an LLM support chatbot with retrieval (Weaviate + embeddings), improving first-contact resolution **5% → 50%**.
- Won **1st place** in an internal hackathon building a webinar networking/contact-card product; secured adoption buy-in from **2 C-level execs** and **5 PMs** via stakeholder demos.

## PROJECTS

### Slug Loop | *React, Material UI, ExpressJS, C, Azure*

March 2023 – Present

- Served **20,000+ users** by shipping a real-time UCSC bus tracking web app with a React frontend, ExpressJS APIs, and Azure deployment.
- Placed **Global Top 10** in the Google Developer Student Challenge out of **2,000+ projects** (2023).

### DispatchAI | *NextJS, FastAPI, Twilio, RetellAI, Hume, Google Maps*

June 2024 – Present

- Built and demoed a low-latency voice agent for emergency call intake + escalation (Twilio + FastAPI + NextJS), including intent routing and structured handoff to responders.
- Won **Grand Prize (\$50k)** at UC Berkeley AI Hackathon (**900+ participants**) by integrating RetellAI + Hume and deploying an end-to-end real-time pipeline (fine-tuned Mistral 7B).

### TalkToABank | *NextJS, OpenAI, RetellAI*

November 2024

- Won **1st Overall** at HackUTD (**1,000+ participants**) by building a phone-based AI banking assistant with multi-agent routing (NextJS, OpenAI, RetellAI).

## TECHNICAL SKILLS

**Languages:** Python, JavaScript, TypeScript, C/C++, Java, HTML, CSS, Dart

**Machine Learning:** PyTorch, TensorFlow, Keras, NumPy, Pandas, Scikit Learn, Matplotlib, OpenCV

**WebDev/Cloud:** React, NextJS, Vue, Angular, Flask, FastAPI, ExpressJS, Azure, GCP, AWS, PostgreSQL, GraphQL

**Tools and Libraries:** OpenAI, LangChain, LlamaIndex, Git, Docker, Firebase, Supabase, MongoDB