

NITHISH KANNAN M

+91 78451 36609 | nithishkannanm1@gmail.com | Chennai, Tamil Nadu, India | [GitHub](#) | [LinkedIn](#) | [nithishkannanm.com](#)

EDUCATION

Vellore Institute of Technology (VIT), Chennai

2024 – 2028

B.Tech, Computer Science and Engineering | CGPA: 8.34/10

TECHNICAL SKILLS

Languages & Frameworks: Python, C, C++, FastAPI, LangChain (LCEL), PyTorch

ML / Retrieval: FAISS, BM25, BGE-M3, Hybrid Retrieval (RRF Fusion), TFLite INT8

Systems & Infra: Linux cgroup v2 & PSI telemetry, OpenTelemetry, MCP (Model Context Protocol), AWS

Web & Tools: React, HTML, CSS, JavaScript, SQL, Git/GitHub

EXPERIENCE

Deep Learning Engineer — AI Club, VIT Chennai

Jan 2026 – Present

- Build and review deep learning systems and research-oriented prototypes as part of the club's core AI engineering team.
- Delivered a speaker session on Cursor, an AI coding assistant, as part of an AI Club workshop.

AI/ML Engineer — AWS Cloud Club, VIT Chennai

Feb 2026 – Present

- Develop ML and cloud-integrated engineering projects on AWS infrastructure within the AWS Student Builder community.

Freelance Web/Infra Consultant — Viha Online

June 2026 – Present

- Ran a technical audit of a Shopify store (Liquid theme, GraphQL Admin API, GSC data) that surfaced a 58% checkout abandonment rate and a desktop-vs-mobile organic search gap (position 31.5 vs. 10.8).
- Queried store analytics via ShopifyQL and a Supermetrics GSC connector to trace an AOV drop to a single underperforming SKU with a 7-8x landing-page traffic gap.
- Built and shipped a custom shop-by-purpose.liquid section, wiring it into the homepage through config/settings_data.json.

RECENT PROJECTS

Modular Memory Unit (MMU) — Hybrid Retrieval RAG Framework

Jan 2026 – Present

- Built a hybrid retrieval RAG framework combining BGE-M3, FAISS, and BM25 with Reciprocal Rank Fusion; benchmarked on legal/compliance documents across 5 custom metrics including faithfulness and cross-document reasoning.
- Shipped a FastAPI backend (HybridRetriever, async/sync thread-pool pattern, SSE streaming); targeting EMNLP/ACL Student Research Workshop.

Ring Zero — Predictive Memory Orchestrator

Samsung EnnovateX Hackathon, Apr – May 2026

- Co-built a predictive memory orchestrator combining a multi-modal next-context transformer with a PPO-trained cgroup v2 tier policy.
- Outperformed the Android App Standby Buckets baseline: 90% vs 86% warm rate and 192ms vs 217ms latency at a 400MB memory budget.

ex_mem — Contradiction-Aware Shared Memory Server (MCP)

Jun – Jul 2026

- Designing an MCP-based memory server that treats changed facts as evolution with preserved history rather than overwrite, targeting a gap where all 22 systems evaluated on MemoryAgentBench's FactConsolidation benchmark score $\leq 7\%$.
- Built on a hybrid SQLite FTS5 + sqlite-vec retrieval stack with an LLM-judge-based contradiction classifier.

Plateau — Semantic Circuit Breaker for AI Agents

VIT Chennai Frontier AI Hackathon, Jul 2026

- Built a detector that trips on semantic information stagnation (not textual repetition) in autonomous agent loops, using offline sentence-embedding comparison with a self-calibrating baseline.
- Validated with 121 passing tests and a live two-machine demo; shipped as the installable package plateau.

CERTIFICATIONS & ACHIEVEMENTS

- Machine Learning Specialization — Stanford University / DeepLearning.AI (Supervised Learning, Advanced Learning Algorithms, Unsupervised Learning & Reinforcement Learning).
- Generative AI Engineering — IBM (Transformer-based Language Modeling, Fine-Tuning Transformers, Foundational Models for NLP).
- 3rd Place, BIS National Hackathon 2025 — Empath, an emotion-aware adaptive learning platform (React, FastAPI, HuggingFace, Supabase).
- Published adaptive-rag-router on PyPI: a pip-installable, embedding-similarity routing layer for RAG pipelines with confidence-based refusal and OpenTelemetry tracing.