Kartik Makkar

San Jose, CA 95134

(408) 324-6029 | kartikmakkar.in@gmail.com | linkedin.com/in/makkarkartik

Summary

Senior AI/ML Engineer & Platform Architect with 14+ years shipping enterprise platforms across collaboration, observability, and healthcare. Architects production AI systems that turn enterprise documents and application data into governed knowledge, chat, search, and agentic analysis experiences. Known for improving accuracy, reliability, and delivery velocity through data-aware routing, deterministic grounding, protocol-based ingestion, hybrid retrieval, and configuration-driven agent delivery. Pairs hands-on model work (RAFT / LoRA fine-tuning, RAG evaluation, LangGraph agents) with production architecture across Python, Java/Spring, Angular, MongoDB Atlas, and cloud platforms. **M.S. Computer Science (AI), Georgia Tech, 2024.**

AI/ML Fine-Tuning Projects

Healthcare-domain SLM — 12B RAFT fine-tune for grounded RAG that knows when to refuse. Fine-tuned a Google Gemma 12B adapter to test whether a self-hosted, domain-tuned generator can hold its own against a frontier API on specialized grounded Q&A at a fraction of inference cost. Built the retrieval, RAFT data-generation, training, and evaluation pipeline end-to-end (Modal GPUs, MLflow), benchmarked head-to-head against `GPT-5.4` and `GPT-5.4-mini`.

- **Retriever-first, retriever-aware training.** Tuned and *froze* a hybrid retriever (dense + BM25 + RRF + cross-encoder rerank; hit@5 ∼0.78) before touching the generator, then built every RAFT example from the retriever's *real* top-5 output — its actual near-misses and genuinely unanswerable cases — so the training distribution matched deployment. Trained as full-precision **bf16 LoRA** on multi-GPU (8×H200) DDP (Modal), val token-accuracy 0.934.
- **Bias-balanced evaluation.** Designed a three-model-family LLM-judge panel (independent majority, answer-order-swapped) to cancel judge self-preference across the teacher, student, and candidate model families, run at *full coverage* (1,572 grounded rows × both frontier tiers) via the Anthropic Batch API. The tuned 12B **tied GPT-5.4-mini head-to-head** and trailed full GPT-5.4 on answering — after showing overlap metrics (ROUGE/BERTScore) reward teacher-mimicry, not correctness.
- **Calibration — the deployable edge.** Built a deterministic answer/refuse confusion matrix over all 1,880 cases, scored by *balanced* accuracy (raw accuracy rewards "always answer"): the 12B led calibration **79.5% vs. 67.4% / 57.3%** and correctly refused **74% of unanswerable questions vs. the full frontier's 16%**, while ranking most faithful to source by RAGAS — the no-hallucination property a high-stakes RAG copilot needs. Designed the serving path: vLLM LoRA serving, eval-gated adapter promotion, FinOps guardrails.

Experience

Cisco Systems **Nov 2018 – Present**

San Francisco Bay Area, CA

Software Architect **Nov 2024 – Present**

AssessmentX AI — Cisco's enterprise AI layer for knowledge ingestion, structured / semantic search, RAG chat, and multi-skill agentic analysis; in production across 40+ hospital systems (∼10 facilities each, 400+ assessment projects). Three decoupled production agents on a shared, domain-agnostic LangGraph SDK — new specialists ship as declarative skill manifests (5 graph templates, three-tier model resolution, Pydantic-validated outputs, bounded retries) rather than hand-rolled plan/act/validate loops.

- **Interactive multi-intent chat agent.** A schema-aware planner that classifies intent per turn and serves three modes from one graph: RAG generation for unstructured questions (hybrid retrieval — vector + BM25 + RRF + BGE / LLM rerank), an NL→MongoDB aggregation fast path for structured-data questions, and an analysis mode that surfaces the available analysis types and lets the user launch the separate, HTTP-invoked analysis service from within the chat. Conversation memory, safe caching, and deterministic grounding; collapsed a 3-call route / classify / plan chain into one schema-aware planner (∼3→1 LLM calls), holding grounded answers sub-2s (∼3–3.5s vector TTFT) while cutting per-turn token cost.
- **Supervisor-orchestrated analysis service.** An independent, HTTP-invoked supervisor-pattern main graph: a supervisor coordinates specialist sub-agents with Send-based parallel fan-out, each a declarative skill on the SDK with structured outputs and bounded retry, and **streams live graph-topology decisions to the client over SSE**. 3 production specialists; custom graphs when templates don't fit. **Saved delivery architects 40–50 hours of manual research per engagement** across analysis types.
- **Governed spec-to-code studio (Assessment Studio).** A separate agent that authors new assessment pipelines (parse → score → render Python) from a markdown spec, gated by an AST allowlist → SHA-256 human approval → frozen-pack snapshot → memory / time-capped subprocess sandbox, so no ungoverned LLM output reaches production. **Collapsed ∼4 weeks of engineering lead time per new assessment type into ∼25 minutes**, and cut generation cost by routing codegen to cheaper models (GPT-4o / GPT-5.x rather than Claude Opus / Sonnet).
- **Debugger Studio — operate, trust & optimize in production.** Self-hosted, HIPAA-aware, LangSmith-grade observability for all three agents: live DAG / span trees, token & cost accounting, LLM-as-judge scoring, and run comparison — used to benchmark agent quality, estimate cost, and drive cost / performance optimization. Keeping traces in-network preserved **data sovereignty** and removed ∼**$20K/yr in departmental cross-charges**; grounding, refusal, and tenant isolation keep every answer cited to retrieved data or deterministic computation.
- **Agentic SDLC framework.** The AI-native toolkit used to deliver the platform — ∼55 multi-IDE skills (Cursor, Claude Code, GitHub Copilot, Windsurf, VSCode AI), an age-encrypted secret vault, a TypeScript MCP server (31 tools / 10 resources / 5 prompts), and CSDL / OWASP security gates.

