

Shubhankar Gupta

AI Engineer | LLM Agents (LangGraph, MCP), RAG, Evaluation, Python, Cloud Run

Delhi, India (UTC+5:30) | Remote (contractor or EOR) or relocation | shubhankar15august@gmail.com

github.com/shubhankar360 | linkedin.com/in/shubhankar-gupta-a73aa21a0 | shubhankar360.github.io | solquara.com

SUMMARY

AI engineer who builds and runs LLM systems end to end: retrieval, tool servers, evaluation and production serving. Sole engineer behind **solquara.com**, a live AI video-ad studio with a generative-media pipeline and an accounts platform covered by **125 end-to-end checks**. Publishes open-source agent, retrieval and evaluation work with measured results: a LangGraph policy guard kept an eager agent at **40/40** compliant conversations (28/40 without it), and hybrid search in Postgres reached **nDCG@10 0.712 on BEIR SciFact**. Self-taught engineer with eight prior years running operations and client service for a family business.

SKILLS

Languages: Python, TypeScript, JavaScript (ES2022), SQL, PHP 8

LLMs and generative AI: Anthropic Claude API, OpenAI API, Gemini API, prompt engineering, tool and function calling, structured JSON output, multimodal generation (image, video, TTS)

RAG and search: Retrieval-augmented generation, embeddings, hybrid search (BM25 + dense vectors), reciprocal rank fusion, reranking, pgvector, HNSW, citation grounding

Evaluation: LLM and agent evaluation, retrieval metrics (Recall@k, MRR, nDCG), pass@k, Cohen's kappa, McNemar test, bootstrap confidence intervals

Agents: LangGraph (state graphs, checkpointing, human-in-the-loop interrupts), tool calling, policy guards, Model Context Protocol (MCP), JSON-RPC

Backend and data: FastAPI, Pydantic, Node.js, REST APIs, PostgreSQL, SQLite, MySQL, webhooks (HMAC), web crawling, data pipelines

Cloud and DevOps: Google Cloud Run, Terraform, Artifact Registry, Secret Manager, Workload Identity Federation, Docker, GitHub Actions CI/CD

Machine learning: PyTorch, Hugging Face Transformers, PEFT (LoRA fine-tuning), scikit-learn, pandas, NumPy, NLP, classification

Testing and delivery: pytest, Vitest, end-to-end testing, Git, performance profiling, WebGL

Currently learning: LangChain, LlamaIndex

EXPERIENCE

Founder and Lead Engineer, Solquara (solquara.com)

Jul 2026 – Present

AI video-ad and automation studio, remote. Sole engineer: architecture, AI pipeline, testing, deployment and incident response.

- Designed the **generative-media pipeline** (keyframe image, then image-to-video, then voiceover). Costed the models first: for a 25-second Short, keyframe-then-animate was **~4x cheaper** than an all-in-one narrated-video workflow.
- Built the **accounts, plans and billing platform** (auth, 4 plan tiers, usage limits, signed billing webhooks with replay protection) and covered it with **125 end-to-end checks**, including IDOR, session and throttling tests, plus **112 checks against production**.
- Rebuilt the public site as **40 routes** in vanilla JavaScript and WebGL. Homepage went from 23 stylesheets to 2, 26 scripts to 1, and **123 kB to 36 kB** of HTML.
- Profiled a scroll-driven UI, found one CSS effect using **~80% of the frame budget**, and took it from **6 fps to 55 fps**.
- Built a Python **B2B prospecting pipeline** (OpenStreetMap extraction, resumable crawler, Companies House verification) that produced **~920 verified prospects**.
- Found a bug every automated test missed (a duplicate animation hid the whole pricing table) and added a build check that fails if any price is not rendered in the HTML.

Independent AI Engineer, Open-source projects

2024 – Present

- Published the agent, voice, fine-tuning, retrieval and evaluation projects below. Each has its own test suite and CI, and runs on a clean checkout with no API key.

Operations and Client Services, Family handicraft and handloom business, Delhi

2016 – 2024

- Ran daily operations end to end (inventory, fulfilment, pricing, after-sales) without a supervisor. Handled **20+ client conversations a day** and used sales and stock data to decide what to buy.

