Swapnil Sharma

551-998-8215 | swapnil.sh@nyu.edu | linkedin.com/in/swapsharma | github.com/swappysh

Summary

AI/ML engineer with 6+ years spanning backend, NLP, and production **LLM systems**, **multi-agent architectures**, and **evaluation infrastructure**. Founding team member at HippocraticAI (unicorn), scaling to **1M+ patient calls/week**; at Aleph building an **agentic FP&A assistant** that surfaces data issues, reasons over financial models, and recommends targeted workflow fixes.

Education

New York University — New York, NY
Master of Science in Computer Science — *Aug 2021 – May 2023*

Indian Institute of Technology — Mandi, India
Bachelor of Technology in Computer Science and Engineering — *Aug 2013 – May 2017*

Professional Experience

Aleph (Series B, AI-native FP&A)
Senior Research Engineer — *Oct 2025 – Present*

- Built **LLM-as-a-judge** evals and tooling for Aleph's agentic FP&A assistant across chart generation and natural language financial queries; helped stabilize a production customer eval suite and improve one client's correctness by **∼15 percentage points**.
- Implemented key capabilities for an agentic FP&A assistant on **Google ADK** that surfaces data-fix and modeling recommendations; delivered action approval flow (human-in-the-loop) for safe sheet/matrix writes, metric-aware financial reasoning, extensible skills, and **MCP/tool-calling**.
- Helped cut end-to-end response latency from **>10 min** to **<1 min**; reduced peak-load error rate from **90%** to **<5%** and **Docker** build time by **50x** for the chart/image service.

HippocraticAI
AI Engineer, Senior AI Engineer — *June 2023 – Sep 2025*

- Founding team member scaling HippocraticAI from 0 to **unicorn status**; co-architected the initial product and **Polaris multi-agent LLM constellation** for safety-critical, long-form patient voice conversations.
- Led development of core specialist agents (**Checklist, Labs & Vitals, Nutrition, Policy**) and a Conversation Navigation agent for context-window-aware flow management; applied **RAG** and **tool-calling** to enable empathetic, human-like patient interactions.
- Promoted to **Senior AI Engineer** to lead LLM fine-tuning; improved DOB extraction accuracy from 70% to **>90%** by fine-tuning **70B and 405B-scale models**, including **quantized/QLoRA** variants, via **PEFT** using **LLaMA Factory** and **TRL** on multi-cluster GPUs.
- Platform scaled to **1M+ patient calls/week**; built evaluation frameworks and an **automated retraining pipeline** (backend + UI) for continuous model improvement; co-authored the Polaris LLM paper ([arXiv]).

Microsoft
Applied Scientist — *May 2022 – Aug 2022*

- Applied **OpenAI Codex** to automate unit test generation for **JavaScript**, achieving an average **45% high-quality test case** rate per file.
- Shipped the pipeline as a **VS Code extension** (**TypeScript**), enabling test generation directly from the editor.

Nference
Data Scientist | Senior Data Scientist — *May 2019 – Aug 2021*

- Engineered efficient storage and retrieval pipeline for structured biomedical datasets, directly enabling new client onboarding and driving **∼15% incremental revenue**.
- Developed hybrid biomedical search platform in **Go** with **NLP**-driven query translation (**NER**, dependency parsing); improved result ranking via TF-IDF, PMI, and importance sampling.

Directi
Platform Developer | Zeta – Payments And Accounting Team — *Aug 2017 – May 2019*

- Engineered high-throughput payments backend in **Java**; improved **transaction latency** (∼ **5x**) for ∼ **15M TPD** system via DB-level optimizations and Redis caching.

Research & Publications

Stochastic Code Generation [arXiv] | NYU (Dr. Ranganath), 2022 — TimeControl autoencoder-decoder with contrastive learning for semantic code representations; comparable to CodeParrot on HumanEval.

Stereotypical Biases in BERT Models [arXiv] | NYU (Dr. Bowman), 2022 — Analyzed bias mitigation in ELECTRA, DeBERTa, DistilBERT vs. BERT on StereoSet & CrowS-Pairs; measurable reduction, biases persist.

Technical Skills

Languages: Python (primary), TypeScript, SQL, Java, Go
AI/ML: PyTorch, LLM Fine-tuning (QLoRA, LoRA, PEFT), LLaMA Factory, TRL, RAG, Multi-Agent Systems, Google ADK, LLM-as-a-Judge, MCP, Hugging Face
Serving & Infra: SGLang, vLLM, AWS, PostgreSQL, Redis, Kafka, Docker
Observability: Grafana, Datadog