Monis Shaikh

monis.shaikh.0212@gmail.com — +91 8878618087 — LinkedIn — Github

Education

VIT Bhopal University **Bhopal, MP**
B.Tech in Computer Science, Specialization in Cloud Computing *Expected May 2026*

Experience

Codefeast **Remote**
Gen AI LLM Engineer Intern *Sept 2025 – Present*

- Authored highly complex instruction data for a production LLM, creating scenarios that averaged 4+ sequential API calls and consistently achieved a 95% quality acceptance rate, thereby accelerating the overall fine-tuning timeline.
- Conducted error analysis on LLM-generated API calls, providing critical feedback for the RLHF process to significantly improve instruction-following fidelity and reduce tool-use errors.

ASolution **Remote**
Software Developer Intern *May 2024 – Aug 2025*

- Engineered a complete website redesign using React.js and Tailwind CSS, achieving a 30% reduction in load time and ensuring 100% responsiveness across all mobile and desktop viewports.
- Boosted key site performance metrics by optimizing image assets and implementing code splitting, raising the Lighthouse performance score from 60 to 81 and the SEO score from 74 to 92.
- Managed end-to-end deployment via cPanel, ensuring 99.9% uptime by implementing a robust backup and recovery plan to mitigate potential site failures.

Projects

InsightEngine — AI-Powered Document Intelligence Platform *Tech: Python, FastAPI, GCP, TinyBERT, Pinecone*

- *Advanced to the second round of Bajaj Finserv's HackRx 6.0, placing 242nd out of an initial field of over 46,000 registrants.*
- Produced a novel RAG pipeline using a HyDE generator and a TinyBERT cross-encoder, increasing the model's overall answer accuracy by 233% over the initial baseline.
- Optimized the retrieval and ranking algorithm, reducing the average end-to-end query response time by 23.5%.
- Deployed the full application on Google Cloud Platform (GCP) with a high-performance FastAPI backend, implementing a two-stage cache that cut latency for repeated queries by 95%.

DeepMetrics — AI Model Benchmarking Tool *Tech: Next.js, FastAPI, TensorFlow, PyTorch, ONNX*

- *MERNxAI Hackathon Runner-Up Winner*
- Led an automated benchmarking infrastructure to evaluate TensorFlow, Pickle, PyTorch, and ONNX models, which measured 5+ key performance metrics (e.g., latency, memory, CPU usage) and reduced manual testing efforts by 85%.
- Developed a high-performance FastAPI backend and a responsive Next.js frontend, implementing lazy loading for heavy components to achieve a 40% improvement in overall site load time.

Skills

- **Languages:** Python, JavaScript, TypeScript, C++
- **AI/ML:** PyTorch, TensorFlow, Pandas, NumPy, Scikit-learn, Hugging Face
- **Frontend:** React.js, Next.js, Tailwind CSS, Streamlit, GSAP
- **Backend & Databases:** FastAPI, RESTful APIs, MySQL, MongoDB, Pinecone
- **Cloud/DevOps & Tools:** GCP, AWS, Docker, Git, Linux, Figma, Jira