VIDUSHI ANAND

Gurugram, Haryana, India

vidushianand09@gmail.com | https://www.linkedin.com/in/vidushii-anand/ | https://github.com/vidushi2709

EDUCATION

Bachelor of Technology in Computer Science specialized in AIML — **Gurugram, Haryana, India**
Dronacharya College of Engineering | 8.3 — *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.
- Developed **30+ RAG evaluation scorers** and **pipelines** for automated benchmarking and quality assessment of retrieval-augmented generation systems.

TECHNICAL PROJECTS

- **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 the model on **custom corpora** with tokenization, mini-batch gradient descent, and cross-entropy loss; tracked losses and generated sample outputs to assess model performance.
  - Fine-tuned for **spam classification** by replacing the final layer with a 2-node classifier; experimented with tokenization, gradients, and mini GPT models.
- **TerminaLoRA: CLI Assistant using TinyLlama and LoRA** | GitHub | Python, PEFT — June 2025
  - Adapted a 1.1B TinyLlama model with LoRA to translate natural language into shell commands.
  - Developed a compact adapter architecture enabling fast inference and deployment in resource-constrained environments.
  - Built logging and dry-run execution features to ensure command safety and traceability during automation.
- **Law Vector: AI-Powered Legal Document Search Tool** | GitHub | *Sentence Transformers, Pinecone, FastAPI, Docker, PyPDF2, Perplexity* — May 2025
  - Co-developed a semantic search engine for legal documents, achieving **5× faster** retrieval than keyword search.
  - Implemented a **document chunking + transformer embedding pipeline**, integrating with **Pinecone DB** for scalable, high-precision vector search.
  - Deployed a **FastAPI backend** to handle real-time legal queries with **<300ms latency**, ensuring smooth user experience.
- **Transformer-Based Language Model (LLM) Development** | GitHub | *PyTorch, Dataset: WMT-14* — May 2025
  - Engineered a **transformer model from scratch** for robust, scalable language generation.
  - Fine-tuned on the **WMT-14** German-to-English dataset, surpassing baseline BLEU scores and improving translation accuracy.

TECHNICAL SKILLS

- **Languages & Databases:** Python, Java, C, R, SQL, MongoDB
- **AI/ML & NLP:** Supervised & Unsupervised Learning, CNNs, RNNs, LSTMs, Transformers (BERT, GPT), Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Transfer Learning (ResNet, YOLOv8), Model Optimization, Vector Search & FAISS
- **Frameworks & Tools:** TensorFlow, PyTorch, OpenCV, NumPy, Pandas, Scikit-learn, Streamlit, FastAPI, Docker, Postman, Git, Kaggle, Google Colab

LEADERSHIP & ACHIEVEMENTS

- **Co-Lead, Deviators Club:** Oversaw logistics, challenge design, and event coordination for Debug Decrypt 2.0, a three-day DSA event with over 110 teams that saw a 35% increase in turnout from the previous year.
- **SIH 2025 Grand Finalist:** Worked on an edge-based IoT controller for real-time phase imbalance detection and dynamic load switching in distribution feeders.
- **Impact Challenge IIT-M (Top 40):** Won a ₹50K grant for creating a biodiversity tech solution leveraging DETR and ViT for high-precision image-based detection.