Sakalya Mitra

AI Engineer | Evals & Agent Infrastructure | Production LLM Systems

sakalyamitra@gmail.com | +91 98308 87243 | linkedin.com/in/sakalya-mitra | github.com/Sakalya100 | sakalyamitra.com

Relocating to London | Requires Skilled Worker sponsorship

Summary

AI engineer who builds the layer that makes non-deterministic models safe to depend on: evals infrastructure, agent orchestration, prompt tooling and the observability around them. I have shipped 8 LLM systems to real customers and know first-hand how they fail in production, which is why evaluation-driven development became the delivery standard on my team rather than a nice-to-have. Comfortable fine-tuning and self-hosting an open-weight model under a hard data constraint, designing durable state that survives partial failure, and treating model output as untrusted by default.

Experience

Sanscritic (formerly ScaleGen, TheAgentic) — **San Francisco, CA (Remote)**
Senior AI Engineer, Technical Lead — *April 2025 – Present*

- **Shipped 8 LLM systems to real customers** and owned them after go-live, from an ambiguous problem statement through to production and support. Live examples include regulatory document review at three US pharmaceutical companies and a scouting platform with a professional football club.
- **Built the evals infrastructure:** datasets from held-out historical cases with strict temporal cutoffs to prevent leakage, LLM-as-judge through a four-stage adversarial critic pipeline including a blind grounded critic, and prompt regression suites wired into the development loop so every change is validated against a fixed set before release. Raised validated accuracy **from a 53% baseline to 71.3%** on one system and **from 75% to 89%** on another.
- **Owned the prompt and model lifecycle:** prompt versioning and GEPA-driven optimisation of the retrieval path scored against a fixed benchmark, plus framework and model selection on fit, output quality, latency and cost across OpenAI, Anthropic and Azure OpenAI.
- **Fine-tuned and self-hosted where the data could not leave:** served Qwen3-30B-A3B on a single H200 in an isolated client environment, adapted with supervised fine-tuning and then GRPO against agreement with known-correct human review outcomes.
- **Designed the agent building blocks and chose between them:** embeddings and hybrid retrieval, persistent memory across long multi-turn sessions, deterministic tool dispatch with structured output contracts, and long-running state that survives container restarts through checkpointing and idempotent resumption, so a partial failure never duplicates or loses work.
- **Wrapped the model in things that are deterministic:** eligibility gating and quantitative guardrails around model output, failure-aware fallbacks when a provider or tool call dies mid-run, and human-in-the-loop at the decisions a business will not delegate to a model.
- **Instrumented everything:** Logfire and OpenTelemetry tracing on every agent step and tool call for per-step latency and cost attribution, then used those traces to remove the real bottlenecks rather than the assumed ones.
- **Built the backend underneath:** FastAPI services, PostgreSQL modelling and query tuning, and queue-backed orchestration running at 10 concurrent workers and 50+ multi-minute jobs per day.
- **Stayed close to the customer:** run requirements directly with non-technical operators, push back when a request will not work, and lead a team of 4–5 engineers through code review.

FutureSmart AI — **Remote**
GenAI Engineer — *February 2024 – April 2025*

- Built and shipped a production LangGraph agent for Glamira Jewellery, live to **10K+ daily users across 80+ countries**, routing between knowledge-base answers, shopping assistance, ticket creation and websocket handoff to live agents.
- Built the shopping assistant's retrieval layer: indexed **2M+ product images** with a multimodal embedding model in Qdrant, supporting image-based and guided choice-based product discovery, and tuned the index and query path to **sub-2-second retrieval** at that scale, multilingual throughout.

Selected Systems

Regulated Document Review Agent — *Self-hosted and fine-tuned, live at three US pharmaceutical companies*

- Automated FDA Pre-IND, NDA and IND validation over submissions of 100–200+ documents, several thousand pages each, replacing a review a human previously took 2–3 months to complete.
- Served an open-weight model rather than a frontier API because the content could not leave the client environment: Qwen3-30B-A3B on a single H200, adapted with SFT for domain structure and output format, then GRPO against agreement with known-correct human review outcomes.
- Built hybrid retrieval over Qdrant, dense search merged with lexical matching by Reciprocal Rank Fusion, after pure embedding search kept missing regulatory terms where a near-synonym will not do.

