

SAHIL GUPTA

LEAD OF COGNITIVE INTELLIGENCE, NEURAL ARC INC

gsahil.sg20@gmail.com Pune, India sahilempire.com github.com/sahilempire linkedin.com/in/sahilempire
huggingface.co/sahilempire

SUMMARY

AI engineer with two years of shipping LLM products end to end: orchestration, agent workflows, RAG, fine-tuning, billing, and the infrastructure that keeps them fast, cheap and safe. Creator of cachellm (npm, PyPI) and the RedSec security models.

EXPERIENCE

NEURAL ARC INC Pune, on-site

Lead of Cognitive Intelligence

Jul 2025 – Present

Designing generative AI systems and multi-agent architectures for enterprise customers and owning them from prototype to production: LLM orchestration, tool calling, memory, RAG, Dockerized backends on AWS and GCP, and usage-based billing with Stripe.

Quantum Coder

Mar 2025 – Jul 2025

Core LLM features and agent workflows for early-stage AI products, with the backend APIs and cloud deployments behind them.

AMPERSAND Pune, on-site

SDE I

Mar 2025 – Present

AI-powered web and app features across the stack with React, AWS and modern backend frameworks. Promoted from intern on the strength of delivery and ownership.

AI and App Development Intern

Dec 2024 – Feb 2025

Frontend components, backend APIs and cloud services for AI-integrated application features.

LVLUP VENTURES Remote

Venture Scout

Oct 2025 – Present

Sourcing and evaluating early-stage startups for the fund.

INDIARESULTS.COM Jaipur, on-site

Data Engineer Intern

Jun 2023 – Aug 2023

Data processing and ETL pipelines on production datasets.

FASTLEGAL Jaipur, remote

Data Engineer Trainee

Jan 2023 – Feb 2023

Data integration and ETL work.

PROJECTS

cachellm OPEN SOURCE

Wraps your Anthropic, OpenAI or Gemini client and injects prompt-cache breakpoints automatically. On npm and PyPI.

RedSec RESEARCH

Security-specialized LLMs fine-tuned with QLoRA for authorized red-team work. Models and datasets on Hugging Face.

MCP Apps Sandbox WRITING

An evidence-based walkthrough of where an MCP App's HTML runs inside Claude.ai and what it can reach.

SKILLS

LLM Systems. Orchestration, tool calling, memory, and RAG pipelines that turn a model into a product, with cost and latency budgets built in from day one.

AI Security. Red-team testing, prompt-injection evaluation, and security-specialized models built for authorized penetration testing and defensive research.

Data Engineering. ETL pipelines, dataset curation, and evaluation harnesses that keep training data clean and models honest.

Agent Workflows. Multi-agent architectures for enterprise use cases, owned end to end from research and prototyping to production deployment and monitoring.

Full-stack Apps. Next.js, TypeScript, and Python applications with Supabase, Stripe billing, and Dockerized backends on AWS or GCP.

Large language models, LLM orchestration and agent workflows, Retrieval-augmented generation, Prompt caching, AI security and red teaming, Fine-tuning with QLoRA, Next.js and TypeScript, Python, Data engineering.

EDUCATION

Jaypee University of Engineering and Technology

B.Tech, Computer Science, 2020 – 2024

Oracle Agentic AI Certified Foundations Associate

Goldman Sachs Software Engineering Virtual Program, 2023

