

Sanjay Budhala

sanjaybudhala.com sanjaybudhala@gmail.com linkedin.com/in/sanjaysingh23

About me

Lead ML Engineer with 4 years building distributed, high-throughput production systems at enterprise scale: a multi-LLM router handling 1M+ requests/day, agentic AI orchestration (DAG workflows, multi-agent), an LLM-as-judge framework that caught a live agent-quality regression, and contributions to an AutoML platform.

🏆 *Sprinklr Manager's Spotlight Award (Q3 FY24) — reliability & robustness of core ML services*

Skills

AI Agents & Orchestration: DeepAgents, LangChain, LangGraph, Agentic workflows, Tool use/function calling, Multi-agent systems, Prompt engineering & optimization, Multi-LLM routing, RAG, LLM-as-judge evaluation

Infrastructure & Tools: Azure, GCP, Docker, Kubernetes, Kafka, Redis, Elasticsearch, MongoDB, gRPC, MLOps

Languages: Python (Primary), Java

Education

Delhi Technological University

Aug 2018 – May 2022

Bachelor of Technology (B.Tech) in Computer Science

- **CGPA:** 9.20/10
- **Coursework:** Machine Learning, Software Engineering, Operating Systems, DBMS, OOPS

Experience

Lead Machine Learning Engineer: AI/ML

Sprinklr

Gurugram, Haryana

July 2022 – Present

Architecting deepagents agentic workflow copilots that automate dashboard analytics through natural language—handling intelligent summarization, query routing, and conversational Q&A.

- Unified Multi-LLM Request Router:

- Solo-designed and built a unified microservice abstracting multiple LLM providers (Azure, OpenAI, Gemini, in-house) behind a single API for both HTTP and Kafka workloads—enabling rapid product innovation across all Sprinklr AI products.
- Drove over \$100,000 in annual cost savings by conceptualizing and implementing granular telemetry and automated mitigation for LLM usage.
- Ensured high reliability and scalability by architecting a fault-tolerant, production-grade service—handling 1M+ daily requests with 99.9% uptime.

- DeepAgents for AI Content Generation Copilot

- Led development of the brand understanding layer, building structured brand profiles to ground downstream agent reasoning and content generation in brand-specific context.
- Built DeepAgents for insight generation, synthesizing social signals, engagement metrics, and external web context into actionable, data-backed content ideas.
- Integrated contextual data sources and information retrieval into agent workflows to enrich insights with relevant brand, audience, and market context.

- Workflow based Agentic Dashboard Copilot

- Built a DAG-based workflow orchestration layer—routing queries through conditional pipelines across 200+ analytics sources with Python REPL tool execution and LLM function calling for code generation and intelligent report selection.
- Built an LLM-as-judge evaluation framework to systematically score agent response quality and track prompt-optimization gains across model upgrades—used it to catch a live regression in agent behavior following an

upstream OpenAI model update (model drift), enabling rapid mitigation before customer impact.

- Implemented sliding-window context management to maintain multi-turn conversation coherence within LLM token limits, with session-level state persistence across user interactions.
- Owned end-to-end intelligent summarization system within a 5-person cross-functional team—automatically distilling actionable insights from 400+ user-created dashboard widgets, reducing time-to-insight by 70%+ based on Limited Availability customer feedback.
- Architected custom field integration layer enabling seamless handling of user-defined metrics with non-standard naming across enterprise customer deployments.

- Platform Infrastructure & Mentoring:

- Contributed to an internal AutoML platform as part of a 4–5 engineer team, automating training, tuning, validation, deployment, and monitoring—100+ daily training requests, cutting model development time 50%+.
- Designed a Python-based distributed rate limiting SDK (Redis + Lua) for the ML org—enforcing fine-grained quotas across time buckets with zero-downtime config refresh via MongoDB, protecting LLM APIs against traffic spikes. [\[blog\]](#) [🔗](#)
- Single-handedly architected and rolled out a secure microservice proxy (HTTP + gRPC) to centralize MongoDB access org-wide—eliminating 1,000x direct connections and mitigating data exposure risks.
- Drove \$10,000/month in savings via Nvidia GPU time slicing; built a GenAI-powered Python SDK for automated unit test generation to raise code coverage across engineering teams.
- Mentored engineers across tenure—from a junior engineer and rotating interns early in the role to 2–3 concurrent junior engineers in a senior capacity.

Machine Learning Intern — Sprinklr, Gurugram, Haryana

May 2021 – July 2021

- Developed a centralized platform to efficiently manage Sprinklr’s ML models and datasets at a unified place.
- Fine-tuned the RoBERTa base model on Hinglish and Hindi datasets, achieving strong accuracy on multi-lingual classification tasks.

Open Source Projects

Mercari AI Shopper — *Multi-Agent AI Shopping Assistant* github.com/sanjay23singh/mercari-ai-shopper [🔗](#)

- Built an end-to-end multi-agent AI system (5 sequential agents) for intelligent product discovery on Mercari Japan—decomposing natural language queries into search strategies, scraping dynamic content via Playwright, and filtering candidates.
- Built a retrieval-augmented generation (RAG) pipeline over scraped product listings to ground LLM-generated recommendations in real listing data, delivering ranked results with data-driven reasoning.
- Implemented selective tool calling: used OpenAI GPT-4.1 native function calling only for structured URL parameter extraction (where validation matters), and JSON output parsing for other agents—balancing capability, cost, and reliability.
- Built fully async, concurrent pipeline with configurable batch semaphores for resource-safe parallel scraping; supports English and Japanese queries with real-time Streamlit UI progress tracking.
- 90+ unit tests covering all major components, demonstrating production-grade engineering discipline.

Publications

Evolution of Modern Deep Image Inpainting Methods

Feb 2022

<https://ieeexplore.ieee.org/document/9754090> [🔗](#)