

Tarun NagaSai

Freelance AI Product Engineer — Agentic AI Systems, RAG & End-to-End Product Delivery

Hyderabad, India (Remote) | +91 9908260773 | tarunnagasai@icloud.com | tarun.avipra.com
linkedin.com/in/tarun-naga-sai | github.com/TarunNagaSai | huggingface.co/TarunNagaSai007

SUMMARY

AI product developer who ships the whole product: the agent loop, the retrieval stack, the GPU inference behind it, and the interface in front of it. Four-plus years of production mobile and web development, now focused on agentic AI systems. Recent work includes Atlas, a financial research platform where a streaming ReAct agent plans, calls tools, and cites its sources live in the UI, and a production translation pipeline that processed 100,000+ verses through a self-hosted vLLM deployment with a 0% error rate. Used to being the only engineer on the project — from database schema to app store release.

SKILLS

Agentic AI: ReAct agent loops, LLM tool use / native function calling (Google Gemini), multi-step planning, agent memory and multi-turn context, streaming agent responses (SSE), cooperative cancellation, agent observability and tracing (Langfuse, Logfire), prompt engineering, safety guardrails and human-in-the-loop review (AI ethics)

RAG & Retrieval: hybrid search (dense + BM25, Reciprocal Rank Fusion), cross-encoder reranking, parent-document retrieval, semantic chunking, GraphRAG (entity and community search), embeddings, pgvector, ChromaDB, LangChain, HuggingFace

LLM Engineering: vLLM inference serving (PagedAttention, KV cache), GPU infrastructure (RunPod), fine-tuning with LoRA / QLoRA (Unsloth, TRL, PEFT), custom model evaluation (ROUGE-L, BERTScore, task-specific exact match), quantization (bitsandbytes), HuggingFace Hub

Backend & Data: Python, FastAPI, TypeScript, NestJS, PHP, Laravel, REST API design, Swagger / OpenAPI, Auth & Role Guards (RBAC), PostgreSQL, MySQL, Supabase, Prisma ORM, ETL pipelines

Frontend & Mobile: Next.js, React, Tailwind CSS, Flutter (4+ years, production), Dart, Firebase (full suite), native iOS / Android bridging (Method Channels)

DevOps & Cloud: DevOps practices across Docker, CI/CD (Fastlane), Vercel, Railway, Google Cloud Platform (Firebase, Vertex AI, Gemini APIs), RunPod, Socket.IO

SELECTED AI PROJECTS

Atlas — Agentic Financial Research Platform

2026 — atlas.avipra.com

FastAPI, PostgreSQL / pgvector (Supabase), Google Gemini, Next.js, React, TypeScript, Tailwind CSS

- Built a streaming ReAct agent on Gemini native function calling: a tool-less planning turn streams the agent's plan first and conditions every later step, then the loop emits typed events (plan, thought, tool call, tool result, sources, token usage) over SSE until the model stops calling tools.
- Engineered the retrieval tool end to end: hybrid search fusing pgvector dense scores with BM25 through Reciprocal Rank Fusion, cross-encoder reranking, and parent-document retrieval — child chunks match, full pages get served, and every passage is numbered for citation.
- Added a GraphRAG tool for structural questions plain retrieval misses: per-document knowledge graphs with entity-traversal local search and community-summary global search.
- Handled the production edges: cooperative cancellation that stops generation within one token of a stop request, a force-answer path when the agent hits its hop limit, idempotent content-hash ingestion, a bring-your-own-API-key flow, and per-turn tracing in Logfire and Langfuse.
- Built the full chat frontend, owning the user experience design: a live reasoning timeline (plan, thoughts, tool call/result pairs), multi-turn history replay, per-conversation token accounting, and dual storage modes.

