Sarthak Kumar Maurya

+917011445426 | sarthakmaurya04@gmail.com | linkedin.com/in/sarthax11 | github.com/sarthexe

Education

Guru Gobind Singh Indraprastha University — Nov 2022 – Present
Bachelor of Technology in Artificial Intelligence and Machine Learning — *Surajmal Vihar, Delhi*

Vivekanand School — Apr 2020 – Apr 2022
CBSE — *Anand Vihar, Delhi*

Experience

AI Intern — Jul 2025 – Jan 2026
Neolytix — *Gurugram, Haryana*

- Collaborated with billing teams to engineer CPT Code Automation and RAG-based claims agents using LangChain, interpreting user requirements to automate 65% of provider ID resolutions and reduce manual query time by 30%.
- Solved persistent model hallucination issues by devising an advanced prompt tuning strategy in Neoscribe, effectively reducing error rates by 35% and restoring clinical accuracy for healthcare end-users.
- Refactored backend modules to resolve 50+ SonarQube critical issues, reducing technical debt and achieving an 'A' grade for code maintainability.

Machine Learning Research Intern — Jul 2024 – Aug 2024
Guru Gobind Singh Indraprastha University — *Surajmal Vihar, Delhi*

- Developed supervised ML models (Naive Bayes, Logistic Regression) for spam detection, achieving 98.3% accuracy and reducing false positives by 15% via rigorous hyperparameter tuning.
- Processed and visualized a 5,000+ email dataset using TF-IDF to identify key patterns, communicating data insights to research mentors to guide feature engineering strategies.

Projects

Mental Health MCP Project | *Python, MCP, Fine-tuned LLM, Docker* — Sep 2025

- Architected a local-first AI agent using the Model Context Protocol (MCP), ensuring 100% of sensitive user PII (journals/logs) remains stored on-device while enabling seamless LLM interoperability.
- Implemented 3 custom MCP Tools for secure journaling and mood tracking, utilizing optimized context retrieval to reduce prompt token usage by 40% compared to standard full-context injections.
- Engineered a low-latency crisis detection guardrail (processing inputs in ¡50ms) that identifies high-risk keywords with 95% accuracy, immediately preempting the model to surface emergency resources.

Text Summarizer | *Python, PyTorch, Pegasus, NLP* — Apr 2025

- Implemented a Transformer-based text summarization pipeline utilizing the Pegasus model; reduced manual summarization efforts by 40%, enabling analysts to process 500+ more documents monthly.
- Optimized model hyperparameters to achieve ROUGE scores of 40.17 (R1), 20.6 (R2), and 40.11 (RL).
- Created a flexible and modular pipeline for data preprocessing and inference, ensuring consistent code quality and a 30% increase in deployment speed.

YouTube RAG-based Q&A System | *Python, LangChain, RAG Systems, Gradio* — Mar 2025

- Developed a Retrieval-Augmented Generation (RAG) system allowing users to ask context-aware questions on YouTube content, improving information retrieval relevance by 40%.
- Integrated OpenAI and LangChain with YouTube transcription and semantic search, increasing answer accuracy by 35% and reducing response latency by 20%.

Technical Skills

Languages: Python, C/C++, SQL (Postgres), JavaScript, HTML/CSS
Frameworks: Hugging Face Transformers, PyTorch, LangChain, FastAPI, MCP
Developer Tools: Git, Docker, Azure, Jupyter, Google Colab, uv, MLflow
Key Skills: Generative AI, Prompt Engineering, NLP Pipelines, Web Crawling, Deep Learning, Predictive Analysis
Certifications: AWS Certified AI Practitioner, Machine Learning Specialization (Coursera), RL Onramp (MathWorks)