

Mayank Saini

Member of Technical Staff, AI Engineer at Mirian AI
Agentic Systems • Financial Data Platforms • LLM Routing

@itsmayank.saini21@gmail.com

itsMay21

Mayank Saini

+91 8955071955

WORK EXPERIENCE

Member of Technical Staff, AI Engineer

Mirian AI, New York

Aug'26 - Present

FinClose: Agent-Enabled Financial Close Platform

- **Own FinClose end to end**, Mirian AI's financial close platform and core product, from finance domain modelling through production operation.
- **Took it live with a client**, cutting their close from **three weeks to six days**, run on live ERP and bank feeds.
- **Architected** close cycles, checklists, cross-task dependencies, and **preparer/reviewer segregation of duties** over a tenant-isolated schema with immutable audit trails.
- **Built the agentic layer**: agents prepare, reconcile and explain close tasks, each action **reviewable and reversible**.
- **Hardened** it for clients: role-based access, scoped auth, **tracing and audit logging** across APIs and agents.

Mirian Operating Data Product

- **Designed and built independently** the canonical data layer for financial operations that every Mirian agent consumes.
- **Built ERP-agnostic ingestion**, from **NetSuite** to **QuickBooks**, plus **bank SFTP feeds**, payroll and client portals, into **one clean data layer** with source lineage.
- **Turned** data integration from per-client custom work into a **reusable product layer** for every new agent and client.

Machine Learning Engineer, R&D

PwC US AC, Bangalore

Jul'24 - Aug'26

PwC One: Enterprise AI Platform

- **Architected PwC One**, PwC's first enterprise AI platform, onboarding **50+ AI tools** from Advisory, Tax and Assurance into one agentic layer.
- **Designed a LangGraph** supervisor-as-tool agent architecture with **MCP**, **A2A** and Redis checkpointing.
- **Built a 3-stage context pipeline** (pruning, compaction, summarization) with **LiteLLM** caching and **Langfuse** tracing.

Agent Engineering Hub (AEH)

- **Authored and led adoption** of AEH, PwC US's standard harness for agent-assisted development across practices.
- **Drove 53% faster** execution on development work and moved teams from sprint cycles toward **weekly releases**.

Tax AWM: Agentic Tax Review

- **Built a registry-based multi-agent framework** to extract and cross-check K-1s, footnotes and capital summaries.
- **Cut manual review time by 84%** with explainable compliance checks and audit trails.

EDUCATION

B.Tech, Electrical Engineering

IIT Kanpur, Minor in Machine Learning

2020 - 2024

R&D PROJECTS AT PWC

Spreadsheet Sense: LLM for Spreadsheets

PwC US R&D

- **Built** a Python library for LLM Q&A and sanity checks on multi-sheet Excel, using a sheet/table/row data model, **RAG** and **agent tools** for validation.
- **Reduced** evidence-sheet validation time by **85%**.

NexusAI: Multimodal Query Routing

PwC US R&D

- **Built** a LangGraph multi-agent framework that routes text, document, image and audio queries to the **cheapest capable model**, across LLMs and SLMs.
- **Routed 81-86%** of queries to cost-efficient models, cutting high-cost LLM use by **62%** and tokens by **48%**.
- **Resolved** multimodal workflows **3x faster** through parallel subtask execution and query pre-filtering.

RouteLLM: Cost-Aware LLM Routing

PwC US R&D

- **Built an embedding-based query classifier** with calibrated thresholds, routing hard prompts to GPT-4 and simpler ones to Mistral and Llama.
- **Cut LLM spend by 67%** at **86%** of GPT-4's answer quality.

RESEARCH & PATENTS

- **Spreadsheet Sense: LLM-Driven Spreadsheet Analysis and Q&A**, *Filed patent application, PwC US*
- **Towards Resource-Efficient Multimodal Intelligence: Learned Routing among Specialized Expert Models**, *arXiv 2025*
- **NexusAI: One Supervisor, Many Modalities: Adaptive Tool Orchestration for Autonomous Queries**, *arXiv 2026; filed patent application, PwC US*

EARLIER EXPERIENCE

AI-ML Intern

Samsung R&D Institute, Delhi

May'23 - Jul'23

- **Built** a Smart TV content recommender (GPT, BERT, SVD) at **87% accuracy**, nominated best project of 75 interns.

TECHNICAL SKILLS

- **Languages**: Python, C, C++
- **Agents**: LangGraph, LangChain, Deep Agents, CrewAI, Claude Agent SDK, MCP, A2A
- **Agent Runtime**: sandboxed code execution, ephemeral agent runtimes, FastAPI, FastMCP
- **LLMs**: OpenAI, Claude, Hugging Face, LiteLLM, RAG
- **Observability**: Langfuse, Datadog
- **ML**: TensorFlow, BERT, FinBERT, scikit-image
- **Cloud & Tools**: Azure, AWS, Redis, Git, Claude Code