Yahuah Bible Platform — Production AI Translation Pipeline

2026 — yahuahbible.com

Python, vLLM, Gemma 27B (full precision), RunPod (NVIDIA RTX 6000 Blackwell 96GB), Laravel

- Translated 100,000+ Bible verses into French, German, Italian, and Brazilian Portuguese while preserving Hebrew and Greek sacred nouns — 54,409 automated word replacements at a 0% error rate.
- Built the accuracy machinery: a custom ETL and feature engineering pass extracting 31,000+ verses into a high-precision semantic map, Ratcliff/Obershelp sequence matching for positional replacement, a similarity guardrail flagging misalignments, and deterministic noise filtering that removed 7,300+ false positives ahead of human-in-the-loop review and an admin publish gate.
- Ran DAG orchestration on RunPod with a gated GPU mutex — CPU preprocessing in parallel, one pipeline at a time on the vLLM server — plus async egress to the Laravel API so containers terminate the moment they finish.

Gemma Fine-Tune — Domain Assistant with Custom Evals

2026 — HuggingFace Hub

Unsloth, LoRA / PEFT, TRL (SFTTrainer), HuggingFace Transformers and Hub, Google Colab

- Fine-tuned Gemma E2B on ~8,600 authored instruction examples across three task types, including chain-of-thought reasoning; validation loss converged from 0.41 to 0.164 with no overfitting on a single 16GB T4.
- Wrote a task-specific evaluation suite instead of naive exact match — extracted EM plus token F1 for factual fields, ROUGE-L plus BERTScore for long-form output, and reasoning step coverage — then published both the adapter and the merged model to HuggingFace Hub with model cards.

EXPERIENCE

AI Product Engineer — Freelance

Dec 2025 – Present

Independent, Remote

- Design and deliver agentic AI products end to end: rapid prototyping, architecture, retrieval, inference infrastructure, APIs, and UI.
- Built and shipped Atlas and the complete Yahuah Bible platform — the AI translation pipeline detailed above plus its Laravel backend, admin panel, and an offline-capable PWA wrapped in Flutter for iOS, Android, Windows, and macOS.
- Shipped Horeb, an offline-first multilingual Flutter Bible app with an AI chat module, on a strictly layered Bloc / service / repository architecture — serving 10,000+ readers after deep performance optimization for low-end devices.

Full Stack Developer

Sep 2024 – Nov 2025

SterlingWhiz Solutions Private Limited, Hyderabad — Remote, Full-time

- Sole developer on Project Orion — architected the full product pipeline from scratch across three surfaces: a Flutter mobile app, a React web app, and a NestJS backend.
- Designed the data layer on Supabase-hosted PostgreSQL with Prisma ORM, and shipped a fully documented API surface with Swagger / OpenAPI.
- Implemented role-based access control with NestJS Auth Guards and Role Guards, and built enterprise-grade third-party integrations.

Associate Software Developer — Lead Flutter Developer

Dec 2021 – Jul 2024

ContenTerra Software, Hyderabad — On-site, Full-time

- Led Flutter development across multiple parallel client projects using agile methodologies, shipping more than four applications to both the Play Store and the App Store.
- Built HealthPass by TruNord end to end: bridged Flutter to native health SDKs over Method Channels (Apple HealthKit, Google Fit), integrated the Aktivo SDK, and delivered activity scoring, diet plans, and doctor-on-call video features.
- Built BumpedIn, a proximity social app: radius-based people discovery on background location tracking, real-time chat over Socket.IO, and Firebase Realtime DB, Functions, and FCM.

EDUCATION & CERTIFICATIONS

- Bachelor of Technology (B.Tech), Computer Science — IITDM Vizag, 2016 – 2020.
- Claude Code in Action — Anthropic, issued Jan 2026 (Credential ID qdad5mx8qp8n).

PUBLICATIONS & OPEN SOURCE

- Five technical articles on Medium (medium.com/@tarunnagasai007) spanning LLM pipeline engineering, generative UI research, model quantization, and Flutter architecture; open source includes an experimental Rust garbage collector.