

# Manoj Mali

Principal AI Engineer | Agentic Systems and Production AI

Pune, India | manoj@manojmali.com

manojmali.com | linkedin.com/in/manojpmali | github.com/MinsterMind

Software engineering leader with 13+ years of experience, specializing in production AI, agentic systems, RAG and real-time conversation intelligence. Combines architecture strategy, hands-on delivery and technical leadership with evaluation-driven development and quality, latency and cost optimization.

## Experience

**Staff Software Engineer | Gruve** • Apr 2025–Present

**Principal Software Engineer | Symbl.ai** • 2020–2025

- Scaled a conversation intelligence platform from 60 to 5,000+ concurrent sessions through distributed architecture optimization.
- Reduced infrastructure costs by 50%; established engineering practices that reduced production defects by 20%.
- Led engineers building RAG systems and real-time AI assistants, integrating retrieval, vector search and streaming communication into production products.

**Software Engineer III | Chalo** • 2017–2020

- Migrated monolithic services to microservices processing 1M+ GPS records daily; built a ticket-booking platform serving 40,000+ monthly active users.
- Implemented CI/CD pipelines that reduced manual deployment intervention by 80%.

**Senior Software Developer | Crowdfire** • 2015–2017

- Developed an article recommendation system processing 500,000+ articles daily and an analytics service tracking 100,000+ social profiles.

**System Engineer | Tata Consultancy Services** • 2013–2015

- Worked on enterprise database systems, schema design and server technologies for client projects; received the Covance Best Problem Solver award in 2014.

## AI Engineering and Technical Leadership

**Agents and context:** Multi-agent orchestration, tool/function calling, MCP, memory, prompt engineering, LangGraph, LangChain, LlamaIndex; durable workflows and human approval gates.

**Retrieval:** RAG, ingestion and chunking, embeddings, hybrid search, reranking, grounded generation, ChromaDB and pgvector.

**Evaluation and LLM Ops:** Offline/online evals, golden datasets, regression gates, LLM-as-judge, agent task success, model benchmarking; LangSmith, Langfuse and OpenTelemetry.

**Model adaptation and inference:** PyTorch, Hugging Face, SFT, LoRA/QLoRA, PEFT, vLLM; OpenAI, Anthropic and Gemini APIs; latency and token-cost optimization.

**Reliability and voice:** Guardrails, red-teaming, tool validation, context boundaries, retries and timeouts; streaming ASR, turn detection and barge-in.

**Architecture and leadership:** Technical strategy, roadmaps, ADRs, design reviews, mentoring, cross-functional alignment and business-metric ownership.

**Engineering stack:** Async Python, FastAPI, Java, Node.js, SQL, Kafka, Redis, PostgreSQL; AWS, GCP, Kubernetes, Docker, Terraform, CI/CD and GitOps.

## Education and Recognition

Shivaji University | 2008–2012

TCS .NET coding challenge winner, 2013