

VIDUSHI ANAND

Gurugram, Haryana, India

[vinspace](https://vinspace.co) | vidushianand09@gmail.com | <https://www.linkedin.com/in/vidushii-anand/> | <https://github.com/vidushi2709>

EDUCATION

Bachelor of Technology in Computer Science specialized in AIML

Dronacharya College of Engineering | 8.4

Gurugram, Haryana, India

Aug. 2023 – 2027

EXPERIENCE

SOFTWARE AI ENGINEER INTERN

June 2025 – October 2025

API.market (MagicAPI Inc | Noveum.ai)

San Francisco, California, USA - Remote

- Fixed, optimized, and deployed production-ready APIs using **Python, Docker, Cog, FastAPI, and Postman**.
- Designed, built, and productionized a high-performance vector search system for low-latency, large-scale semantic retrieval.
- Built 30+ automated RAG evaluation metrics to benchmark retrieval and generation quality, enabling systematic model comparison and debugging

TECHNICAL PROJECTS

- VoxLoad: Speech-Based Cognitive Load Estimation** | [GitHub](#) | Python, Whisper, FFmpeg March 2026
 - Built a **multi-agent pipeline** to estimate **cognitive load** from speech.
 - Extracted **acoustic signals** including **pause patterns, filler-word frequency, and speech-rate features**.
 - Implemented **per-speaker z-score normalization** and **multi-signal aggregation** to produce normalized cognitive-load scores with confidence.
- Memory: Persistent AI Memory System** | [GitHub](#) | Python, DSPy, ChromaDB February 2026
 - Built an **LLM-powered memory layer** with persistent vector storage, semantic retrieval, and multi-user isolation.
 - Implemented **384-dim embedding recall with top-5 similarity search** and historical memory queries.
 - Designed a **ReAct-based memory agent** supporting ADD / UPDATE / SUPERSEDE operations with versioned memories, timestamps, and background writes.
- SafeToSay: Clinical AI Agent System** | [GitHub](#) | LangGraph, MedGemma, Mistral, FastAPI, Python January 2026
 - Built a **5-agent clinical QA pipeline** that answers only when safety and evidence thresholds are met; otherwise abstains.
 - Orchestrated **MedGemma-4B** generation with 4-bit quantization and Mistral reasoning agents using LangGraph state machines.
 - Increased safety pass rate from 53.6% → 98.2% across 56 evaluation queries (+44.6pp) using runtime evaluation agents.
- Build LLM Playground: GPT-2 from scratch** | [GitHub](#) | PyTorch, Python August 2025
 - Implemented **GPT-2** with multi-head attention, positional embeddings, layer normalization, and feed-forward networks
 - Pretrained on custom corpora using tokenization, mini-batch gradient descent, and cross-entropy loss.
 - Fine-tuned for **spam classification** by replacing the final layer with a 2-node classifier; experimented with tokenization, gradients, and mini GPT models.

TECHNICAL SKILLS

- Languages & Databases:** Python, Java, C, R, SQL
- AI/LLM Systems:** LLM Agents, Multi-Agent Systems, Retrieval-Augmented Generation (RAG), Vector Embeddings & Semantic Search, Prompt Engineering, LoRA Fine-Tuning, LLM Evaluation
- Frameworks & Infra:** PyTorch, Hugging Face Transformers, LangGraph, LangChain, DSPy, LiteLLM, ChromaDB, FAISS, NumPy, Pandas, Scikit-learn

LEADERSHIP & ACHIEVEMENTS

- Co-Lead, Deviators Club:** Organized **Debug Decrypt 2.0**, a 3-day DSA competition with 110+ teams, and conducted tech skill sessions for club members.
- SIH 2025 Grand Finalist:** Worked on an edge-based IoT controller for real-time phase imbalance detection and dynamic load switching in distribution feeders.
- Kaggle Notebook Expert:** Top 3% globally (Rank 1,569 / 60K+)