PROJECTS

handoff-graph

Python, LangGraph, FastAPI, Claude

- Card-support agent in **LangGraph**: intent routing, an identity-verification gate, a policy guard before every tool call, and human handoff with interrupt() and resume. Graded on final account state plus an independent audit of the action log.
- Across 40 conversations, an eager planner stayed at **40/40** with the guard, but broke policy in **12 of 40** without the gate and guard. 15 tests; Docker image and API smoke-tested in CI.

agentcheck

Python, Anthropic SDK, LLM evaluation

- Evaluation harness for tool-using agents: **36 tasks** in a deterministic 7-tool environment, graded on final state, a policy check over the action log, and the answer.
- Showed a transcript-only grader agreed with ground truth **worse than chance (Cohen's kappa -0.40)** and passed 16 policy-breaking runs, and that a 96% pass@1 agent falls to **75% pass@8**. 22 tests; CI regenerates the report.

voice-latency

Python, Vosk streaming STT, TTS, voice activity detection

- Benchmarked every stage of a voice-agent turn offline on 48 clips in four English accents. A 300-500 ms silence threshold cut off **every** caller who paused mid-sentence; only 1200 ms caught them all, so endpointing, not the model, set the latency floor.
- Streaming recognition left only ~30 ms of work after the turn ended, but word error rate was **2.1% on the US voice vs ~11% on UK, Australian and Indian English**. 7 tests; CI runs a Linux smoke benchmark.

lora-intents

Python, PyTorch, Transformers, PEFT (LoRA)

- Learning curves on Banking77 (77 intents) from 5 examples per intent to all 10,003: TF-IDF, a frozen MiniLM encoder with logistic regression, a rank-8 **LoRA** adapter and a full fine-tune, in plain PyTorch on CPU.
- The frozen encoder matched or beat LoRA at every data size (**93.0% vs 92.5%** on all data, 80.3% vs 78.0% at 5 per intent) and trained in minutes; a full fine-tune edged ahead only with all the data (93.2%).

pg-hybrid-search

TypeScript, PostgreSQL, pgvector, Docker

- Hybrid search in one SQL query (BM25 + pgvector, fused with RRF), benchmarked on BEIR SciFact: **nDCG@10 0.712** vs 0.643 for vectors alone.
- Found pgvector's default ef_search silently returns **40 rows for LIMIT 100**. Tests run on PGLite and real Postgres in CI, which caught two portability bugs.

aistudio-mcp

Node.js, MCP, Gemini API

- Zero-dependency **MCP server** giving Claude 9 tools: Gemini text and vision, image generation, Veo 3.1 video, TTS and the Files API. Long renders resume instead of timing out.
- **12 end-to-end tests** run the real server against a local mock of Google's API; CI on Node 18, 20 and 22.

rag-document-qa

Python, FastAPI, Claude, Terraform, Cloud Run

- Citation-grounded document Q&A. Measured that a similarity threshold could not separate off-topic questions (0.907 vs 0.905) and replaced it with a coverage guard that can.
- Packaged for **Google Cloud Run with Terraform** (Artifact Registry, Secret Manager, keyless GitHub Actions deploys via Workload Identity Federation). Found and closed a **file-read vulnerability** in its own API while hardening it. 69 tests.

Also on GitHub: job-radar (collects ~12,500 job postings a run from 9 sources with an eligibility verdict for each; 67 tests), hybrid-rag-eval (BM25 beat hybrid retrieval on Recall@3 on a small corpus and ran 40x faster), promptdesk (support agent with RAG and sentiment-aware escalation).

Earlier ML projects: heart disease classification (88.5% test accuracy, recall 0.92), wine quality prediction (13 models; 93% accuracy, 0.95 ROC-AUC), content-based movie recommender (TMDB 5000).

EDUCATION

Bachelor of Arts, Tourism, Indira Gandhi National Open University (IGNOU)

2020

CBSE Class XII, Air Force Bal Bharati School, New Delhi

2014

Engineering is self-taught; the evidence is public at github.com/shubhankar360.

LANGUAGES

English (fluent, professional), Hindi (native)