

Gowtham Sai G

+91 9391758678 | gowthamyadav023@gmail.com | gowthamsai.in | linkedin.com/in/gowtham-sai-yadav |
github.com/gowtham-sai-yadav

EXPERIENCE

AI Engineer (Internship)

Jan. 2026 – May 2026

Scaler AI Labs · clients: OpenAI, xAI, Meta | Python, PyTorch, TRL, FastAPI, MCP, ClickHouse

Bengaluru, India

- Delivered **5 LLM agent environments** (Shopify Admin, HelloSign, Asana MCP, Hootsuite, MS Excel) against hard client deadlines, **driving the requirement calls myself** to scope vague briefs. **Asana MCP validated at a perfect 1.0 on 52 tasks**; HelloSign cleared QA at **92.94%**.
- Built **Pulse**, a production LLM pipeline (desktop agent → S3 → Celery/Redis → LLM → **ClickHouse**) for a **20,000-user** client org; on request I re-cut daily roll-ups to hourly and shipped **per-org prompts**, tunable without code changes.
- Built a distributed **MCP server** at **sub-200ms p99**, adopted by OpenAI and xAI teams, plus a **crawler** (Playwright, Redis) at **100K+ pages/day** and **10K+ graded task suites** at **98%+ accuracy**, cutting review cycles **50%**.

Software Development Engineer 1 (promoted from SDE Intern, Mar. 2025)

Aug. 2024 – Aug. 2025

Wealthy (Fintech) | React, Next.js, GraphQL, REST, Node.js, PostgreSQL, Redis

Bengaluru, India

- Owned the internal CRM used daily by **7,700+ partners** behind **INR 4,000+ Cr AUM** and **INR 190+ Cr** monthly gross sales; turned PM and ops-user workflow complaints into shipped changes, **lifting conversion 37%**.
- Traced sluggish dashboards to data refetched on every tab switch; profiled first, chose **stale-while-revalidate** over websockets, and cut **API latency 32%** under peak load. Built the CRM's first version as an intern; my component and API patterns became the **full-time team's standards**.

OPEN SOURCE: DEBUGGING THE LLM STACK IN PUBLIC

HuggingFace TRL #6534 (merged): found a **silent training-corruption bug** where GRPO let a reward function's length-1 return broadcast across every completion. Filed [issue #6533](#) with repro and root cause, shipped the guard and a regression test, and **replicated the fix in RLOO**.

TRL #6503 (merged): defaulted GRPO to the importance-sampling-corrected KL gradient; a maintainer pushed back, so I reproduced their script, found my own “no-op” claim wrong (**~21% on-policy**), and said so. **TRL #6477 (merged)**: completed the `GRPOConfig.loss_type` docs. **EleutherAI lm-evaluation-harness #3950 (open)**: repointed the **67-subject TMMLU+ benchmark** at a parquet-native dataset so it loads on `datasets>=4`.

entangle (21 stars, 16 releases, [GitHub](#), [site](#)): Go CLI that hands a live coding-agent session to another machine, rewriting every pinned path so it resumes. Encrypted three-word transfer, Homebrew tap, VS Code extension, and a **zip-slip** security fix.

PROJECTS

Viveka, Meta x Scaler OpenEnv Hackathon (Top 9, All-India) | PyTorch, GRPO, TRL, Unsloth | [GitHub](#) | [Demo](#) 2026

- **Top 9 of 70,000+ registrations**. An RL environment teaching agents to **pause and confirm before irreversible actions**, graded by a **6-component deterministic reward** with **no LLM-as-judge**.
- Baselines flat-lined at **0.02/1.0** until I traced it to a **multiplicative rubric** and redesigned it as a mixed form; trained Qwen-1.5B via GRPO to **45% of Claude Sonnet's hardest-tier score at 0.05% of the parameters**.

gravel, an LSM-tree storage engine written from scratch | [Go](#) | [GitHub](#)

2026

- Hand-rolled a skip-list memtable, **CRC32 WAL**, SSTables with xxhash bloom filters, and leveled compaction: **~295K sequential writes/sec**, **~45% faster than BadgerDB**. **Fault-injection tests** crash the engine mid-write to prove recovery.

LEADERSHIP

Chairperson, Yugaantar (SST's First Student-Led Fest)

Sep. 2025 – Nov. 2025

Scaler School of Technology · Earlier: first President, Student Council

Bengaluru, India

- Ran an **INR 55L** budget with **21 HODs** and **200+ organizers** to **6,500+ registrations from 800+ colleges** and **2,500+ footfall** (*Times of India*); SST took **9 podium finishes** against IITs, NITs, and BITS. Rebuilt the sponsor pipeline after **every sponsor withdrew**, and replaced the proshow vendor **3 days out**.

TECHNICAL SKILLS

Languages: Python, Go, Java, TypeScript, C++, SQL | **Core CS**: DSA, OOP, OS, DBMS, Distributed Systems
AI & LLM: PyTorch, HuggingFace, TRL (GRPO/DPO/SFT), Unsloth, QLoRA, RL envs, LLM evals, MCP, RAG
Backend: FastAPI, Django, Spring Boot, Node.js, REST, GraphQL, gRPC, Celery | **Frontend**: React, Next.js
Data & Infra: PostgreSQL, ClickHouse, Redis, MongoDB | Docker, AWS, Kubernetes, Linux, CI/CD

EDUCATION

Birla Institute of Technology and Science (BITS Pilani) | B.Sc. Hons. in Computer Science

Aug. 2023 – Present