

# Sanatan Khemariya

7898009986 | [sanatankhemariya@gmail.com](mailto:sanatankhemariya@gmail.com) | [linkedin.com/in/sanatank](https://linkedin.com/in/sanatank) | [github.com/sanatren](https://github.com/sanatren)

## EDUCATION

### Jaypee University of Engineering and Technology

Bachelor of Technology in Computer Science and Engineering

Guna, India

Aug. 2022 – Jul. 2026

### IIT Guwahati (Certification Program)

Credit-Linked Program in Data Science

Guwahati, India

Jan. 2025 – Jan. 2026

## EXPERIENCE

### AI Engineering Intern

BharatNiti AI

May 2025 – November 2025

Remote

- Architected a comprehensive **financial research analyst** for investment funds and traders, leveraging financial data from 6,262+ Indian companies with a FastAPI backend achieved **10× performance improvements** over traditional RAG systems
- Built **multi-layer intelligent caching** with PostgreSQL primary + SQLite fallback and sliding window retention — achieved **95%+ cache hit rates** and reduced storage growth by 60%
- Engineered **parallel document processing** pipeline with ThreadPoolExecutor, reducing fresh docs 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
- Created **temporal intelligence engine** for complex time-based financial queries with 90% accuracy

### Azazel - Legacy Model Enhancement | LLMs, OpenAI, FAISS, LangChain, FastAPI | [Live Demo](#) | [GitHub](#)

- Built AI toolkit transforming legacy language models into multimodal assistants with text generation, image analysis, code execution, and web search capabilities
- Developed robust **FastAPI backend** supporting streaming chat, document processing, and multimodal request routing, ensuring sub-100ms response latency under load
- 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](#) | [GitHub](#)

- 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 | [GitHub](#)

- 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

Led the only selected team from Madhya Pradesh among 2,600+ teams nationwide

TATA Technologies

## 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](#)