

Tushar Soni

AIML Engineer

+91 892-077-0370 | tusharsoni.info@gmail.com | GitHub [🔗](#) | LinkedIn [🔗](#) | Website [🔗](#)

Summary	B.Tech. final year AI/ML specialist with hands-on LLM, RAG, fine-tuning, and AI app dev experience. Strong OOP & algorithms foundation. Production AI systems expert in retrieval optimization, post-training, quantization, local inference.	
Education	Guru Gobind Singh Indraprastha University (GGSIPU) New Delhi, Delhi	Sep 2023 - Jul 2027 (Expected) B.Tech (AIML)
Skills	AI/ML/DL: PyTorch, Tensorflow, scikit-learn, Transformers, LLMs, PEFT, Agentic AI, RAG, ONNX Frameworks & Libraries: FastAPI, NumPy, LangChain, LangGraph, Unsloth, CrewAI Databases & Retrieval: MongoDB, MySQL, FAISS, ChromaDB, Redis Cloud: AWS, GCP DevOps: Git, Github Actions (CI/CD), Docker, Kubernetes Languages: Python, C, JavaScript, TypeScript, HTML, CSS Operating Systems: Linux (Ubuntu, Fedora), Windows	
Experience	LearnNex Powered by Wipro Dice ID Data Science Intern	Oct 2025 - Dec 2025
	<ul style="list-style-type: none">Analyzed customer behavior and churn patterns, identifying high-risk segments and key churn drivers to inform targeted retention strategies.Established Logistic Regression as the baseline model and improved predictive performance by 11.5%, achieving 87% accuracy through feature engineering and model optimization.Engineered features from tenure, service subscriptions, and payment behavior, improving model input quality and enabling more effective identification of high-risk customers.Built and evaluated Artificial Neural Network (ANN), Logistic Regression, and Random Forest models, comparing classification performance to identify the best-performing approach.	
Publications & Articles	Tushar Soni , GRAG: Graded Retrieval-Augmented Generation for Precise Chunk Retrieval (Under review at TMLR) Simple to Advance RNN , Jan 23, 2026 🔗 Attention is Good , Apr 21, 2026, 🔗	
Projects	ManuIndex: Document-Aware RAG Retrieval System 🔗 <ul style="list-style-type: none">Built a document-aware retrieval system that uses three-layered architecture to route queries to the most relevant documents, achieving a 0.7332 faithfulness score and a 1.958s average end-to-end time in benchmark tests.Implemented hybrid search combining dense vector and sparse text indexes, outperforming standard RAG methods on 2 datasets with a 0.5617 context F1@3 score.Deployed a practical RAG pipeline for policies, reports, and contracts using ONNX Runtime, beating 5 RAG variants on benchmark evaluation suites. n99: Intelligent Movie Ticket Availability Tracker 🔗 <ul style="list-style-type: none">Built a two-tier architecture using Discord.py (bot) and APScheduler (monitoring service), where the bot handles user interactions instantly while the monitoring service runs in the background to check cinema websites.Implemented Redis session management and PostgreSQL advisory locking to prevent duplicate watch creation and ensure atomic data operations.Integrated Playwright browser automation pipeline to scrape movie availability.Containerized Deployment on AWS EKS (EC2-based) with Docker, automatic failover, and horizontal scaling capabilities.	
Achievements	Top 10 Finalist - SSCBS ACE(M) CodeFest 2.0 (2026): Alignify: Employee Engagement AI Platform Runner Up - SIH (Internal) 2025: AI-Driven Public Health Chatbot for Disease Awareness	