

Vaishak Girish Kumar

vaishakgkumar@gmail.com | +1-716-292-4977 | vaishakgkumar.com | github.com/Gustav-Proxi | linkedin.com/in/vaishakgkumar

AI/ML Engineer and Researcher building reliable medical VLMs, agent infrastructure, and LLM observability systems.

Education

University at Buffalo (SUNY) <i>M.S. in Engineering Science (Artificial Intelligence)</i>	Buffalo, NY <i>Expected May 2027</i>
Presidency University <i>B.Tech. in Computer Science and Engineering</i>	Bengaluru, India <i>Aug 2020 – May 2024</i>

Experience

Presidency University, School of CSE, Undergraduate Researcher, Medical AI	Bengaluru, India Jan 2023 – May 2024
- Co-authored a peer-reviewed system proposal for explainable multimodal diagnostic support in resource-constrained clinics; defined the architecture, deployment constraints, and evaluation plan. - Engineered an evaluation harness comparing 4 prompt-adaptation strategies across reasoning, translation, and QA; identified feedback-based refinement as the strongest multi-agent approach.	
Tonic.AI Open-Source Program, Open-Source Developer, Scitonic LLM Pipeline	Remote Jan 2024 – Jul 2024
- Fixed 17 bugs across tokenization, preprocessing, and API modules, driving the pipeline failure rate down 20%. - Refactored 1,500+ lines of Python and vectorized pandas preprocessing for a 30% speedup; added regression tests to 10+ modules.	

Projects

FOVEA-Net <i>PyTorch, MedNeXt, MONAI, LangGraph, Hugging Face</i>	Independent Research
- Architected a 5-stage grounded chest-CT VLM combining MedNeXt encoding, region proposal, mask-conditioned attention, confidence gating, and structured report generation. - Deployed a private Docker QC workflow organizing 23 candidate findings into 69 review images for radiologist evaluation; model validation remains in progress.	
AgentOS <i>Python, vLLM, PyTorch, CUDA</i>	Independent Systems Research
- Built per-agent token and KV-cache resource controls over vLLM V1, including hard budgets, eviction policy, priority, preemption, and snapshot/restore paths. - Validated the scheduler with 84 CPU tests, 44 real-torch KV tests, and an RTX A4000 runtime covering token metering, hard-budget aborts, and paged-KV behavior.	
SalienceFormer <i>PyTorch, PEFT, LoRA, Hugging Face</i>	
- Added hippocampal salience gating and memory consolidation to a frozen Gemma-2B; reached 11.83 perplexity versus an approximately 18.0 WikiText-2 baseline, a 35% reduction, with about 15M trainable parameters. - Evaluated 15 ablation variants to isolate each component and released reproducible weights and configurations on the Hugging Face Hub.	
VectorLens <i>Python, FastAPI, SentenceTransformers, React</i>	PyPI
Published a local-first RAG observability package that attributes generated sentences to retrieved chunks; bounded analysis to 7 attribution calls per trace and added drop-in instrumentation for 5 LLM providers and 5 vector-database clients.	
AgentTraceDAG <i>Python, SQLite, FastAPI, React</i>	
Packaged a local time-travel debugger that records agent steps in an immutable SQLite DAG; its replay UI exposes context, tool payloads, and token hotspots through interceptors for 4 agent frameworks with zero telemetry.	

Technical Skills

Languages: Python, SQL, TypeScript
ML and Vision: PyTorch, Hugging Face Transformers, scikit-learn, OpenCV, MONAI, MedNeXt
LLM Systems: vLLM, LangGraph, LangChain, RAG, PEFT/LoRA, SentenceTransformers
Infrastructure: FastAPI, PostgreSQL/pgvector, SQLite, Qdrant, Redis, Docker, Next.js, AWS, GCP, ROCm, Git

Publications

- AI-DxMH: Artificial Intelligence Diagnosis for Modern Health** — first author, IJCRT, 2024. Peer-reviewed.
- On-the-fly Prompt Optimization in Multi-Agent Systems: A Comparative Study** — second author, IJCRT, 2024. Peer-reviewed.