

MUSTAFA SHABBIR MASVI

Enterprise AI Architect | Agentic Systems, RAG and GraphRAG

Bengaluru, India | +91 72595 26231 | hello@mustafamasvi.com | linkedin.com/in/mustafamasvi | mustafamasvi.com

SUMMARY

AI architect with 10 years in software engineering and 2+ years leading **Generative AI** delivery. Helped build In Time Tec's AI practice from its start and delivered **~14 AI projects and POCs** for government and commercial clients in the Netherlands, UAE, South Korea, Australia, the USA and India. Specialises in **knowledge graphs (GraphRAG)**, **text-to-SQL** and **multi-agent systems**, and owns the path from requirements and solution design to production, with **evaluation, observability and Responsible AI guardrails** built in. Works alongside global sales on client calls and demos; **~80% of proposals** written have converted into projects.

CORE SKILLS

- **Generative and Agentic AI:** LLMs (GPT-4o, Claude, open-weight Qwen and Gemma), multi-agent orchestration, LangGraph, LangChain, Langflow, LlamaIndex Workflows, MCP, Google ADK, ReAct agents, agent handoffs, tool calling, agent memory, human-in-the-loop, prompt engineering
- **Retrieval and Knowledge:** RAG, GraphRAG, Neo4j knowledge graphs, hybrid search (BM25 + dense), reciprocal rank fusion, Azure AI Search, vector databases (Qdrant, Pinecone), text-to-SQL, semantic layers
- **LLMOps and AI Observability:** RAGAS, DeepEval, Promptfoo, LLM-as-judge, Langfuse, LangSmith, Opik, OpenTelemetry, offline eval harnesses, regression gates, token cost tracking
- **Responsible AI and Governance:** AI governance, guardrails, Azure AI Content Safety, prompt-injection and safety testing, confirmation gates for AI actions, tenant isolation, RBAC, architecture decision records (ADRs), NFR and threat reviews
- **Voice and Document AI:** Deepgram, ElevenLabs, on-premise Arabic/English speech, Azure Document Intelligence, Docling, Korean and Arabic content
- **Cloud and Platforms:** Azure AI Foundry, Azure OpenAI, Azure Logic Apps, AWS, on-premise LLM hosting, FastAPI, microservices, WebSocket/SSE streaming, Celery, RabbitMQ, Redis, PostgreSQL, SQL Server, Docker, CI/CD (Azure DevOps)
- **Engineering:** Python, C#/NET, ASP.NET Web API, TypeScript, NestJS, React, Angular
- **Architecture and Leadership:** solution and reference architecture, requirements discovery, pre-sales demos, RFP responses, effort estimation, stakeholder management, design authority, Agile/Scrum, mentoring and hiring

WORK HISTORY

Technical Architect (AI / GenAI) | In Time Tec

Dec 2022 – Present | Bengaluru, India

- Joined the AI practice at its start, helped build the team, and delivered **~14 AI projects and POCs**, from short proofs of concept to multi-month production programmes across **agents, RAG, knowledge graphs and text-to-SQL**.
- Wrote **technical proposals, solution architectures and effort estimates** for new business, with **~80% converting into projects**, including an **8-month AI platform programme** in the UAE.
- Partnered with the **global sales team** on client calls: explained solutions, ran **live demos** and shaped scope to help win deals.
- Act as **design authority** for GenAI delivery, approving architecture and production releases across **5+ concurrent client programmes**; ran pre-release reviews (security, database, ingestion, graph) that caught release blockers such as a path-traversal flaw.
- Grew the AI team through **mentoring, internal agent-engineering training and an AI-assisted graduate hiring programme**.
- Set **AI-assisted coding standards** for engineering teams and built a **model-routing plugin** that sends each task to the smallest capable model, keeping AI tooling spend in check.

Staff Engineer | Odessa Technologies

Oct 2019 – Dec 2022 | Bengaluru, India

- Improved LeaseWave platform performance by **over 60%** by profiling with JetBrains dotTrace and rewriting slow SQL; **promoted to the Core Platform Team**.
- Built **Roslyn-based static code analysers** that flag common defect patterns at build time.
- Ran daily Scrum and design alignment for **15+ engineers** and stakeholders.

Senior Software Engineer | In Time Tec

Jul 2016 – Oct 2019 | Bengaluru, India

- Built **C#/NET and ASP.NET** modules and **REST Web APIs** for client integrations.

KEY PROJECTS

Customer Knowledge Assistant (GraphRAG) | Netherlands

AI Architect | 2026 – Present

- Raised overall answer quality from **64% to 98%** (faithfulness 65% to 99%, multi-hop reasoning 52% to 98%) by fixing **hybrid retrieval** and adding deterministic counting tools.
- Outperformed a **competing SQL-plus-vector build** on **66%** of real business questions under independent scoring and matched or beat it on **86%**; our codebase was then chosen as the base for the joint production build.

