Gourav Shokeen

+91 9213395804 | gouravshokeen41@gmail.com | linkedin.com/in/gourav-shokeen | github.com/gourav-shokeen | gouravshokeen.me | Greater Delhi Area, India

SUMMARY

B.Tech CSE (AI & ML) student at SGT University (GPA 8.8/10) who designs, builds, and ships production machine learning and Generative AI systems end-to-end. Built **Katha-LM**, a Hindi transformer LM trained from scratch (5.19M params, val perplexity 2.51), and **Artha**, a deployed Hinglish sentiment-analysis API (Macro F1 0.9304); co-founder & lead developer of **ReVive**, an AI physiotherapy platform incubated under IHFC, IIT Delhi. Strong in deep learning, retrieval-augmented generation (RAG), Large Language Model (LLM) fine-tuning & integration, agentic workflow automation, and model deployment (FastAPI/Docker). Seeking an AI Engineering internship to build and deploy LLM-powered products at scale.

EDUCATION

Shree Guru Gobind Singh Tricentenary (SGT) University *Aug 2024 - Aug 2028 · Gurugram, Haryana*
B.Tech — Computer Science & Engineering (AI–ML Specialization) · GPA: 8.8 / 10

PROJECTS

Katha-LM — Hindi Story Language Model (from scratch) *2026*
Python · PyTorch · Custom BPE Tokenizer · FastAPI · Docker · Next.js · github.com/gourav-shokeen/katha-lm
- Built a decoder-only transformer from scratch in PyTorch (bigram → MLP → transformer; val perplexity 26.8 → 16.3 → 13.1), then scaled a 5.19M-param config to val perplexity **2.51** on 100k Hindi stories (Kaggle T4).
- Engineered a custom Hindi BPE tokenizer (vocab 756) at **1.63 fertility vs GPT-2's 7.02** on Devanagari (~4× fewer tokens/word); shipped to production via FastAPI on HF Spaces (Docker) + Next.js 14 on Vercel.

Artha — Hinglish Sentiment Analysis API *2026*
Python · PyTorch · xlm-roberta · HuggingFace · FastAPI · Docker · Next.js · artha.social
- Scraped & processed 300k+ real Hinglish comments; applied domain-adaptive pretraining (DAPT) on xlm-roberta-base, then a 3-stage fine-tune (DAPT → v1 F1 0.7856 → neutral upsampling + lr 5e-6 → v2 **Macro F1 0.9304**).
- Deployed a production REST API: FastAPI on HF Spaces (Docker) + Next.js on Vercel, with bearer-token auth, model versioning, and private HF Hub model hosting.

RFQ-to-Quote Agent — B2B Sales Automation *2026*
n8n · Gemini (Flash-Lite / Flash) · LlamaParse · Playwright · Telegram · Gmail · Google Sheets
- Built an n8n agent that turns inbound RFQs (PDF / Excel / email) into branded PDF quotes end-to-end: LLM line-item extraction → in-context catalog matching & pricing over 75 SKUs → Playwright PDF → Telegram human-approval gate → in-thread Gmail reply + Google Sheets logging.
- Routed extraction/scoring to Gemini Flash-Lite and drafting to Flash; enforced strict structured-output schemas with confidence-based exception flagging so low-confidence matches surface for review instead of failing silently.

Medical Reference Agent — Self-Correcting Agentic RAG over Clinical Literature *2026*
Python · LangGraph · LangChain · Chroma · FastAPI · Next.js · github.com/gourav-shokeen/medical-rag-agent
- Built a self-correcting agentic RAG system (LangGraph state machine) over **491K** clinical references that cites sources, refuses out-of-scope questions, and rewrites its own weak queries; scored **62.2%** on the MIRAGE medical-QA benchmark with a 7-8B model (GPT-3.5 ref. 71.6%).
- Ran an embedding ablation (general vs. MedCPT) that exposed a retriever/reranker misalignment costing **4.4 pts** of accuracy; built a resumable RAGAS / DeepEval / Langfuse eval harness, embedding 491K snippets on a 6 GB GPU.

EXPERIENCE

ReVive — Co-Founder & Lead Developer *2026 - Present · IHFC, IIT Delhi*
AI physiotherapy platform incubated under IHFC, IIT Delhi · React Native · Node.js/Express · MongoDB · Redis · Socket.IO · WebRTC
- Co-founded and lead development of a cross-platform patient + therapist app: 7-step clinical onboarding, real-time booking against live availability, and in-session WebRTC video consultations over a Socket.IO signalling namespace.
- Engineered the booking engine with 30-minute slot generation and a 3-layer double-booking guard (Redis lock + DB uniqueness constraint + partial-index backstop).

SGT University — Tuskers Club · Event Coordinator *Feb 2026 - Apr 2026 · Gurugram*
- Planned and executed 5+ public-speaking & debate events end-to-end; managed logistics and multi-channel outreach.

SKILLS

Languages: Python, JavaScript, TypeScript, C, SQL, HTML, CSS
AI / ML: Machine Learning, Deep Learning, Generative AI, Large Language Models (LLMs), Neural Networks, Retrieval-Augmented Generation (RAG), Agentic RAG, LLM Integration (Groq/Llama), LLM Fine-Tuning, Prompt Engineering, Embeddings, Vector Databases, Cross-Encoder Reranking, Domain-Adaptive Pretraining (DAPT), Transformer Architectures, LLM Pretraining (from scratch), Tokenizer Design (BPE), MLOps / LLMOps, Model Deployment & Serving, NLP, Agentic Workflow Automation (n8n)
Frameworks: PyTorch, HuggingFace Transformers, LangChain, LangGraph, Sentence-Transformers, scikit-learn, Pandas, NumPy, spaCy, FastAPI, Next.js, React, React Native
Tools & Platforms: Git/GitHub, Docker, REST APIs, n8n, HuggingFace Spaces, Vercel, Chroma (vector DB), RAGAS, DeepEval, Langfuse, Ollama, Supabase, MongoDB, Redis, WebRTC, WebSockets, Kaggle

AWARDS & CERTIFICATIONS

- Winner — AI/ML Hackathon & Coding Competition, SGT University (organized by Samatrix Consulting), 2026.
- ACIC SGT University Funding — secured funding for Click2Drive (2025) · Top 1000, IIT Delhi College Youth Ideathon (2025).
- Certified: Master Python & AI — Data Generation, Predictive Modeling & Advanced Analytics (2025); Programming in C (2024).