Arjun Sivasankar

Trier, Germany | arjun.sivasankar@gmail.com
linkedin.com/in/arjun-sivasankar-381380155 | github.com/Arjun-Sivasankar | arjun-sivasankar.github.io

PROFILE

AI Engineer with 2+ years of hands-on machine learning, LLM engineering and data science in industry and research, plus 1.5+ years of design engineering in industry. At Docs4D, a healthcare AI start-up, built and piloted SuperDoc, a LangGraph multi-agent clinical assistant with a voice agent for a German university hospital, on a self-hosted LLM stack. M.Sc. in Computational Modeling and Simulation from TU Dresden (grade 1.7) and co-author of an IEEE paper.

KEY ACHIEVEMENTS

- **Clinical AI assistant piloted with a university hospital:** built SuperDoc, a LangGraph multi-agent assistant for chat and voice targeted at 400 patients, grounded in ESC guidelines for cardiology, hypertension and diabetes; owned most of the AI work in a team of four engineers that built three products (SuperDoc, MedScribe, DocScan).
- **Self-hosted LLM stack with sub-second retrieval:** served a LoRA-fine-tuned open-weight LLM through vLLM on in-house NVIDIA GPUs so patient data never reached third-party APIs; batched cross-encoder reranking kept retrieval under 1 s with no loss in context recall.
- **Knowledge-graph-augmented LLM fine-tuning:** for the M.Sc. thesis, LoRA-fine-tuned Llama and Meditron models on 53,561 MIMIC-IV admissions with PyTorch DDP on 3 H100 GPUs; adding UMLS knowledge-graph evidence raised micro-F1 from 0.459 to 0.503 and rare-code F1 by up to 46%.

TECHNICAL SKILLS

LLMs & Agents: LangGraph, LangChain, RAG (Qdrant, hybrid search, reranking), MCP tool calling, LoRA fine-tuning (Hugging Face PEFT), vLLM, LLM APIs (Azure OpenAI, Anthropic Claude, Gemini)

ML & Data Science: PyTorch, TensorFlow, scikit-learn, XGBoost, LightGBM, pandas, NumPy, SQL, NLP (NER), time-series forecasting, A/B testing

Speech & Documents: Whisper (faster-whisper), NVIDIA Parakeet, speaker diarization, text-to-speech, OCR (PaddleOCR-VL, Azure AI Document Intelligence), structured output (Pydantic)

Evaluation & MLOps: RAGAS, LLM-as-judge, Langfuse (prompt versioning), MLflow, model benchmarking and selection, pytest, Docker, Kubernetes, Terraform, CI/CD (GitLab CI, GitHub), Git LFS (data versioning), AWS, Azure, GCP

Software & APIs: Python (async), FastAPI, REST APIs, microservices, TypeScript, React, Next.js, PostgreSQL (Supabase), Redis

Regulated AI: ISO 14971, IEC 62304, EU MDR, EU AI Act, GDPR, docs-as-code documentation (GitLab)

Languages: English (fluent), German (basic), Hindi, Malayalam, Gujarati

PROFESSIONAL EXPERIENCE

AI Engineer
Mar 2026 – Nov 2026

Docs4D GmbH (healthcare AI start-up), Trier, Germany

- Designed SuperDoc's agent architecture with a German university hospital partner, where it was piloted: grounded in clinical guidelines through RAG, with MCP tool calling for patient details and vital signs and a rule-based validation layer around the generation core for EU MDR conformity.
- Built its real-time voice agent as back-end services behind a gateway, integrated into the Guardian iOS and Android app (Parakeet streaming STT with Whisper fallback, LLM, Qwen3-TTS with CosyVoice 2 fallback); an evaluation agent checked every response before it was spoken, at about 1.5 s end to end.
- Built the RAG pipeline end to end (document parsing, hierarchical chunking, Qdrant hybrid search, cross-encoder reranking) and LoRA-fine-tuned Qwen3-30B-A3B for the clinical domain, served via vLLM on in-house NVIDIA GPUs.
- Designed the speech pipeline of MedScribe, a real-time clinical transcription service, moving from chunked faster-whisper to streaming NVIDIA Parakeet TDT 0.6B v3 because it suited real-time streaming and was multilingual; transcripts were turned into SOAP notes with NER and LLM structured output.
- Built DocScan, an OCR service for patient documents, choosing PaddleOCR-VL over DeepSeek-OCR 2 on public benchmarks and our own tests, and built Frontline, the Next.js (TypeScript, React) dashboard integrating both.
- Set up evaluation and observability with Langfuse tracing and prompt versioning, MLflow, and RAGAS plus LLM-as-judge scoring of RAG answers; A/B-tested agent prompts and benchmarked open-weight LLMs to select the baseline model.
- Deployed the back-end services with FastAPI in Docker with SQL migrations and row-level security on Supabase, and carried out ISO 14971 risk assessment along the IEC 62304 lifecycle.