- Designed a **three-store retrieval architecture** (relational, **knowledge graph**, hybrid search) over customer visit reports; led five client review rounds that cut the schema from 59 to 41 entities, lifting its independent review score from **6.4 to 8.5**.
- **Tech Stack:** LangGraph, ReAct agent, Neo4j, Azure AI Search, PostgreSQL, Azure OpenAI, Langfuse, REST/SSE APIs

AI Chief of Staff | Abu Dhabi

AI and Backend Lead | 2026 – Present

- Gathered and refined requirements with a government executive office through repeated working sessions, turning a broad brief into **defined use cases**: a single view of hundreds of projects, daily briefings, proactive escalations and voice Q&A.
- Architected a **sovereign, fully on-premise multi-agent platform** (briefing, escalation and Q&A agents) on **self-hosted open-weight LLMs**, with **Arabic and English voice**, RAG over project documents and an event-driven pipeline for deadline monitoring.
- Built a **human-in-the-loop guardrail** (the executive approves every outbound action), defined data-ownership boundaries, and integrated work across **multiple teams and partners**, including a 12-person AI, backend and mobile team.
- **Tech Stack:** LangGraph, Qwen and Gemma (self-hosted), NestJS microservices, on-premise STT/TTS, RAG, native iOS client

Sustainability Awards Knowledge Platform | UAE

Lead Architect | 2025 – 2026

- Architected a **three-layer AI platform**: an application-intelligence engine, a **multilingual tech-support bot** and an **internal knowledge chatbot**, producing scored reports for the final judging team.
- Support bot served **7 languages and 200+ concurrent users** during peak submission windows, using **hybrid retrieval with RRF**, **Azure AI Content Safety** guardrails and automatic escalation to human agents.
- Cut internal chatbot latency from **5–6 seconds to under 3 seconds** by replacing agentic retrieval with deterministic retrieval, with tracing and evaluation through Langfuse, Opik and RAGAS.
- **Tech Stack:** LlamaIndex Workflows, Azure OpenAI GPT-4o, Qdrant, Redis, FastAPI, Azure AI Content Safety, Azure Logic Apps, Langfuse, Opik, RAGAS

Automated Due Diligence AI Engine | UAE

Lead Architect | 2025 – 2026

- Designed a **three-level evaluation engine** (eligibility screening, efficacy scoring, deep due diligence) that assessed **8,000+ applications** in under 15 minutes each, cutting **manual review effort by over 80%**.
- Added real-time **claim fact-checking**, file-metadata forensics and **vector-based plagiarism detection**, giving **100% automated audit coverage** and catching tampered attachments.
- Standardised **LLM-based scoring** across prize categories to reduce subjective bias in judging.
- **Tech Stack:** LangGraph, Azure OpenAI, Celery, RabbitMQ, Qdrant, Tavily, Azure Document Intelligence, ExifTool, PostgreSQL, Docker

Xero Conversational Accounting Assistant | Australia

Technical Architect | 2025

- Architected a **production AI accounting assistant** that lets small businesses and accountants query and update live Xero books in plain language, through a **custom MCP server exposing 50+ accounting, payroll and reporting tools**.
- Designed **per-tenant ReAct agents** with intent routing (a small model picks the prompt module to cut token use), **mandatory confirmation before any create, update or delete**, and strict **tenant isolation** across multiple organisations.
- Turned model output into **charts, tables and dashboard cards** and one-click PDF business-insight reports, with usage metering; deployed to Dev, Test and Prod on Azure.
- **Tech Stack:** LangGraph, MCP (custom TypeScript server), Azure OpenAI GPT-4o / GPT-4o-mini, FastAPI, WebSocket, React, Azure Document Intelligence, Azure DevOps, Docker

DataInsight AI Platform: Visual Agent and Workflow Builder | Internal Product

Architect | 2025

- Architected the company's **visual, low-code AI workflow builder** (built on Langflow), where teams compose **multi-agent and RAG pipelines** from 100+ components without glue code; a standard RAG flow goes from idea to running in **under 30 minutes**.
- Made every flow deployable as a **REST API or MCP server** across **15+ LLM providers and 10+ vector stores**, with built-in **LLM observability** (Langfuse, LangSmith) and **RBAC**; used as the foundation for client AI builds.
- **Tech Stack:** Langflow, LangGraph, LangChain, MCP, FastAPI, React, PostgreSQL, Redis, Langfuse, LangSmith, Docker, Azure Pipelines

Other Agentic Builds

- **Medical receptionist voice agent**: hands-free booking, rescheduling and reminders with **sub-second voice response**. Tech Stack: LangGraph, Deepgram, ElevenLabs
- **React code generator**: turns stated requirements into a runnable, customised React application. Tech Stack: LangChain, AWS-hosted LLM, FastAPI, Docker

EDUCATION

B.Tech, Information Technology | Dayananda Sagar College of Engineering, Bengaluru | 2016