Sanatan Khemariya

7898009986 | sanatankhemariya@gmail.com | linkedin.com/in/sanatank | github.com/sanatren

Education

Jaypee University of Engineering and Technology — Guna, India
Bachelor of Technology in Computer Science and Engineering — *Aug. 2022 – Jul. 2026*

IIT Guwahati (Certification Program) — Guwahati, India
Credit-Linked Program in Data Science — *Jan. 2025 – Jan. 2026*

Experience

AI Engineering Intern — May 2025 – August 2025
BharatNiti AI — *Remote*

- Architected a comprehensive **financial research analyst** for investment funds and traders, leveraging financial data from 6,262+ Indian companies with a FastAPI backend and real-time streaming frontend — achieved **10–50× performance improvements** over traditional RAG systems
- Built **multi-layer intelligent caching** with PostgreSQL primary + SQLite fallback, content-based SHA-256 invalidation and sliding window retention — achieved **95%+ cache hit rates** and reduced storage growth by 60%
- Engineered **parallel document processing** pipeline with ThreadPoolExecutor, reducing fresh company processing from 10–20 minutes to **2–5 minutes** with 99% fault tolerance
- Implemented **AI-powered query optimization** with semantic expansion and multi-query processing — delivered **5–10× more relevant results** with 92%+ query relevance
- Developed **real-time cost tracking** system with token-level granularity across OpenAI/Vertex AI APIs, achieving **60–90% cost reduction**
- Created **temporal intelligence engine** for complex time-based financial queries with 90% accuracy

Projects

Azazel - Legacy Model Enhancement | *LLMs, OpenAI, FAISS, LangChain, Streamlit* | Link

- Built AI toolkit transforming legacy language models into multimodal assistants with text generation, image analysis, code execution, and web search capabilities
- Integrated GPT-4o-mini vision with **LangChain, RAG and FAISS vector storage**, improving response accuracy by **25%** on complex queries
- Reduced **API costs by 30%** through prompt optimization, conditional model routing, and response caching

Legal Document Analyzer | *BART, Transformers, PPO RL, Hugging Face* | Link

- Engineered NLP system for legal document summarization using Domain adapted BART model and custom transformer architecture
- Fine-tuned **facebook/bart-large-cnn** on legal datasets, improving summary coherence for domain-specific legal text
- Constructed **custom transformer** with **Byte Pair Encoding** (50k vocabulary), demonstrating multi-head attention mechanisms

Easy-Notes - AI Writing Assistant | *FastAPI, TensorFlow, Transformer, Supabase* | Link

- Designed full-stack AI note-taking platform with intelligent text suggestions and translation capabilities
- Trained LSTM next-word prediction model achieving **66% accuracy** and Transformer-based English-to-Hindi translation model reaching **94.6% accuracy**
- Implemented RESTful API with FastAPI serving ML predictions under 100ms latency, optimized PostgreSQL queries reducing response times by **25%**

Additional Experience

TATA Innovant GenAI Hackathon Finalist — TATA Technologies
Led the only selected team from Madhya Pradesh among 2,600+ teams nationwide

Technical Skills

Languages: Python, C++, Swift
Frameworks: FastAPI, LangChain, OpenAI SDK, PyTorch, TensorFlow, Scikit-learn
Databases: PostgreSQL, MySQL, Vector Databases
Tools & Technologies: Google Vertex AI, AWS S3, Supabase, Git, GitHub, Hugging Face, Google Colab

Certifications & Achievements

Kaggle — Pandas for Data Analysis | Link
Kaggle — Generative AI Intensive | Link
LeetCode: Solved 300+ problems across algorithms and data structures | Profile