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

**Kaggle:** https://[www.kaggle.com/hartzbyte](http://www.kaggle.com/hartzbyte)

# SUMMARY

AI/ML Engineer with hands-on experience designing, training, and scaling deep-learning solutions for NLP, computer vision, and generative AI. Skilled in Python, PyTorch, TensorFlow, and modern MLOps (Docker, AWS, FastAPI). Proven at building end-to-end pipelines, optimizing model performance, and integrating large-language-model (LLM) services into production applications. Agile collaborator who bridges research prototypes and real-world impact.

# KEY PROJECTS

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

 Architected an intelligent conversational system combining Mistral 7B GGUF for multilingual dialog with Stable Diffusion v1.5 for local image generation on consumer GPUs.

 Implemented token-aware memory management with persistent conversation histories, enabling contextual understanding across sessions and intelligent conversation trimming to prevent token overflow.

 Built CORS-enabled FastAPI backend with asynchronous endpoints, VRAM-optimized inference (attention slicing, CPU offload), and robust session management for chat/image persistence.

 Developed responsive Streamlit frontend featuring real-time status monitoring, automatic language detection (English/Hindi/Spanish/French), and dynamic conversation state management.

## LLM-Powered Research Assistance System

 Created a multi-agent pipeline for NER, question answering, summarization, and sentiment analysis using BiLSTM, Transformer, and BiDAF architectures (PyTorch).

 Automated large-scale preprocessing, benchmarking, and evaluation; delivered reusable modules for rapid experimentation and production integration.

 Improved inference reliability with standardized logging, exception handling, and unit tests.

# EXPERIENCE

## ML Intern | Tensaw Technologies June 2025 - August 2025

 Developed and deployed ML models for payment-risk scoring and anomaly detection in healthcare-fintech datasets (Python, scikit-learn, TensorFlow).

 Executed data preprocessing, feature engineering, and model evaluation; presented insights to stakeholders in weekly sprint reviews.

 Built Dockerized pipelines on AWS, adhering to HIPAA-compliant data-governance standards.

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

 Shipped scalable MERN-stack web apps; optimized backend APIs, secure auth/payment flows, and relational + NoSQL schemas.

 Integrated RESTful AI microservices and managed CI/CD (Git, GitHub Actions) to cut release time by 25%.  Collaborated in Agile sprints, performing code reviews and mentoring junior interns on best practices.

# 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) ( [Deep Learning Specialization Certificate](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/))

# SKILLS

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

**AI/ML s Deep Learning:** PyTorch | TensorFlow | Keras | Stable Diffusion | CNN | RNN | LSTM | Transformers | LLM fine-tuning | Hugging Face | scikit-learn

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

**Frontend s Databases:** Streamlit | React.js | Node.js | Express.js | MongoDB | PostgreSQL | Redis