Software Engineer IV **Sep 2022 – Nov 2024**

Built a healthcare assessment-automation platform that compressed weeks of solution-architect effort into minutes of repeatable ingestion, scoring, recommendations, and deliverable generation for hospitals, providers, and life-sciences customers.

- **Assessment automation platform.** Architected and shipped the platform end-to-end across Java 21 / Spring Boot 3.5 microservices, a Python AI layer, and Angular 18 UI; automated spec-template ingestion, third-party HIMSS scoring intake, score computation, dashboards, and DOCX/PPTX deliverables.
- **Priority-aware recommendations.** Led parallel workstreams for a new assessment type across ingestion, data-collection UI, scoring, dashboards, deliverables, and practitioner workflows; modeled personas, themes, services, use cases, and leadership priorities to align maturity recommendations with business goals.
- **Cross-stack integration & security hardening.** Unified project lifecycle and document workflows across Java services and the Python AI service, including per-project AI flags and Feign-proxied integration. Ported Flask services to FastAPI, moved secrets into AWS Secrets Manager, and drove CSDL / SonarQube remediation across Spring, OAuth, CSP, and OWASP ZAP findings.

Software Engineer IV **Feb 2022 – Sep 2022**

AppDynamics, Customer Engineering. Internal team transfer from CX Collaboration — customer-facing engineering role on observability strategy for enterprise customers.

- **Technical SME, customer engagement lifecycle.** Embedded with enterprise customers as an architecture-level observability SME. Mapped application architecture, defined KPIs with engineering leaders, translated executive observability goals into delivery requirements, and stayed engaged through implementation, including custom AppDynamics plugins.

Software Engineer III/IV **Nov 2018 – Feb 2022**

CX Collaboration, Cross Domain Center of Excellence. Engineering lead on the Collaboration Workflow Manager (CWM) — Cisco CDCoE's platform for onboarding and provisioning users onto Cisco collaboration infrastructure for enterprise customers including a Big Four U.S. bank and a global technology company.

- **CWM architecture lead.** Owned the data-extraction and data-aggregation microservices for an enterprise collaboration provisioning platform. Drove cluster consolidation, data-migration services, and customer rollouts for a Big Four U.S. bank and a global technology company.
- **Webex Contact Center agent-statistics dashboard.** Designed and shipped a full-stack dashboard for live agent, queue, and contact metrics using Angular, Spring Cloud, and MongoDB, with ownership from UX through backend delivery.

Infosys **Aug 2011 – Nov 2018**

Bangalore, India & SF Bay Area

Full-Stack Engineer **Aug 2011 – Nov 2018**

Enterprise provisioning platform for a global financial-services (banking) client.

- Engineered the workflow-automation core of an enterprise provisioning platform for a major global bank, handling user, device, voicemail, and contact-center provisioning. Modeled each request as a directed graph of Java service calls through jBPM workflow nodes and exposed capabilities through Spring REST and SOAP-WS APIs; scaled to 500K+ requests across a 5-year engagement.
- Engagement technical lead and customer-facing liaison for the US engagement with a major global technology company. Owned integration with platform APIs and helped drive the re-architecture from a monolithic Spring / Tomcat application to a microservices platform.

Education

M.S. Computer Science (Artificial Intelligence), *Georgia Institute of Technology* — Atlanta, GA **Dec 2024**

B.Tech. Electronics & Communications, *Punjabi University, Patiala* — Punjab, India **Jan 2011**

Technical Skills

AI / ML — Systems & Retrieval: Agentic SDK development, multi-agent orchestration, LangGraph state machines, Model Context Protocol (MCP), LLM codegen (AST-gated), type-agnostic Document IR ingestion, hybrid search (cosine / dot-product + BM25 / TF-IDF, variable weights), Reciprocal Rank Fusion, LLM / BGE re-ranking, small-to-big hierarchical retrieval, NL-to-MongoDB / NL-to-SQL, GraphRAG, embeddings (Matryoshka, OpenAI, Google)

AI / ML — Modeling, Eval & Safety: Fine-tuning SLMs / LLMs, LoRA / QLoRA, RAFT, synthetic data generation, RAG evaluation, LLM-as-judge, pairwise evaluation, RAGAS, AI guardrails, grounding / hallucination control, prompt-injection defense, tenant isolation, HIPAA-aware AI observability

AI / ML Tools: LangChain, LangSmith, PyTorch, HuggingFace Transformers, TRL, PEFT, Databricks, MLflow, pandas, NumPy, scikit-learn, XGBoost, TensorFlow, AWS Bedrock, AWS SageMaker, GPT-4o / Claude / Gemini / Llama / Azure OpenAI

Backend & Databases: Python (FastAPI / Flask), Java 21, Spring Boot 3 / Spring Cloud, Maven, REST / gRPC / SSE, microservices, distributed systems, event-driven architecture, system design, MongoDB Atlas, PostgreSQL, DuckDB

Frontend: Angular 18, TypeScript, standalone components, signals

Security: HIPAA-aware design, CSDL compliance, OWASP ZAP, Trivy, Semgrep, Anchore, OAuth, JWT, AWS Secrets Manager, age-encrypted secret vault

Development Tools: Cursor, Claude Code, GitHub Copilot, Windsurf, Conda, Docker, Kubernetes, OpenShift, Terraform, CI/CD, Jenkins, AWS (S3 / ECR / Secrets Manager / OpenSearch / EC2 / IAM), GCP, Azure, Jira, Confluence, SharePoint, Splunk