

Rishabh Jain

github.com/rishbjain1 | [LinkedIn](#) | rishbjain@gmail.com | +1 408-772-5230 (WhatsApp) | +91 97874-44595

MCP / Applied AI Developer | Chennai, India — Open to relocation anywhere

Summary

Applied-AI engineer who builds Model Context Protocol platforms and the evaluation harnesses that keep agents honest. Shipped three public AI systems solo: a 14-tool MCP server running a full generative film pipeline, a hybrid-retrieval RAG service whose CI fails builds on unsupported claims, and a deployed multi-model LLM gateway. At Teleparty, I built a 13-agent LLM pipeline that ships production dashboards end to end — I take a problem, build across the stack, measure it, and publish the evidence.

Technical Skills

MCP & Agentic Systems

- **Model Context Protocol** servers & tools, **agentic workflows**, tool-calling, human-in-the-loop gates, developer platform & DX
- **Anthropic** & **OpenAI** APIs, Claude Code / Agent SDK, multi-agent orchestration, guardrails, traceability, streaming (SSE/HTTP)

Evaluation & RAG

- **Evaluation harnesses**, LLM-as-judge, citation verification, CI quality gates, benchmark & experimental design, cost & latency tracing
- **RAG**: embeddings, vector search (Chroma), BM25, reciprocal-rank fusion, cross-encoder reranking, PyTorch

Languages & Backend

- **Python**, **SQL**, **JavaScript/TypeScript**, **Next.js/React**, **Bash**; FastAPI, REST APIs, auth

Machine Learning

- **PyTorch**, **TensorFlow/Keras**, Hugging Face; classification, clustering, time series, computer vision

Infra & Delivery

- **Docker**, **GitHub Actions** CI/CD, Terraform (AWS ECS Fargate), Cloudflare Workers, Git/PR workflow

Data

- **Athena/Presto** (~50B-events/month tables), **Superset**, **PostHog**, Linear, GA4, BigQuery, pandas, partition tuning

AI Systems & Open Source

studio-mcp — MCP platform for AI film production

2026

github.com/rishbjain1/studio-mcp | [live demo](#)

- Built a 14-tool Model Context Protocol server that runs a generative film pipeline end to end — shot planning, look-locking, image generation, QC, animation, and edit — over stdio and streamable-HTTP, callable from any MCP client.
- Designed the agentic workflow with a style-drift QC gate and human approval between generation stages, plus per-LLM-call cost and latency tracing, so every run is inspectable and every failure traceable.
- Shipped a Next.js/TypeScript console with an LLM **Director** that drives the tools autonomously, a multi-agent orchestration graph (Agent SDK), and a Langfuse-traced benchmark suite — deployed live.

creative-rag — hybrid RAG with a CI evaluation gate

2026

github.com/rishbjain1/creative-rag

- Built hybrid retrieval over a 470-chunk film-craft corpus: dense (Chroma) + BM25 fused by reciprocal rank, reranked by a local bge cross-encoder (PyTorch), served via FastAPI with key auth and a CLI; grounds the MCP platform's prompts over a verified cross-service HTTP integration.
- Engineered a citation-verification evaluation harness wired into CI — the build fails when a generated claim is not supported by the retrieved sources; containerized with Docker + Terraform for AWS ECS Fargate.

fusion — multi-model LLM gateway

2026

github.com/rishbjain1/fusion — *deployed on Cloudflare Workers*

- Built and deployed an OpenAI-compatible gateway that fans one prompt across a frontier-model panel and synthesizes one answer — a drop-in base-URL swap for any OpenAI-client application.
- Built an LLM-judge benchmark harness: the frontier-panel fusion matched the best solo model's quality at ~28% lower cost per task (\$0.14 vs \$0.19), while an honest run on a cheaper-model panel underperformed by 2 points — published both the win and the failure.

- Run a 200-page Obsidian LLM-wiki (Karpathy pattern) operated by Claude Code — schema-enforced pages, automated ingest/lint/consolidation routines, and a ~500-word hot cache answering ~80% of questions in one read.
- Implemented graphify, a knowledge-graph tool that turns code, docs, and media into a queryable graph — every edge marked source-stated vs inferred, so the audit trail stays honest — with community detection.
- Front the system with Hermes, a Telegram agent that routes requests across three tiers: a chat agent, Claude Code sessions, and scheduled background workers.

Also shipped: [finance-os](#) — agentic portfolio system: canonical state + append-only ledger, an LLM advisor that flags rule violations pre-trade, and a scheduled agent fleet. [insider-signals](#) — 7 agents aggregate public market signals (SEC executive-trade filings, on-chain flows) and vote on a consensus call flagging equities to research (GitHub Actions, Supabase). [sf-flow-demo](#) — an LLM triage agent driving Salesforce Service Cloud over REST.

Experience

Teleparty

May 2026 — Present

Remote

Data Analyst (Analytics Engineering)

- Rebuilt the trial-to-paid conversion metric growth leadership relies on — caught a 2.5× overstatement (86% published, 30–40% real) traced to an undercounting event, and aligned 10 SQL pipelines to the fix.
- Re-mapped marketing attribution across the full subscription-revenue base — lifted attributed coverage from 24% to 33% and surfaced a \$67k-per-quarter unmapped-country bug that skewed CAC and channel-spend decisions.
- Built a 13-agent, human-gated LLM pipeline (plan, build, validate, review, publish) wired to the team’s Linear via MCP — a tagged issue invokes it to scope, build, and ship the deliverable into the BI stack directly.
- Cut a five-minute query to 91 seconds via partition pruning on a ~50B-events/month Athena table; corrected a ~50% metric error across a 14-chart dashboard, fixing it once in a shared calculated field instead of chart-by-chart.

90 Pictures Productions

Jul 2023 — Sep 2023

Chennai, India

Assistant Director Intern, Film Production

- Generated 50+ gen-AI storyboard concepts the director used to ideate shots in pre-production; drove the daily shooting schedule for a five-person AD team.

Shuchi Beverages Pvt Ltd

Sep 2022 — Jul 2023

Chennai, India

Financial Operations Analyst (Family Business)

- Organized and analyzed financial and legal records spanning ~\$22M in claims across 11 lender banks, tracked receivables, and prepared documentation for tribunal proceedings; kept bankers, counsel, and auditors aligned through the process under high ambiguity.

Ernst & Young

Jun 2022 — Aug 2022

San Francisco, USA

Technology Risk Analyst Intern

- Ran SOX IT general controls (access/identity, change management, backups) for enterprise clients — the entitlements, secrets, and governance discipline that keeps integrations safe.

Shreadyplast — LaunchX Entrepreneurship Program

Sep 2019 — Sep 2020

Chennai, India

Founding Team Member, Sustainability Innovation

- Developed a plastic-shredding system with a digital-wallet recycling incentive under MIT faculty mentorship; cut plastic waste ~25% and enabled resale of shreds to manufacturers.

Education

AI/ML Certificate Program

Nov 2024 — Oct 2025

Python, Keras, TensorFlow; ML, deep learning

California Institute of Technology

B.S. in Commerce, Management Information Systems

2019 — 2024

Santa Clara University, Santa Clara, CA

GPA: 3.65/4.0 | Minors: Business Analytics, Real Estate, Theatre

Interests

Classically trained actor: Romeo (*Romeo & Juliet* — Prague, Venice), and Edgar (*King Lear*) + Orsino (*Twelfth Night*) on European tour — Amsterdam, Vienna, Genoa, Bremen · AI filmmaking · building personal tools — a web dashboard command center.