- **Cut review cycles from 2–3 months to 3–4 days** at **85–90% agreement with expert human reviewers**, measured on a held-out set of previously reviewed submissions, with a human confirming every finding. *Stack:* FastAPI, Qdrant, PostgreSQL, Celery, Redis, Vision LLMs, DeepSeek OCR, Docker

Autonomous Freight Operations Platform *In production at a US freight brokerage, running live shipment operations*

- Built a multi-agent platform running freight appointment scheduling and proof-of-delivery collection end to end across email, automated voice calls and carrier portals, including a **Fortune 100 retailer's scheduling portal** and **OpenDock**, with a coordinator selecting the channel per load.
- **Built a hybrid deterministic plus LLM validation engine** reconciling AI-extracted delivery documents against the source rate confirmation: local normalisation and fuzzy matching first, targeted LLM semantic-equivalence calls only for genuinely ambiguous fields, and explicit business-rule overrides, emitting a **field-level PASS/FAIL report with per-field reasoning** as the auditable record a human signs off before anything reaches the TMS.
- Ran the collection loop as a durable multi-day workflow that survives restarts and horizontal autoscaling: webhook-triggered scheduled request, escalation on carrier non-response, then validation, backed by a PostgreSQL job store with misfire grace and coalescing plus **advisory-lock guards so each scheduled action fires exactly once across N container instances** rather than once per instance.
- Built the dependency-graph engine behind a drag-and-drop workflow builder: topological ready-node resolution, load-time cycle detection, conditional branch gating, parallel execution of independent nodes, variable passing between nodes, and live execution-log streaming over WebSocket.
- Gated every irreversible external action, document submission, outbound email and appointment booking, behind an async human-in-the-loop approval step with timeout and reject-with-feedback, so model output is treated as untrusted before it touches a customer's system of record. *Stack:* FastAPI, PydanticAI, PostgreSQL, SQLAlchemy, Alembic, APScheduler, Playwright, Steel, Logfire / OpenTelemetry, Turvo TMS API, Docker

Agent Platform and Regression Harness *Reusable infrastructure behind the client work*

- Built the shared platform layer so a new domain starts from working patterns: reusable agent, workflow and retrieval blueprints, prompt versioning, and durable queue-backed job orchestration with bounded concurrency and checkpointed state.
- Architected the reasoning layer on a Recursive Language Model, working over large multi-document context directly rather than a costly multi-hop retrieval loop, cutting token cost and latency roughly 10× while reducing citation errors.
- Wired regression benchmarking into the development loop: measured on APEX Agents and Vals.AI Finance, accuracy on complex long-document tasks moved **from 75% to 89%**, with every architectural change validated against a fixed set before release. *Stack:* PydanticAI, FastAPI, PostgreSQL, Redis, Celery, Qdrant, object storage, Docker

Technical Skills

Languages	Python, SQL, TypeScript, Java, C++
Agents & LLMs	LangGraph, PydanticAI, LangChain, multi-agent orchestration, tool calling, structured output contracts, persistent memory, prompt versioning, GEPA, guardrails, SFT and GRPO fine-tuning, multi-provider (OpenAI, Anthropic, Azure OpenAI) and self-hosted open-weight
Evals & Observability	Eval dataset design, LLM-as-judge and adversarial critic pipelines, prompt regression suites, leakage control, baseline benchmarking, PyTest, Logfire / OpenTelemetry tracing, latency and cost attribution
Retrieval & Data	Embeddings, Qdrant, pgvector, hybrid dense plus lexical retrieval, Reciprocal Rank Fusion, PostgreSQL (modelling, query tuning, GIN indexes), MongoDB, Neo4j, SQLAlchemy, Alembic
Backend & Infra	FastAPI, queue-backed async messaging (Celery, Redis), bounded concurrency, checkpointing and idempotent retries, Docker, Kubernetes, GCP, AWS, Azure, GitHub Actions, CI/CD

Education, Research & Open Source

VIT Bhopal University | Integrated M.Tech, Artificial Intelligence (CSE), **Gold Medallist**, CGPA 9.67, 2020–2025. **10 peer-reviewed papers** (full list) plus **Best Research Paper Award**, ICAIA 2024 (Springer). **Open source:** core contributor to **CortexOn** and **AgenticBrowser**, both agent systems.