

Soma Sundar Mamilla

Sunnyvale, CA, USA | somasundar.m1@outlook.com | <https://www.linkedin.com/in/soma-mamilla> | <https://www.github.com/Sundar6264-ux> | <https://somasundar.dev>

SUMMARY

Software engineer with 4.5+ years of experience in trading systems, retail automation and AI products. Spent two years as the sole full-time developer on a C matching engine handling 5-6 million transactions a day at the National Stock Exchange. Now builds and runs automation, a staff app on the App Store and Claude-powered tools for retail clients, owning each system from requirements to production.

CORE SKILLS

Languages: Python, C, C++, SQL, JavaScript, Bash, Dart, Embedded C, Verilog HDL

Backend & APIs: FastAPI, Flask, Django, SQLAlchemy, Alembic, Pydantic, REST APIs, Webhooks, OAuth2, Role-based access control, SQLite, Redis

AI & Machine Learning: Anthropic Claude API, Agentic tool use, Structured outputs, Prompt engineering, LangChain, Google OCR, PyTorch, CUDA, Self-supervised learning, Pandas, NumPy, Scikit-learn

Cloud & DevOps: AWS (Lambda, S3, API Gateway, CloudWatch, IAM), Microsoft Azure, Docker, GitHub Actions, Jenkins, Git, SVN, Tailscale, systemd/launchd

Low-Latency Systems: Linux, Shared memory (SMFS), LBM messaging, TCP/IP, Multicast, GDB, perf, Valgrind, CPU pinning and IRQ balancing

Testing & Practices: PyTest, Integration testing, Swagger/OpenAPI, Code review, Agile/Scrum, Claude Code (AI-assisted development), Technical documentation

EXPERIENCE

Software Engineer, KloudC LLC

Dec 2025 to Present

Backend engineer on a 15-person distributed team building systems for B2B retail and FinTech clients, turning needs like multi-vendor purchasing and inventory reconciliation into running cloud services.

- Replaced a simple sales-average reorder method with a forecast that accounts for out-of-stock days, recent sales and the day of the week; in a backtest over three periods it modestly reduced forecast error (WAPE 0.475 to 0.462) and under-forecasting (bias -12.6% to -7.8%).
- Replaced a 3-hour manual purchasing audit across 40 vendors with an automated order system that finishes in 2 minutes, using 30-day sales trends and delivering results by email and Google Drive.
- Tested time-series forecasting (Nixtla, Prophet) on the store's real sales data, found it too unstable for production, and shipped a simpler deterministic forecast because it was more reliable.
- Since Aug 2026, run an in-house operations platform for a retail client alongside its Toast POS (FastAPI backend, React dashboard, Flutter staff app) used daily by store staff; own requirements, design, testing and releases, and directed Claude Code agents that wrote most of the code.
- Built the sales and inventory data feeds from Toast (orders, menus, stock, labor) with nightly backfills checked against Toast's own totals; they have processed over 110,000 sold items and 570,000 inventory snapshots across 120 days.
- Added an "Ask Inventory" assistant on the Claude API that answers staff questions in plain English through 14 read-only reports, sending simple questions to a cheaper model and harder ones to a stronger one, plus an AI-written weekly summary of the store's numbers.
- Added a staff hub with tasks and photo proof, announcements with read receipts and team chat that meets Apple's rules for user content, plus push reminders for orders and closing and per-user access control across 19 features.

Software Engineer Intern, KloudC LLC

Jul 2024 to Apr 2025

Primary engineer for the data sync behind a 6,000+ SKU retail inventory, in a small team covering everything from API middleware to the company website.

- Built the Instacart sync (Python, Pandas, SFTP) that sends only the last 24 hours of Clover changes, applies price markups and flags out-of-stock items, running on a schedule with no manual steps.
- Cut CI build and test time 60% (20 min to 8 min) with parallel jobs and caching, and automated S3-triggered Lambda reporting, saving approximately 15 engineering hours per week at 99.9% data accuracy.
- Learned LangChain and Google's OCR APIs in two weeks to clear a critical invoice backlog, and became proficient with GitHub Actions CI/CD over roughly three months of self-directed learning.

