HARSH GUPTA

**Gurugram,Haryana • +91-7081771202•**[**harshmail281199@gmail.com**](mailto:harshmail281199@gmail.com) **LinkedIn:** https://[www.linkedin.com/in/harsh-gupta-dev/](http://www.linkedin.com/in/harsh-gupta-dev/) **Github:**https://github.com/Hartz-byte

**Portfolio:** https://portfolio-3-kappa-rose.vercel.app/

# SUMMARY

AI/ML Engineer skilled in designing, training, and deploying deep learning models for NLP, Computer Vision, and Generative AI. Proficient in Python, PyTorch, TensorFlow, and modern MLOps tools like Docker, AWS, and FastAPI. Experienced in building end-to-end AI systems and software solutions that bridge research and real-world impact. Actively developing automation and AI-powered applications to solve practical problems.

# KEY PROJECTS

## AI Job Agent - Automated Job Search and Application System

 Developed a full-stack job application assistant integrating LinkedIn, Indeed, Wellfound, and Internshala job listings.  Leveraged LLMs to auto-tailor resumes and cover letters per job, boosting application effectiveness.

 Built a Streamlit dashboard for real-time job search & application tracking with SQLite backend, ensuring scalable and reliable operations.

 Personally utilized the system to apply to multiple companies, demonstrating its real-world impact successfully.

## EchoVision – Memory-Enabled Multimodal AI Chatbot (LLM + Stable Diffusion)

 Architected an AI chatbot combining Mistral 7B LLM for multilingual conversation with Stable Diffusion v1.5 for local image generation.

 Implemented token-aware memory with persistent conversation histories and efficient token trimming.

 Designed a FastAPI backend optimized for VRAM usage with async endpoints and session management.

 Developed a responsive Streamlit frontend with real-time status, auto language detection (English/Hindi/Spanish/French), and dynamic state management.

# EXPERIENCE

## ML Intern | Tensaw Technologies

## June 2025 - August 2025

Developed and deployed ML models for payment-risk scoring and anomaly detection using Python, scikit-learn, and TensorFlow.

Built Dockerized AWS pipelines adhering to HIPAA standards. Presented insights and collaborated in agile sprints.

## Full Stack Engineer | Sinqlarity Jan 2023 - March 2025

Delivered scalable MERN-stack web apps with optimized backend APIs and secure payment flows. Integrated RESTful AI microservices, managed CI/CD to reduce release time by 25%.

Mentored junior developers and performed code reviews.

# SKILLS

**Programming:** Python | JavaScript | TypeScript | SQL | Pandas | NumPy

**AI/ML s Deep Learning:** PyTorch | TensorFlow | Keras | Transformers | LangChain | RAG | Stable Diffusion | LLM fine-tuning | NLP Pipelines | scikit-learn | Web Scraping | LSTM | CNN | RNN

**MLOps s Cloud:** Docker | AWS (EC2, S3) | FastAPI | Git | CI/CD | Microservices | GPU Optimization

**Full Stack s Databases:** Streamlit | React.js | Node.js | Express.js | MongoDB | PostgreSQL | Redis | SQLite

**Additional:** MCP | Automated Job Search Systems | LLM-driven Document Tailoring | Rate Limiting & Error-Handling | Agile Collaboration | Communication | Teamwork

# EDUCATION AND CERTIFICATIONS

 B.Tech Coursework Completed (2019–2023), Graphic Era Deemed to be University  Machine Learning Specialization – Coursera (Andrew Ng)

 Deep Learning Specialization – Coursera (Andrew Ng) ([Link](https://coursera.org/share/5ba2cd0a301f55cbe2f829c0940a39a3))

# RESEARCH PUBLICATIONS

 Transformers Beyond NLP: A Survey on Vision, Speech, and Multimodal Applications ([Link](https://hartz-byte.github.io/survey-transformers/))

 A Survey on Hyperparameter Tuning, Regularization, and Optimization in Deep Neural Networks ([Link](https://hartz-byte.github.io/survey-deep-hpo/))  A Beginner’s Survey of Deep Neural Networks: Foundations and Architectures ([Link](https://hartz-byte.github.io/survey-paper-dnn/))