Sarthak Langde

Lisbon, Portugal | sarthak.langde@gmail.com | +351 910 633 592 | Github | Linkedin | Website

Principal AI Architect | LLM Systems | RAG | Inference Optimization | Fine-Tuning

AI systems leader with 10+ years of experience architecting and deploying production LLM systems across retrieval, fine-tuning, and inference optimization. Proven track record of designing high-performance AI infrastructure, reducing inference cost by up to 60%, and delivering complex AI systems under strict latency and reliability constraints. Operates at the intersection of model architecture, evaluation, and infrastructure to deliver reliable AI systems at scale.

Core Strengths

- LLM System Architecture: RAG, hybrid retrieval, fine-tuning strategy, tool-augmented systems
- Inference Optimization: quantization, batching, Triton, ONNX, GPU efficiency
- Evaluation & Alignment: LLM-as-judge, regression testing, quality frameworks
- Multi-Agent Systems: orchestration, memory, workflow automation
- Infrastructure: Kubernetes, multi-cloud (AWS, GCP, Azure), private cloud deployments
- Data Systems: large-scale ingestion, document pipelines, vector search (Qdrant, pgvector)

Experience

Principal AI Architect Aug 2025 - Present

Kerto AI, Lisbon, Portugal

- Led system design for AI-driven marketing automation platform for healthcare client, managing campaigns across a €20k/month budget pilot in Spain and Portugal
- Built structured generation pipeline producing TOF/MOF/BOF campaign strategies across multiple AI-defined personas, enforcing consistency across outputs under budget and targeting constraints
- Integrated with Meta Ads API for automated publishing and reporting, enabling end-to-end campaign execution from strategy to deployment

Head of AI Mar 2024 - August 2025

10xStudio, Amsterdam, Netherlands

- Architected tool-augmented LLM systems combining retrieval, structured query generation and multi-step reasoning under strict latency constraints (<20s end-to-end)
- Developed Warren AI (Investing.com), enabling natural language to FinQL translation (100+ operators) with reliable tool execution and failure handling under production constraints
- Solved schema grounding and query correctness challenges for LLM-generated structured queries
- Implemented hybrid retrieval (BM25 + vector search) with domain-specific preprocessing, improving answer relevance across production systems
- Designed LLM evaluation frameworks with multi-metric scoring (~200+ test cases) to track regression and improve response quality over time for immersion-based roleplay bots

Head of Engineering Dec 2022 - Feb 2024

Stochastic AI, Boston, USA

- Deployed RAG customer support system for Airstream (Thor Industries, Fortune 500), processing 60TB of scanned manuals with strict requirement of <6s response latency
- Led architecture of xChat, a hybrid RAG + fine-tuned LLM system for enterprise knowledge retrieval and support with continuous evaluation pipelines built-in

- Designed retrieval + generation separation, using hybrid search for grounding and fine-tuned models for controlled response behavior
- Solved large-scale ingestion challenges for unstructured PDFs, balancing cost, accuracy, and processing time
- Built continuous learning loop using user feedback and human corrections, resolving ~90% of failing queries per retraining cycle
- Automated fine-tuning pipelines via Kubernetes, enabling regular model updates without manual intervention
- Architected multi-cloud deployment across AWS, GCP, and Azure to meet enterprise data constraints
- Scaled engineering team from 1 to 15 and led end-to-end delivery of production systems

Deep Learning Engineer Sep 2021 - Dec 2022

Stochastic, Boston, USA

- Built inference optimization systems achieving up to 50x speed improvements across large LLM workloads using quantization, dynamic batching, and GPU-level optimizations
- Fine-tuned a 13B LLM in $1000 that outperformed BloombergGPT 50B on their own benchmarks.
- Implemented high-performance inference pipelines using Triton, ONNX, TensorRT, and Kubernetes
- Developed core components of model acceleration platform for production deployment
- Open-sourced xTuring (2.5k stars), an LLM fine-tuning library and x-stable-diffusion (550 stars), a Stable Diffusion inference optimization library.

Earlier Roles Jul 2015 - Aug 2021

- **AI Advisor,** Brainfish, Sydney, Australia
- **CTO,** Tringo, Stockholm, Sweden
- **Machine Learning Intern,** Sinch, Stockholm, Sweden
- **Senior Software Engineer,** Radius, California, USA
- **Software Engineer,** smallcase, Bengaluru, India
- **ML Research Intern,** CSIR Fourth Paradigm Institute, Bengaluru, India
- **Software Engineering Intern,** Ramrao Adik Institute of Technology, Navi Mumbai, India

Education

MSc. Artificial Intelligence Aug 2020 - Aug 2021

KTH Royal Institute of Technology, Sweden

MSc. Artificial Intelligence Oct 2019 - Jul 2020

Technical University of Berlin, Germany

Bachelor of Engineering in Computer Engineering Aug 2013 - May 2017

University of Mumbai, India