Systems Engineer, Tata Consultancy Services (embedded at National Stock Exchange, India)

Jul 2021 to Aug 2023

Sole full-time developer on a 9-layer C matching engine for the Commodities and Derivatives division, handling 5-6 million transactions per day.

- Used the load-testing tool to find sources of latency noise, cutting P99 latency jitter 30% with CPU pinning and IRQ balancing and reducing end-to-end latency by 0.5 to 1.0ms.
- Audited 500+ message types and updated shared-memory structures in all 9 layers for the encryption rollout with 100% binary compatibility, where a single misaligned byte could crash the engine during trading hours.
- Owned the schema and data integrity of the engine's shared-memory tables (SMFS), which pass data between the Network, Market and Data layers in under a microsecond over LBM messaging.
- Replaced manual SVN deployments (5 to 10 min for minor updates, up to 1 hr for a fresh setup) with the division's first Jenkins pipelines, cutting deployment to under 5 minutes, a 90% improvement.
- Wrote SOPs for and led 40 to 50 disaster-recovery drills over two years with a 0% rollover-failure rate, covering failovers and server migrations, backed by daily post-market backups for 100% data retention.

Junior Software Engineer, NexusTech Solutions

Jul 2020 to Jun 2021

- Helped set up CI/CD in GitHub Actions, adding automated tests to the deployment process.
- Wrote PyTest unit tests for backend services.

PROJECTS

Tailorbird: AI Resume Tailoring Platform

Personal project, Jul 2026 to Present, live in production

- Cut wasted API spend by rejecting poor-fit jobs before any paid tailoring call, after cost experiments showed that stretch-match jobs, not resume size or job-description length, drive reasoning-token cost.
- Saved paid API responses that fail to parse, such as ones cut off at the token limit, to a recoverable file instead of losing the work.
- Found and fixed a bug where a "<" in resume text cut bullets off mid-sentence, caught by comparing renders before and after rather than by reading code alone.

Flockara Job Portal: Private Job Discovery and Matching

Personal project, Sep 2026 to Present, live at jobs.flockara.com

- Launched a private job-search portal that matches live listings against a candidate's resume, with relative-fit scores and filters for title, country, workplace, seniority, posting age, pay and source.
- Used Claude Code to speed up implementation while defining the matching behavior, privacy rules, acceptance criteria and production checks; runs alongside Tailorbird in the Flockara product family.
- Ran quality audits with synthetic resumes and filter combinations to expose bad high-scoring matches, stale or duplicate listings and slow queries, turning each failure into matching and regression-test requirements.

Seeing the Unseen: Stego-Aware Self-Supervised Learning

M.S. Capstone, sole researcher/developer

- Delivered a 60-page thesis and model with 78% top-1 accuracy on CIFAR-10, under 1% accuracy drop under attack (versus a 6-point drop for the standard baseline) and 63% fewer FLOPs than ResNet-18 (1.5G vs. 4.1G).
- Designed Stego-BT, a self-supervised model based on Barlow Twins that stays accurate under hidden LSB steganographic changes that break standard models.
- Trained it in PyTorch on NVIDIA T4 GPUs for 800 epochs on CIFAR-10, using mixed precision for 1.8x higher throughput within a 16GB memory budget.

EDUCATION

M.S., Computer Engineering. University of Bridgeport, Bridgeport, CT. May 2025

B.Tech, Electronics and Communication Engineering. Malla Reddy College of Engineering, Hyderabad, India. Jul 2021

CERTIFICATIONS

AWS Academy Cloud Foundations. AWS Academy Cloud Developing. Microsoft Azure AZ-900. Certified in Geoprocessing using Python (Indian Institute of Remote Sensing).