Avinash Tiwari

✉ 22bds010@iiitdwd.ac.in ☎ +91-8299661089 github.com/avinash4002 in linkedin.com/in/avinash-tiwari-bba572278/

Education

Bachelor of Technology in Data Science and Artificial Intelligence *2022 – 2026*
Indian Institute of Information Technology, Dharwad *CGPA: 9.1/10.0*

Technical Skills

- **Programming Languages:** Python, SQL
- **Machine Learning & AI:** PyTorch, TensorFlow, Scikit-learn, Transformers, YOLO, OpenCV
- **Deep Learning & NLP:** Hugging Face, LLaMA, LoRA/QLoRA, LangChain, Claude, GPT, BERT, Mistral
- **Model Optimization:** Quantization, PEFT, Fine-tuning, High-performance computing
- **Deployment & Engineering:** Docker, FastAPI, Flask, Streamlit, Git
- **Data Management:** Pandas, NumPy, MySQL

Experience

Project Intern *Jan 2025 – April 2025*
Vocab.ai (Remote)

- Built an end-to-end multilingual conversational AI pipeline using AI4Bharat ASR, Mistral LLM for response generation, and Indic TTS
- Fine-tuned Wav2Vec2 on 50+ hours of noisy telephonic data to address poor ASR performance caused by compression artifacts and low bandwidth
- Achieved a 96.76% reduction in training loss, 98.3% reduction in validation loss, and reduced Word Error Rate (WER) to 20%

Projects

Domain-Specific LLM for Financial Analysis

- Fine-tuned LLaMA 2 7B on financial data using LoRA, reducing training time by 73% while maintaining 91% domain accuracy
- Implemented quantization techniques, reducing model size by 62% with only 3% performance degradation
- Deployed the model using FastAPI and Docker on Render, achieving 250ms response time with 99.7% uptime

Technologies: Python, PyTorch, Transformers, PEFT, FastAPI, Docker, Render

Multimodal RAG System for Enterprise Knowledge Base

- Developed RAG integrating text, image, and tabular data for enterprise knowledge retrieval
- Implemented vector embedding chunking strategies with dynamic context window sizing, improving retrieval accuracy by 43%
- Created a custom evaluation framework to assess hallucination reduction, achieving 89% factual accuracy vs. 67% in baseline LLM
- Optimized response generation with 78% lower latency through efficient indexing and parallel query processing

Technologies: Python, LangChain, Llama-Index, FAISS, OpenAI API, Pinecone, ChromaDB, FastAPI, Docker, Streamlit

AI Use Case and Resource Recommendation Agent

- Developed a Python system to provide businesses with targeted AI solutions and the top 5 relevant resources
- Utilized Google Gemini for company profile summarization via web scraping and generating tailored AI use case recommendations
- Integrated multiple APIs to curate top resources per category, employing aiohttp for optimized data retrieval

Technologies: Python, aiohttp, requests, BeautifulSoup, Google Gemini, Google Custom Search API, Hugging Face, Kaggle API, GitHub API, arXiv API

Achievements

- **Top 17 Finalist** – SBI Hack-AI-Thon (2025)
- **Ranked 138** – Amazon ML Challenge (2024)
- **GATE DA Rank: 949** (Top 2% in this Exam)