

ABHISHEK JAISWAL

AI Engineer · Production Agents & LLM Evaluation · Generative AI

Plano, TX (open to relocation / remote) | +1 (551) 285-8560 | abhishek.jaiswal@rutgers.edu | github.com/abhishek2395 | linkedin.com/in/iamabhishekjaiswal

SUMMARY

AI engineer with 5+ years building production AI and data systems, focused on LLM agents and the evaluation that keeps them reliable. I shipped a production multi-agent AI system end to end inside a regulated financial enterprise: planning, tool use, retrieval, and multi-step orchestration, reaching 98% automation on a process that had been fully manual. I also built and open-sourced an LLM evaluation framework that benchmarks 300+ models with an LLM-as-judge across six quality dimensions. I work in Python, ship to production rather than living in notebooks, and care about what happens when a model is wrong. My background spans agent design, RAG, prompt engineering, evaluation, and the data and cloud plumbing underneath.

AI PROJECTS & OPEN SOURCE

LLM Evaluation & Benchmarking Framework

Open Source · github.com/abhishek2395/llm-eval

Python · Claude (LLM-as-judge) · OpenRouter (300+ models) · Streamlit · live eval dashboard

Open-source framework that benchmarks 300+ language models with an LLM-as-judge. Each response is scored on six quality dimensions, rolled into a composite score and a value index that weighs quality against cost, with a live dashboard for quality, latency, and per-prompt drill-down into the judge's rationale. The evaluation discipline that keeps agents accurate, safe, and cost-efficient in production.

Production Multi-Agent AI System

Production · USAA (financial services)

Claude Sonnet · Snowflake MCP · Python · agent orchestration · retrieval · tool calling

Designed and shipped a 4-agent system end to end: document extraction, context-aware planning, tool use and query execution via MCP, and automated output generation. I wrote the orchestration, defined acceptance thresholds, instrumented quality metrics, and ran iterative evals until it was reliable enough to replace a manual process. 98% automation across regulated, compliance-bound cycles.

EXPERIENCE

Senior Analyst, Product & AI Engineering

Jun 2021 – Present

Tata Consultancy Services | Client: USAA (financial services) | Plano, TX

- Built and shipped a production multi-agent AI system end to end: agent design, planning, retrieval, tool use via MCP, and multi-step orchestration, reaching 98% automation on a previously manual process.
- Created the evaluation layer that made the agents trustworthy: defined acceptance thresholds, built evaluation datasets, and used LLM-as-judge scoring to measure accuracy, relevance, and hallucination across model versions and prompt variants.
- Shipped to production in fast Agile cycles and partnered with business stakeholders to turn real workflows into deployable agent capabilities with clear success metrics.
- Implemented governance and responsible-AI safeguards for a regulated environment: data handling, access controls, audit logging, and human-in-the-loop review on sensitive steps.
- Led a 4-person data engineering team in Agile delivery and built the Python and SQL pipelines (Snowflake) the agents and analytics depend on.

Data Analyst, Ad Intelligence Product

May 2020 – Dec 2020

Airfind Corp | Remote, US

- Drove 15% revenue growth by optimizing ranking signals on an early-stage product. Defined KPIs and ran A/B tests to compare variants and instrument measurement.

Research Analyst, Automation & Data Quality

Jun 2017 – May 2019

Media.Net Software Services | Mumbai, India

- Cut processing time 40% by building Python and SQL automation pipelines for data quality and analytics workflows.

SKILLS

Agents & LLMs: Production AI agents · Agent design (planning · tool/function calling · memory) · Multi-step orchestration · RAG & retrieval · Prompt engineering · LLM APIs (Claude / OpenAI) · Snowflake MCP · LangChain (familiar)

Evaluation & Safety: LLM-as-judge evaluation · Eval pipeline design · Accuracy / hallucination / reliability scoring · Acceptance thresholds · Responsible AI · Audit logging · Human-in-the-loop · Production monitoring

Engineering: Python (Pandas · NumPy · scripting) · SQL · REST APIs · Git · Docker · AWS (SageMaker · S3) · Snowflake · dbt · production shipping in Agile

Data & Domain: Data pipelines · ETL/ELT · Experimentation & A/B testing · Financial services & regulated workflows · Data governance

EDUCATION

MS, Information Technology & Analytics | Rutgers University – Newark | GPA: 3.83 | 2020

BE, Electronics & Telecommunications | KJ Somaiya College of Engineering | 2017

CERTIFICATIONS

AWS Certified AI Practitioner (AIF-C01) | Amazon Web Services | Active

AI for Product Management | Pendo · Google Cloud · Mind the Product