NIDHUSHAN KANAGARAJA

nidhushank6@gmail.com | +1 (503) 421-2971 | **Open to Relocation**
Portfolio | LinkedIn | LeetCode | GitHub | HF Dataset

RESEARCH

Mnemosyne: Time-Aware Episodic Memory for RAG Agents Sep 2025 – Present
Independent Research – Long-term Memory Architectures for Personal AI

- Engineered a scalable memory substrate for 98K StackOverflow events achieving 95–97% schema-valid ingestion.
- Designed time-aware hybrid retriever (lexical + dense + decay) giving 15% precision@5 on temporal QA.
- Consolidated month-scale histories into 1K episodic units, reduced memory footprint 10× preserving semantic recall.
- Built FAISS/HNSW index for 100K+ MiniLM embeddings, compressed 303MB to 60MB with sub-15ms latency.

Moonshine: Text-Conditioned Procedural Content Generation (AAAI 2025) Jan 2024 – Apr 2025

- Developed a text-conditioned PCG pipeline using a synthetic dataset of 70K map–description pairs.
- Achieved 85% semantic alignment by training the Five-Dollar Model and discrete Diffusion Model as T2M baselines.
- Created dataset splits with 49K training, 14K test, and 7K validation examples for controlled generative evaluation.
- Evaluated models using BLEU (54.7), ROUGE-L (33.2), METEOR (19.5), SPICE (11.3), and CLIP similarity.

EXPERIENCE

AI/ML Engineer | Giinius | New York, NY May 2024 – May 2025

- Fine-tuned open-weight LLMs using LoRA and p-tuning; improved grounding precision by 15%.
- Integrated DPR with NVIDIA NeMo for retrieval; raised context match scores by 18%.
- Reduced inference time from 1.2s to 850ms using pruning, distillation, and quantization.
- Improved LLM response quality by routing queries through LoRA adapters with retrieval-aware model selection.

Software Engineer | Citibank (TCS) | Chennai, India Jul 2022 – Jun 2023

- Engineered Python/Shell frameworks for SFTP automation, housekeeping, and outbound pipelines.
- Led testing and delivery of modules supporting compliance-ready financial transactions.

ML Engineer Intern | HighRadius | Chennai, India Jun 2021 – May 2022

- Developed payment delay prediction model (92% accuracy) and credit risk scoring chatbot.

PROJECTS

Mnemo: AI-Powered Diary & Voice Journal System Nov 2025 – Present

- Designed an LLM-driven diary system converting free-form notes into structured logs using on-device 3B/8B models.
- Constructed a multimodal pipeline with faster-whisper STT and Piper TTS for speech journaling and voice playback.
- Deployed the system on an Oracle Cloud VM using FastAPI, Nginx, HTTPS, and Docker for low-latency operation.

SkillBarter: C2C Peer Service Marketplace | NYU Capstone May 2024 – Dec 2024

- Led a team of 6 to design and develop a peer-to-peer service exchange platform using time-credit economy.

EDUCATION

New York University Sep 2023 – May 2025
M.Sc. in Computer Science – *Merit-Based Scholarship, Co-author: AAAI 2025 Paper* GPA: 3.75/4.0

SRM University Jun 2018 – May 2022
B.Tech in Computer Science – *First Class with Distinction* GPA: 3.68/4.0

SKILLS

AI/ML: Generative Models, Transformers, NLP, RAG, Reinforcement Learning, Model Optimization
Frameworks & Libraries: PyTorch, TensorFlow, NVIDIA NeMo, LoRA, ONNX, LangChain, OpenAI API
Programming: Python, C/C++, Shell Scripting, JavaScript, SQL
Backend & Systems: FastAPI, RESTful APIs, Docker, Linux, GitHub Actions, CI/CD Pipelines