NIKHIL KASUKURTHI

Staff Machine Learning Engineer

nikhil.kasukurthi@gmail.com

Göttingen, Germany · Work Authorization: Available (Germany)

Professional Summary

Staff Machine Learning Engineer with 8 years experience building **deep learning models and inference systems**. Designed evaluation frameworks for LLMs (open-sourced KARMA). Built optimized model serving via vLLM on Kubernetes, cutting inference costs by 60%. Published researcher (3 peer-reviewed papers) with hands-on experience in PyTorch, vLLM, Kubernetes and Docker

Work Experience

Independent ML Research

Göttingen, Germany | Nov 2025 – Present | *4 mos*

- Built LLM training pipeline from scratch (BPE tokenizer, pre-training, instruction tuning, RLHF, evaluation) as part of Stanford CS336
- Published technical deep-dive on distributed training optimization; benchmarked H100 SXM vs PCIe, identifying 2x cost-efficiency difference
- Trained models with DeepSpeed ZeRO-2 setup on **SLURM**-managed clusters on Runpod
- Profiled GPU utilization and memory bottlenecks using **Nsight Systems** and **PyTorch Memory Profiler**
- Built Voice Activity Detection (VAD) at the RTP packet level for a telephony product. Exported ASR models via **ONNX** for optimized edge inference, reducing costs over third-party ASR providers

Lead Data Scientist | Eka.care

Bengaluru, India | Jan 2022 – Oct 2025 | *3 yrs 10 mos*

Healthcare company building AI-powered tools for 100K+ doctors across India

LLM Evaluation & Benchmarking

- Designed and open-sourced KARMA-OpenMedEvalKit, an evaluation library for LLMs in Indian healthcare scenarios. Released datasets and models
- Gates Foundation is adopting KARMA for healthcare bot evaluation in collaboration with Eka.care
- Evaluated custom-built agent harnesses on **Tau-Bench2** with LLM generated clinical scenarios/environments, and overall benchmarks on **HELM** and MedHELM

Model Optimization & Inference

- Deployed a custom **Speech LLM** (Whisper + Gemma 2) for medical transcription via vLLM plugins. Reduced STT/ASR inference costs by 60% versus third-party API providers
- Architected inference serving for multimodal LLMs via **vLLM** on Kubernetes torch-compiled models on **RayServe**
- Applied **TensorRT** optimization for latency-critical paths. Cut inference cost by 50% vs AWS SageMaker
- Built **PySpark** pipelines for feature creation. Used **Apache Beam** to unify scattered records across DBs for unified patient profiles, across engineering and data teams

Deep Learning Products

- Architected **MedAssist**, an LLM client with remote **MCP** server support. Adopted by Apollo Hospitals (India's largest hospital chain)
- Early adopter of MCP, open-sourced MCP server to provide LLMs with Indian medical context
- Designed multi-modal retrieval pipeline using **ColQwen-2.5** and **Vespa** with hybrid search (BM25 + dense retrieval). Top-3 retrieval accuracy +24% over text-only embeddings
- Built query decomposition and understanding layer on **ElasticSearch** to normalize medical shorthands; nDCG@10 +55%
- Built **contrastive learning–based semantic retrieval models** to map unstructured clinical text to structured medical ontologies; diagnosis coding +30%, medication coding +80%
- Directed team to build medical data collection platform (Django). Gathered 1000+ medical protocol documents and 100+ hours of speech data. Managed annotation protocols with medical professionals

Leadership

- Managed a cross-functional team of 3 (1 engineer, 2 data scientists)
- Owned the AI product roadmap end-to-end. Scoped problems with product and clinical stakeholders, prioritized research bets, and drove delivery across evaluation, inference, and LLM products.
- Conducted **product interviews** and query log analysis to surface gaps between user behavior and system performance. Findings shaped the research roadmap
- Delivered internal tech talks on LLM APIs, driving team-wide adoption of Bedrock and MCP servers
- AWS Bedrock featured Eka.care as an early reference customer for healthcare AI, coordinating with AWS global teams

Data Scientist | Udaan

Bengaluru, India | May 2021 – Dec 2021 | *8 mos*

India's largest B2B e-commerce marketplace

- Built **learning-to-rank models** (gradient-boosted trees) for product search. Improved search-to-cart conversion by 10%, validated through A/B testing across business verticals
- Built 3D point cloud pipeline (DGCNN) for LiDAR-based volume estimation of warehouse shipments; 40% cost savings

Visiting Researcher | National Centre for Biological Sciences (NCBS) – TIFR

Bengaluru, India | May 2020 – Mar 2021 | *11 mos*

Concurrent with role at SigTuple

- Developed PrISM (Precision for Integrative Structural Models) using Variational Autoencoders, a novel unsupervised technique to score integrative models
- Won Best Poster Award at NCBS Annual Talks 2021
- Published in Bioinformatics (Vol. 38, Issue 15, August 2022)

Data Scientist III | SigTuple

Bengaluru, India | May 2018 – May 2022 | *4 yrs*

Healthcare AI startup building diagnostic products

- Built retinal disease detection products end-to-end, from annotation and model development through clinical validation and CE certification
- Published 2 papers at IEEE ISBI 2019
- Technical lead for two diagnostic products (Fundus, Urine analysis). Defined and executed research and engineering roadmap, coordinating across science, engineering, and clinical teams
- Led ML platform team, architected multi-model inference DAG using TF Serving, Kubernetes, and Cloud Functions. Improved turnaround time by 60% and reduced costs by 40%

Publications

PrISM: Precision for Integrative Structural Models | Bioinformatics, Vol. 38, Issue 15, August 2022. V. Ullanat, **N. Kasukurthi**, S. Viswanath

Deep Learning for Weak Supervision of Diabetic Retinopathy Abnormalities | IEEE ISBI, July 2019. M. Ahmad, **N. Kasukurthi**, H. Pande

Dynamic Region Proposal Networks for Semantic Segmentation in Automated Glaucoma Screening | IEEE ISBI, July 2019. S. Shah, **N. Kasukurthi**, H. Pande

Technical Skills

ML Systems: PyTorch, DeepSpeed, vLLM, TensorRT, ONNX, RayServe, TorchServe, TensorFlow

LLM / NLP: Transformers, BERT, Whisper, Gemma, NER/NEL, RLHF

Infrastructure: Kubernetes, Docker, SLURM, AWS, GCP, Apache Beam, PySpark

Profiling & Optimization: Nsight Systems, PyTorch Memory Profiler, GPU benchmarking

Search & Retrieval: ElasticSearch, Vespa, FAISS, ColQwen, RAG

Languages: Python, Go

Education

B.Tech – Computer Science and Engineering | VIT University, Vellore, India | 2014 – 2018 | CGPA: 8.39/10

Awards

Hackathon Winner | AWS GenAI Hackathon, August 2024. Built appointment booking agent through tool use

Impact Award | Eka.care, 2023. For major organizational impact

Best Poster Award | NCBS Annual Talks, 2021. For PrISM research presentation

Interests

- Scuba Diving - Open Water Certified
- Trekking - summited Chandrashila - 12,083 ft | 3,683 m in Himalayas
- Formula 1