GAURAV POONIWALA

Email: gtpooniwala@gmail.com Linkedin: gaurav-pooniwala Website: gtpooniwala.github.io Github: gtpooniwala

AI engineer with 10+ years building production agentic systems. I ship LLM-driven workflows that integrate into real products with strong reliability, evaluation, and human oversight, with experience in early-stage startups and high stakes workflows.

PROFESSIONAL EXPERIENCE

NEXCADE, *London, UK* 2025

AI platform for freight forwarders ($2.4M pre-seed): automated quotation workflows with LLM agents

Founding AI Engineer

Technical scope: Multi-agent systems · LiteLLM · Pydantic · Temporal · PostgreSQL · AWS · Docker · Prompt engineering

- Owned and built a multi-agent system to automate quotation process for freight forwarders (from reading RFQ emails to creation of final quote) handling $50k+ in daily quote volume.
- Built a custom agent framework with standardized templates for prompts, examples, schema validation, monitoring and evals, enabling rapid development of multiple specialized sub-agents.
- Designed a Temporal-orchestrated multi-agent workflow for long-running jobs involving up to 20 LLM calls.
- Implemented a LLM self-healing layer with checkpointing, automatic error detection, and iterative retries/fallbacks, improving robustness to messy real-world inputs and preventing cascading failures.
- Added human-in-the-loop approval gates before any external actions (supplier/customer communications), balancing automation with correctness, auditability, and trust in high-stakes workflows.
- Developed a natural-language rules engine allowing users to define preferences/constraints and conditional application logic in plain text, enabling customer-specific behavior without code changes.
- Established internal AI-assisted development guidelines, and led the team-wide migration to Claude Code.

KIRO, *London, UK* 2024

Founders Factory-backed fintech: AI financial coach with RAG + agentic workflows

Tech Lead (Now Technical Advisor)

Technical scope: Multi-agent RAG systems · LangChain · Semantic retrieval · Pinecone · Next.js · Early-stage infrastructure

- Led development of multi-agent RAG chatbot, contributing to securing $200k in pre-seed funding.
- Designed scalable infrastructure and retrieval algorithms, supporting over 1,000 users.
- Implemented custom agentic workflows to improve response accuracy, reduce hallucinations, and enhance overall user experience.

REPHRASE AI, *Bangalore, India* 2021 – 23

Tech startup acquired by and integrated into Adobe Firefly: Personalized GenAI video creation

Principal AI Engineer & Technical Lead

Technical scope: PyTorch · Generative video systems · Multimodal AI · Computer Vision · Neural network training · GCP

- Developed text-to-video technology acquired by Adobe for $50M and integrated into Firefly serving over 30 million users.
- Productionised video generation pipelines for personalised marketing campaigns, including a Cannes award-winning Mondelez campaign with over 100M views.
- Designed and trained custom neural networks to reduce costs by 75% and data requirements by 66%.
- Managed cross-functional teams of 10 employees, delivering AI-generated video projects worth up to $125k.
- Automated data handling and video creation workflows, reducing manual workload by 80% and eliminating engineering involvement in customer deliverables.

SAMSUNG ELECTRONICS R&D, *Seoul, South Korea* 2015 – 21

Research Engineer (Deep Learning)

Technical scope: Deep learning · TensorFlow · PyTorch · Computer Vision · Quantisation · Pruning · Knowledge distillation

- Provided tailored deep learning solutions as an internal AI consultant across 30+ technical departments within Samsung.
- Architected a company-wide neural network optimisation framework, reducing compute cost and latency by up to 80% and data requirements by up to 50% through quantisation, pruning, and knowledge distillation.

SELECTED TECHNICAL PROJECTS

Pushstart (GitHub): Task manager that actively helps users complete to-dos by breaking tasks into step-by-step instructions, scheduling execution sessions, and executing automated workflows with human oversight.

- Capabilities: task decomposition, execution guidance, scheduling, and HITL review before any external action.
- Architecture: LangGraph agents, OpenAI, Anthropic and Google APIs, MCP integrations (Todoist, Google Calendar, Gmail), PostgreSQL, Next.js, Docker.

Personal Agent MVP (GitHub): Multi-agent personal assistant with an orchestrator coordinating tool-specific agents and persistent memory for end-to-end task execution.

- Capabilities: multi-agent orchestration, tool usage, persistent memory, and RAG-based document Q&A.
- Architecture: LangGraph (multi-agent + orchestrator), FastAPI, tool registry, SQLite, Docker.

LBS Club Treasurer AI Agent (GitHub): Treasurer workflow automation that replaces manual back-and-forth by collecting structured funding requests via chat, validating inputs, and auto-submitting approved forms.

- Capabilities: conversational intake, rule-based validation, approval preview, and automated browser execution.
- Architecture: Python, Selenium browser automation, Gradio UI, structured LLM outputs.

ENTREPRENEURSHIP AND LEADERSHIP

ANTLER UK, *London, UK* 2025
Entrepreneur-in-Residence: LLM Reliability, Observability, Evaluation, and Human Oversight

- Analysed how teams detect, assess, and respond to incorrect LLM outputs, to determine best practices and pain points around observability, evaluation, and human review.
- Ran 30+ customer discovery conversations with AI-first teams to validate failure modes and workflows, translating findings into concrete product requirements and prototype iterations.
- Designed evaluation and monitoring patterns for agentic workflows (LLM-as-a-judge, task-specific metrics, statistical models and human feedback), enabling systematic measurement beyond anecdotal testing.
- Built a risk-scoring and routing approach to trigger human review only for high-severity/low-confidence cases, reducing manual review while preserving correctness and auditability.

INDEPENDENT AI CONSULTING, *London, UK* 2024–2025
Fractional AI Engineer / Advisor

- Consulted 15+ startups on productionizing LLM systems (architecture, evaluation, reliability), plus product strategy, early team design, and fundraising support

HACKLBS, *London, UK* 2024
LBS Flagship Hackathon Winner

- Won 1st place (£3,000) for an event planning product; Pitched prototype and business case to VC judges

TECHNICAL SKILLS

- **Agents & LLMs**: Multi-agent systems, tool use, MCP, RAG, observability/monitoring, evaluation/testing, prompt engineering, context engineering, human-in-the-loop
- **Frameworks & APIs**: LangGraph, LangChain, LiteLLM, OpenAI, Anthropic, Gemini
- **Infra & Data**: Python, FastAPI, PostgreSQL, Docker, Temporal, AWS, GCP, Azure

EDUCATION

Indian Institute of Technology - Bombay, *Mumbai* 2011 - 15
Bachelor's with Honours in Electrical Engineering, Minor in Computer Science GPA: 9.92/10

- Institute Rank 1 among 1500+ students; IIT-B Undergraduate Research Award, 2015
- All India Rank 36 in IIT Joint Entrance Exam (among 500k students)
- Inlaks Scholarship (3 awards among 10k applicants)

London Business School, *London*, and **Wharton School**, *Pennsylvania* 2023 - 25
MBA, Technology and AI strategy

- Dean's List recipient (awarded to top 10% students); GRE 339/340 (99th Percentile)
- President of Data and AI Club; BK Birla Scholar