Applied AI Intern / Student Assistant
Nov 2024 – Sep 2025

Vodafone Group Services GmbH, Vodafone Innovation Centre, Dresden

- Worked directly with enterprise customers to build LLM agents and their MCP servers on Azure OpenAI and Azure AI Foundry, from experiment to integration into their products, and on multi-agent architectures for Vodafone Assist.
- Built an invoice workflow for an enterprise customer: OCR with Azure AI Document Intelligence, LLM extraction validated against a Pydantic data model, then LLM summarization and gap analysis of each invoice.
- Built an on-premise service for the 6G-HEALTH research project that transcribed, separated speakers and summarized doctor–patient conversations in real time.

Student Research Assistant (part-time)
May 2023 – Oct 2024

Vodafone Chair Mobile Communication Systems, TU Dresden, Dresden, Germany

- Built PyTorch deep learning models that forecast memory-access intervals in hardware accelerators from time-series data, and co-authored the resulting IEEE ACDSA 2024 paper; wrote quantization-aware training pipelines to test accuracy below 8-bit precision.

Engineer
Jan 2021 – Aug 2022

L&T Technology Services, Chennai, India

- Wrote a Python network-analysis tool that built Design Structure Matrices from correlation data to classify wind-turbine interfaces by criticality, and automated Windchill PLM data extraction with Selenium, with Power BI dashboards for validation; named the company's Rising Star for FY 2021–22.
- Delivered CAD and drafting work and FMEA for the client Vestas (Denmark) across time zones.

Automotive Research Intern (Bachelor's thesis)
Jan 2020 – May 2020

Gannet Engineering Pvt. Ltd., Bangalore, India

- Built and compared regression and machine-learning models of NOx and soot emissions, including XGBoost and LightGBM; the selected model met the target of under 10% prediction error and was used to optimize the NOx–soot trade-off, with test points planned by design of experiments.

EDUCATION

M.Sc. Computational Modeling and Simulation
Oct 2022 – Jan 2026

Technische Universität Dresden, Dresden, Germany

Overall grade 1.7

- **Master's thesis (ScaDS.AI, grade 1.3):** "Multimodal EHR-based Diagnosis Prediction via Knowledge Graph & Neuro-Symbolic Reasoning", generating diagnoses with LoRA-fine-tuned Llama and Meditron models, grounding them to ICD-9 codes with SapBERT and FAISS and adding UMLS knowledge-graph evidence (micro-F1 0.459 to 0.503).
- **Research project:** reconstructed electrolyzer conductivity maps from magnetic-field sensors with invertible neural networks (Optuna); compared PCA, simulated annealing, genetic algorithms and k-means for sensor selection, cutting sensors from 100 to 10 at 1.5% higher loss.

B.Tech. Mechanical Engineering
2016 – 2020

Manipal Institute of Technology, India

CGPA 8.81 / 10

PUBLICATION, PATENT & AWARDS

- Friedrich, S., **Sivasankar, A.**, Matúš, E., Fettweis, G. "Enlarging the Time Budget for Neural Network Based Predictors for Access Interval Prediction." *IEEE ACDSA*, 2024.
- **Granted patent:** "A System for Ejection", a canister-aided gas ejection mechanism (India, granted 2024).
- **Leadership:** Mechanical Head and Deputy Project Manager of the thrustMIT rocketry team (2017–2019), leading 24 students to build a sounding rocket for Spaceport America 2019 (On the Spot Recognition).