ROHAN SITANIYA

Senior Applied AI Engineer · Agent Systems & Evaluation
Mumbai, India · Open to relocation
sitaniya.com · github.com/rsitaniya · linkedin.com/in/rohan-sitaniya · rsitaniya97@gmail.com · +91 7407997776

SUMMARY

Six years building AI systems for domains where a wrong answer is expensive: fraud, credit risk, underwriting and GenAI governance at American Express, then multi-agent workflow automation at Jitterbit. Producing an answer is the easy half. Establishing whether it is right, and how anyone would know when that changed, is the engineering.

PROFESSIONAL EXPERIENCE

Senior Applied AI Engineer — Jitterbit — Oct 2024 – Apr 2026

Agentic workflow automation for enterprise and citizen-developer customers · Founding AI team (1 of 3), reporting to the CTO · Remote

- **iPaaS AI Assistant: Owned the multi-agent system behind Jitterbit's natural-language integration builder** through production rollout: a LangGraph Planner retrieves from 400+ Marketplace recipes and hands a Pydantic-validated plan to an Executor with checkpointed, resumable execution.
- **Retrieval & context:** Each agent got its own context and model, and a token-aware memory layer summarizing between turns kept long sessions bounded, **cutting LLM spend ~40%** with answer quality held. Pinecone-backed retrieval **improved context precision 60%+ and cut latency 30%**.
- **APIM AI Assistant:** Shipped Jitterbit's natural-language API lifecycle agent, handling 10K+ daily interactions so users create and publish APIs without filing a ticket with engineering. A provider-agnostic adapter spans OpenAI, Bedrock and open-source models.
- **Operations:** DeepEval response checks, hallucination and bias detection, Langfuse request-level tracing and model-drift monitoring, on CI/CD-driven Kubernetes microservices so the controls shipped with every release.

AI Engineer — American Express — Aug 2020 – Sep 2024

AI Labs · Emerging AI Products · Model Risk Management

- **GenAI Governance: Scoped and led a 5-member team** building a GenAI evaluation framework for production reviews: 20+ quantitative and 30+ qualitative checks covering hallucination, bias, toxicity, PII, faithfulness and auditability, run against representative production data and wired into automated model gating. **Cut gen-ai governance time by ~60%.**
- **Fraud Embeddings: Designed Transformer-based embedding architectures** that outperformed the legacy RNN representations by ~40bps, and **built GPT-based merchant embeddings** that went into production fraud pipeline.
- **AI Underwriting Automation:** The commercial new-accounts team was reading bank statements manually. I scoped the problem with them, made the case for an NLP approach over adding reviewers, and **shipped transaction-categorization models at ~85% accuracy** to score the credit worthiness of the applicant, **taking that review cycle from days to hours.**
- **Risk Intelligence:** Built an LLM-powered pipeline for the credit-risk unit, reading real-time news and external signals into automated early-warning alerts **preventing sudden, unforeseen multi-million dollar delinquencies.**

ML Engineer Intern — Algonomy — May 2019 – Jul 2019

- **Fine-tuned BERT and GPT-2** for personalized query auto-completion over 1.5 million business analytics queries.

Research Intern — National Taiwan University — May 2018 – Jul 2018

- Genomic analysis of Axolotl RNA-Seq data to identify regenerative epigenomic markers, mapped to the human genome.

SELF-DIRECTED PROJECTS — MAY 2026 – PRESENT

Three systems built on one rule: the agent proposing a change never controls the evidence used to accept it.

Assay — RL environment for coding agents

- In RL training, an agent that can reach its grader learns to game it instead of solving the task. An April 2026 audit found the top 3 Terminal-Bench 2.0 submissions reading their own answer keys.
- Packages a repo at a commit as a pinned, fingerprinted RL task, keeps grading tests out of the agent's reach and reads the reward from the test runner, not the model. It caught a run passing all 133 visible tests with 0 of 18 graded ones fixed, and closed two reward forgeries that changed no code.

Ratchet — Agent-assisted integration delivery (MaDI-Bench)

- Integration scores can pass while the data is empty. An agent's mapping lifted schema F1 to 0.91 on 1,931 real records, cleared tests, a regression check and human approval, and filled 0 founding-year values.
- Built a two-gate delivery loop where the implementer can only propose edits and never touches the evaluator's gold, plus a per-field yield metric that needs no labels. It exposed that field and a second one silently empty on 38% of records; the fix filled 1,056, with the other 875 empty at source.

Vendor vs Valor — Build-or-buy decision support

- Research agents produce citations that look right and are not: the link resolves and the page is on topic, but it does not say what the claim says. A build-or-buy decision made on that research inherits every error.
- Built a research engine that runs build and buy analysis in parallel, forces every claim to quote a saved source verbatim, and lets only an independent verifier set its status, so unsupported claims never reach the report. A challenger then argues the strongest case against its own recommendation before three human review gates.

TECHNICAL SKILLS

Agent Systems & Retrieval: LangGraph, OpenAI Agents SDK, multi-agent orchestration, function calling, MCP, RAG, Pinecone, context engineering, structured outputs (Pydantic)

Evaluation & Observability: Evaluation design, LLM-as-judge, DeepEval, Langfuse, RAGAS, guardrails, hallucination / bias / PII detection, model-drift monitoring

ML & NLP: Transformer architectures, representation learning, BERT and GPT-2 fine-tuning, transaction categorization, sequence models

Backend & Deployment: Python, FastAPI, Docker, Kubernetes, CI/CD, Kafka, Spark, PostgreSQL, AWS, GCP

EDUCATION

Integrated Dual Degree (B.Tech + M.Tech), Biotechnology — IIT Kharagpur **2015 – 2020**

Thesis: Non-invasive detection of human cancers from multi-genomic TCGA signals using deep learning.