

Chirag Singhal

AI Engineer — Prompt Engineering · RAG · Agents · LLM Evals · Automation

Ghaziabad, India · +91 74284-49707 · chirag@oriz.in · github.com/chirag127 · linkedin.com/in/chirag127 · oriz.in

SUMMARY

I design LLM systems that work in production: prompt pipelines with measurable evals, retrieval-augmented generation (dense + BM25 + reranking over vector DBs), and multi-agent orchestration (planner/executor/critic, JSON-schema tool-calling). Shipped a production multi-agent framework with an LLM-as-judge eval harness and a hybrid-RAG service reporting Recall@k / MRR / nDCG. I wire keyless/local LLM providers to eliminate ongoing API costs, and instrument everything with OpenTelemetry so you see exactly what the model does. Claude Certified Associate (Anthropic, 2026). Rate: \$50–120/hr.

AI / LLM SKILLS

Prompt engineering: systematic prompt design, few-shot / chain-of-thought, eval-driven iteration, structured outputs (JSON schema), system-prompt hardening

RAG: dense embeddings, BM25 sparse retrieval, RRF fusion, cross-encoder reranking, vector DBs (LanceDB, pgvector), chunking strategies, retrieval metrics (Recall@k / MRR / nDCG)

Agents & orchestration: planner/executor/critic architecture, JSON-schema function calling, tool registries, provider failover, multi-step reasoning, MCP (Model Context Protocol) server development

LLM evals: LLM-as-judge harness, automated regression, OpenTelemetry tracing, latency + cost profiling

Providers: Gemini, Groq, Mistral, NVIDIA NIM, Cohere, OpenAI-compatible APIs, keyless/local (Ollama, LM Studio)

Stack: Python, FastAPI, Docker, OpenTelemetry, TypeScript, Claude Certified Associate (2026)

AI PROJECTS — PROOF OF WORK

agent-forge · Python · multi-agent · evals

[chirag127/agent-forge](#)

- Multi-agent LLM orchestrator: planner/executor/critic with pluggable JSON-schema tool-calling, provider failover across 6 providers, and an LLM-as-judge eval harness with automated regression. OpenTelemetry spans end-to-end; 67 tests.

rag-lens · Python · hybrid RAG · retrieval evals

[chirag127/rag-lens](#)

- Production hybrid-RAG service: dense embeddings + BM25 → RRF fusion → cross-encoder reranking over LanceDB; FastAPI + Docker. Eval suite surfaces Recall@k, MRR, and nDCG on every run.

Clear-Thought MCP Server · 36★ · TypeScript · MCP

[chirag127/Clear-Thought-MCP-server](#)

- MCP server exposing 11 structured-thinking tools (sequential thinking, mental models, debugging approaches, decision frameworks, scientific method, structured argumentation) to any MCP-compatible AI client.

oz-ai · keyless multi-provider LLM client

[chirag127/oz-ai](#)

- Browser-side LLM client wrapping g4f/gpt4free with multi-provider failover — no API key required. Powers all oriz.in AI features.

Stochastic-Thinking MCP Server · TypeScript · probabilistic reasoning

[github.com/chirag127](#)

- MCP server for probabilistic / stochastic reasoning steps in AI workflows.

EXPERIENCE — AI AT WORK

Software Engineer (AI tooling) — Tata Consultancy Services

Bhubaneswar · Jun 2025 – Present

- Built a hybrid-RAG search and documentation system over the SAP CPQ knowledge base (100+ videos, 40+ decks) — dense embeddings + BM25 + reranking on SAP-approved Hyperspace AI — cutting ticket-resolution and onboarding time.
- Automate testing, CI/CD, and configuration validation workflows in Python; React dashboards for real-time visibility.
- KT search knowledge base (Cloud Quote) ongoing — practical applied-RAG in a large enterprise context.

Software Developer — Full Stack — QRsay.com

Remote · Jul 2023 – May 2025

- Python, Node.js, MongoDB platform; Apache Kafka event-driven order pipeline; Razorpay HMAC-verified web-hooks.

QUALIFICATION & CERTIFICATIONS

B.Tech, Computer Science & Engineering — AKTU

Uttar Pradesh · 2020 – 2024

- College Topper (Rank 1). JEE Advanced 2020 — AIR 11870 (top 1%).

Certifications: Claude Certified Associate — Foundations, Anthropic (2026) · 20+ Udemy Business courses (Claude Code, MCP, Generative AI, ML, EU AI Act) via TCS.