DEEPAK KUSHWAHA

AI Architect & CTO · Head of Engineering · Agentic Systems, RAG, LLMOps · Built & led a 20-person engineering org (NL/IN)

Bengaluru, India · +91-9901549150 · kdeepakkushwaha@gmail.com · linkedin.com/in/deepak8896484842 · deepakkushwaha.tech

SUMMARY

AI Architect and CTO with 9+ years in software and 4.5 years building AI systems in production. I design agentic and multi-agent systems, knowledge-graph and ontology-grounded RAG, and cost-governed LLM platforms, and I still write code every day. As CTO of Metis Labs (The Branding Club) I own two live SaaS products end to end and lead a ~20-person engineering, QA and DevOps org across the Netherlands and India, covering architecture, delivery and people leadership (hiring, 1:1s, performance and career growth).

AI EXPERTISE

Artificial intelligence & ML systems	**4.5 yrs**	agents, retrieval, evaluation, inference cost
RAG (production)	**3 yrs**	hybrid retrieval, grounding, citations, evals
Taxonomy & ontology design · knowledge graphs	**2 yrs**	ontology-grounded RAG, GraphRAG, semantic modelling
Intent classification (independent of LLM)	**2 yrs**	embedding + small-classifier routing before LLM calls
Small language models (SLMs)	**2 yrs**	open-weight models on own GPUs, per-workload selection
Agentic & multi-agent systems	**2+ yrs**	planning, tool use, guardrails, human-in-the-loop, evals

EXPERIENCE

Metis Labs B.V. (The Branding Club) — Jan 2026 – Present · Remote (NL)

Chief Technology Officer — Jul 2026 – Present

- Own technology strategy, architecture, security and delivery for BrandHub (multi-tenant order management) and Koper (AI spend management), reporting to the CEO and board.
- Built Metis Engine, an agentic SDLC pipeline (ticket → spec → agent build → generated QA → validation agent): 92–95% of agent-built features pass client validation; delivery 4–5× faster.
- Architected the AI knowledge layer: taxonomy and ontology design, knowledge graph linked to vector search, and RAG over live production data with grounded, cited answers.
- Run a cost-governed AI stack: intent routing ahead of LLM calls, open-weight SLMs on our own GPUs, BYOK LiteLLM gateway, LLM evaluation and observability.
- Lead ~20 engineers, QA and DevOps across the Netherlands and India; ERP and payments integrations (Exact Online, Auditvize, Mollie, SEPA Direct Debit).

Head of Engineering — Jan 2026 – Jun 2026

- Led the engineering, QA and DevOps org through the carve-out from The Branding Club; set up hiring, 1:1s, goals and performance reviews for ~20 people.
- Rolled out the Claude Code agent platform (engineering constitution, subagents, slash commands, headless runner) for ticket-to-PR automation.
- Built the DevOps/Cloud career framework and Metis Academy (12-week graduate-to-production-engineer programme); ran a GDPR-constrained data remediation with zero customer impact.

The Branding Club — Head of Engineering & AI Architect — Jul 2024 – Jan 2026 · Hybrid

- Built the engineering, QA and DevOps team from zero to 20 and architected BrandHub (later carved out into Metis Labs B.V.).
- Designed the RAG/LLM content and SKU-generation layer (GPT-4, Claude) on a product taxonomy; cut manual content effort 70%, with LangSmith and Helicone evals for quality, hallucinations and token cost.
- Microservices backend load-tested to 100K read/write ops/sec and 100K SKUs per batch (multiprocessing, queue-based workers, sharding).

Intellectyx, Inc. — Technical Lead Jan 2024 – Jun 2024 · Remote (US)

- Led four cross-functional teams building SaaS products for US enterprise clients; owned requirements, design, delivery and executive communication.
- Designed a vector-embedding recommendation engine that lifted learner engagement 40%; AI-assisted tooling raised productivity 35%, contributing to $2M in client revenue.

Turno — Technical Lead (Founding Team) Jul 2022 – Dec 2023 · Bengaluru

- Led engineering for a Facebook and Google partnership: dynamic deployment of 20 tailored landing pages, conversion 1% → 25%; integrated LLM-assisted personalisation to reduce CAC.
- Owned backend, frontend and DevOps for the Agent App and Early Warning System; Best Performer (Engineering) twice.

Sourcewiz (now WizCommerce) — Senior Software Engineer (Founding Team) Sep 2021 – Jul 2022

- Onboarding automation cut B2B customer onboarding from 10 days to 1; led 3–4 engineers to ship the Customer App to production.

Tally Solutions — Software Development Engineer II, I & Trainee Jul 2019 – Sep 2021

- Code design for TallyPrime 1.x and GST web components; led R&D on WhatsApp and payment-gateway integration. Super Coder, Spot and Rookie awards.

PixelAI.in — Co-founder & Engineering Lead (alongside B.Tech) 2017 – 2019

- Co-founded a software consultancy at university; shipped 10+ products for paying clients.

SKILLS

AI architecture: agentic & multi-agent systems, MCP, knowledge graphs (Neo4j), taxonomy & ontology design, GraphRAG, RAG, intent classification, SLMs, fine-tuning

Retrieval & data: pgvector, Pinecone, Weaviate, hybrid BM25 + dense search, LangChain, LangGraph, LlamaIndex

LLMOps: LangSmith, Helicone, Braintrust, evaluation design, token-cost governance, LiteLLM gateways, GPU serving, open-weight model selection, EU AI Act

AI-native engineering: Claude Code (subagents, slash commands, headless runs), Codex, Cursor, Linear-driven agent workflows

Platform: Python, Node.js, TypeScript, PostgreSQL, Kafka, microservices, AWS, Kubernetes, Docker, CI/CD

Leadership: hiring (35+), 1:1s, performance reviews, career frameworks, roadmap, stakeholder and vendor management

EDUCATION & AWARDS

B.Tech, Computer Science, Visvesvaraya Technological University (VTU), 2019 · Best Performer (Engineering) ×2, Turno · Super Coder, Spot and Rookie awards, Tally Solutions · Google India Challenge Scholarship 2018