HIMADRI S CHATTERJEE

Bangalore, India | +91-8017432134 | officialhimadri11@gmail.com
GitHub | LinkedIn | Portfolio

SUMMARY

Founding Software Engineer with 4+ years of experience building and scaling production AI platforms, real-time streaming services, and payment infrastructure on GCP. Deep hands-on understanding of the transformer stack—from custom CUDA kernels and model training to multi-agent orchestration and cloud deployment. Track record of shipping product in fast-moving startup environments, working cross-functionally with ML, product, and design teams.

TECHNICAL SKILLS

Languages: Python, Go, C, C++/CUDA, SQL, JavaScript
AI / ML: PyTorch, CUDA Kernels, Transformer Architectures, Multi-Agent Systems, LLM Tool-Use, Gemini API
Infrastructure & Cloud: GCP (Cloud Run, GKE, IAP, Compute Engine), AWS, Kubernetes, Docker, Terraform
Backend & Data: gRPC, REST APIs, PostgreSQL, MongoDB, Firebase, Stripe Payments
Frontend & Tools: React, Next.js, HTML/CSS, Git

EXPERIENCE

Founding Software Engineer — Convai, Bangalore
Mar 2022 – Present

- Designed and maintained core REST APIs and gRPC streaming services powering real-time conversational AI interactions; achieved sub-200ms end-to-end latency for voice and text channels at scale.
- Owned early-stage infrastructure on Google Cloud Run—service orchestration, autoscaling policies, blue-green rollouts, and performance tuning across 15+ microservices.
- Built the payments platform end-to-end with Stripe: subscription billing, usage-based metering, and secure checkout flows handling thousands of monthly transactions; co-designed the broader payments architecture.
- Architected on-premise deployment strategy for enterprise clients, packaging services as containerized, GPU-enabled environments; successfully deployed to multiple enterprise accounts.
- Co-built the primary frontend application (React), implementing creator workflows, UI components, and third-party integrations used by the entire user base.
- Worked cross-functionally with ML, product, and design teams to ship features on tight timelines while maintaining platform reliability.

Systems Engineer — Tata Consultancy Services, Bangalore
May 2021 – Mar 2022

- Developed and deployed cloud-based infrastructure solutions on Azure for enterprise clients, focusing on compute provisioning, networking, and identity management.
- Built internal automation tools that streamlined operational workflows for the client organization, reducing manual effort in reporting and resource tracking.

Software Developer Intern — BNP Paribas, Mumbai
Jan 2021 – Mar 2021

- Developed full-stack features for an internal invoice management application (Angular, C#, SQL Server), improving data retrieval performance and extending UI functionality.

PROJECTS

Durable AI Agent Workflow Platform — Effectively-Once Execution Engine | GitHub
Python, FastAPI, Celery, Postgres, Redis, OpenTelemetry, Jaeger, Next.js, Docker

- Built a production-style platform that runs AI tasks as durable workflows on FastAPI, Celery, Postgres, and Redis, with Postgres as the source of truth and the queue only scheduling work.
- Hand-built the durable execution engine: commit-before-execute steps, retry with exponential backoff, idempotency keys for effectively-once side effects, and a periodic sweeper that resumes runs stranded by a

worker crash.
- Added human-in-the-loop approval gates with an audit trail, OpenTelemetry distributed tracing to Jaeger, a cost/latency metrics dashboard, and an automated eval harness scoring runs across six quality and cost dimensions.

GPT-2 from Scratch — with LoRA Fine-Tuning & Speculative Decoding | GitHub

Python, PyTorch, BPE Tokenizer, Multi-Head Attention, LoRA, Mixed-Precision Training

- Implemented a GPT-2 (124M) language model from scratch in PyTorch—custom BPE tokenizer, multi-head self-attention, positional embeddings, and full training loop on OpenWebText with mixed-precision (bf16) and gradient accumulation.
- Extended the base model with a LoRA adapter layer for parameter-efficient domain fine-tuning; implemented speculative decoding with a draft model to accelerate inference by ~2x while preserving output distribution.
- Benchmarked training dynamics (loss curves, gradient norms, learning rate schedules) and evaluated perplexity and generation quality against the original OpenAI checkpoint.

Custom CUDA Kernels for Transformer Ops — with INT8 Quantization | GitHub

C++, CUDA, Shared Memory Tiling, Flash Attention, INT8 Quantization, Nsight Compute, PyTorch C++ Extensions

- Wrote custom CUDA kernels for core transformer operations—tiled matrix multiplication with shared memory, fused softmax, and a flash-attention-style fused QKV kernel—benchmarked against PyTorch and cuBLAS baselines.
- Implemented INT8 quantized matrix multiplication kernels with dynamic per-channel scaling, achieving measurable speedups on consumer GPUs while maintaining acceptable accuracy degradation.
- Packaged kernels as a PyTorch C++ extension for drop-in use; profiled memory bandwidth, occupancy, and warp efficiency using Nsight Compute, documenting optimization decisions at each stage.

Work Buddy — Workstream Control Agent (AI Agents Intensive Capstone) | GitHub

Python, Google ADK, Multi-Agent Orchestration, MCP, Gemini / Vertex, FastAPI

- Built Work Buddy, a multi-agent system on Google's Agent Development Kit (ADK) that turns messy project context—meeting notes, Slack, email, Jira, and repo state—into an auditable execution ledger; a coordinator delegates to five specialist sub-agents for ingestion, cross-source reconciliation, and drafting.
- Enforced a human-in-the-loop approval gate before any external write (Jira, Slack, email, GitHub PR), with secret redaction and source-evidence tracking, so the agent never claims work is done, merged, or shipped without tool proof.
- Exposed the full tool set over a Model Context Protocol (MCP) server, added an eval suite covering the safety-sensitive behavior, and shipped a FastAPI dashboard with a Cloud Run deployment path; model-agnostic via Gemini/Vertex with OpenAI, Claude, and Ollama fallbacks.

EDUCATION

Master of Computer Application — Vellore Institute of Technology, Vellore 2019 – 2021

BSc Computer Science — RKMRC, Kolkata 2016 – 2019

CERTIFICATIONS

Google/Kaggle AI Agents Intensive — Capstone Certified (June 2026)