Sarthak Sanjeev

Artificial Intelligence Engineer & Full Stack Developer

Email: sarthaknsanjeev@gmail.com **LinkedIn:** www.linkedin.com/in/sarthaknsanjeev
GitHub: www.github.com/Hooper-at-Dev **Portfolio:** sarthak-portfolio-11.vercel.app

---

Summary

4th year Computer Science student specializing in Deep Learning, AI, ML, and Full-Stack Development with 2+ years of experience. Skilled across CNNs, LSTMs, Transformers, Generative AI, LLMs, RAG, and end-to-end model development using PyTorch and TensorFlow. Built GPT-2 and LLaMA-3–style models from scratch, including ALPACA instruction tuning, showcasing strong mastery of transformer internals. Delivered production AI systems, including an AI background-remover (Next.js + AWS) and a local LLM with RoPE, KV-cache, streaming decode, and WEB-RAG. Proficient in React.js, Next.js, Node.js, Python, and vector DBs (FAISS, Pinecone), with a focus on scalable, real-world deployment.

Education

Bachelor of Technology in Computer Science
Bennett University, India
Jul 2022 - Jun 2026

Certifications

- **IBM Machine Learning Professional Certificate:** Comprehensive specialization covering supervised, unsupervised, and deep learning using Python. **View**
- **DeepLearning.AI — Improving Deep Neural Networks:** Focused on hyperparameter tuning, regularization, and optimization techniques. **View**

Technical Skills

- **AI/ML:** PyTorch, TensorFlow, Scikit-Learn, LangChain, OpenCV, Matplotlib, NLTK, Generative AI, LLMs, RAG, CNN, LSTM, Machine Learning, Fine-Tuning, CI/CD pipelines
- **Programming & DB:** Python, JavaScript, Vector DB,FAISS ,PineCone, Postgres SQL
- **Web Development & Deployment:** React.js, Next.js, Node.js, Flask, TailwindCSS, HTML, CSS, Azure, AWS
- **Tools:** VS Code, Postman, Anaconda, Canva, Git, ChatGPT

Achievements

- Published research achieving 94% accuracy in CNN-based music genre classification.
- Independently developed GPT-2 and LLaMA-3 models and integrated them into real-world projects.

Work Experience

Jr. Software Developer
Walsis eConnect India (P) Ltd, Remote
Oct 2024 - Present

- Responsible for development and deployment of production ready web apps.
- Actively participate in agile workflows, including sprint planning and code reviews, while writing unit tests and debugging code to maintain high-quality, reliable software in a fast-paced startup environment.
- Developed and deployed AI-driven web solutions for **bgZap.com**, managing the full development lifecycle from architecture to production.
- Implemented deep learning models for image processing and integrated them into live web apps.
- Collaborated with backend and design teams to optimize performance, usability, and scalability.

Projects

GPT-2 Instruction Tuned on ALPACA Prompts GitHub

- Recreated the **GPT-2 model architecture** from scratch following OpenAI's original paper.
- Configured a 355M parameter transformer and loaded pretrained weights for fine-tuning on custom datasets.
- Instruction Fine-tuned it on ALPACA prompts using Colabs a100 GPU.
- Demonstrated an in-depth understanding of transformer internals, mathematics and LLM fine-tuning workflows.

bgZap – AI Background Remover WebApp

- Created a production-grade AI-powered tool for seamless **image background removal and deployed on AWS**.
- Integrated the model with a **Next.js front-end** featuring smooth UI transitions.
- Built a secure payment gateway via **Node.js + PayPal API** for monetization and user management.

BLUE – LLAMA 3 style custom LLM GitHub

- Engineered a production-ready local LLM inspired by llama 3 architecture with RoPE, KV-cache, and optimized from scratch.
- Built an advanced conversational pipeline with streaming decode, prompt handling, and cache-safe context flow.
- Integrated optional WEB-RAG via real time web scrapping for external context retrieval and enhanced response quality.