

TUSHAR JAJU

+91-9927280188

| tusharbrisingr9802@gmail.com

| [LinkedIn](#)

| [GitHub](#)

| [Portfolio](#)

SUMMARY

ML Engineer with published research in cross-modal LLM adaptation (Springer, IMPACT 2025). Fine-tuned Mistral-7B with a 5-method curriculum framework achieving 58% video summarization improvement while preventing catastrophic forgetting. Built production pipelines on consumer GPUs: LoRA fine-tuning of 4B–7B models, custom datasets (4,300+ samples), 4-bit quantized inference under 16GB VRAM, and on-device LLM deployment on Android.

PUBLICATIONS & OPEN SOURCE

- T. Jaju et al., “Cross-Modal Transfer Learning in Domain-Adaptive Video Summarization,” *IMPACT 2025*, Springer (presented Dec 6, 2025; proceedings forthcoming)
- Fine-tuned models and adapters published on [HuggingFace](#) (MedScribe SOAP LoRA, Bawarchi recipe generator); curated datasets on [Kaggle](#) (712 clinical samples, 68MB recipe corpus)
- [llmclean](#) – zero-dependency Python library for cleaning and normalizing raw LLM output (PyPI); [GitHub](#)
- Submission to Google MedGemma Impact Challenge 2026 (Main Track + Novel Task Prize)
- Ranked 78th out of 1,100 applicants in AnaVerse 2.0, national hackathon hosted by Celebal Technologies

PROJECTS

Hybrid-Dataset Summariser – Cross-Modal LLM Adaptation

Nov 2025 – Apr 2026

Tech: PyTorch (nightly/cu128), Mistral-7B, PEFT, bitsandbytes, HDF5, SBERT, Whisper | [GitHub](#)

- **Problem & data:** Published finding (IMPACT 2025) that LoRA fine-tuning on academic papers degrades video summarization by 14–50%; curated 4,324-sample cross-modal dataset (2,368 arXiv papers + 738 YouTube lectures + 1,218 SBERT-mined pairs) with full pipeline: yt-dlp scraping, Whisper transcription, GPT-4o-mini summaries, SBERT pair mining, quality filtering, Mistral tokenization, HDF5 packaging
- **Framework:** 3-phase curriculum on RTX 5070 Ti (16GB): LoRA+ asymmetric LR ($\eta_B/\eta_A=8$), OPLoRA orthogonal projection ($k=16$), EWC Fisher regularization ($\lambda=200\rightarrow400$), contrastive alignment; 4-bit NF4 + gradient checkpointing + 8-bit AdamW
- **Results:** Video ROUGE-1 +58% (0.263 $\rightarrow$ 0.417), video ROUGE-2 +272%, passive voice contamination held at 14.1% vs. 31.4% catastrophic baseline, cross-modal consistency +49%, BERTScore flipped negative $\rightarrow$ positive; all results statistically significant ($p<0.005$, paired t-test with bootstrap CI)

Sakhi – Offline-First Voice-to-Form for Community Health Workers

Apr 2026 – May 2026

Tech: Python, FastAPI, Ollama, Gemma 4, Whisper, Cactus SDK, React | [GitHub](#)

- **System:** Offline-first Hindi voice-to-form for ASHA community health workers – home-visit speech becomes structured government health forms (5 NHM/IMNCI visit schemas) with real-time maternal/neonatal danger-sign flagging and referral decisions; Gemma 4 Good Hackathon entry (Health, Ollama, Unsloth, Cactus tracks)
- **Dual deployment:** Health-center mode (Whisper ASR + Gemma 4 E4B function-calling, $\sim 15\text{--}25\text{s/visit}$) and on-device Android field mode (Gemma 4 E2B INT4 via Cactus SDK, $\sim 5\text{min/visit}$, fully offline); six-layer anti-hallucination pipeline with verbatim transcript grounding and clinical range validation
- **Fine-tuning & eval:** LoRA fine-tuned on 1,154 synthetic ASHA conversations; 13/15 (87%) end-to-end audio pipeline, 15/15 text extraction, 133/133 Hindi numeral normalization; validated on real Hindi voice (4 recordings, 2 speakers); live demo on HuggingFace Spaces

MedScribe – Clinical Documentation Workstation

Jan 2026 – Feb 2026

Tech: MedGemma 1.5 4B, MedASR Conformer 105M, LoRA, Gradio, CUDA | [GitHub](#) [HF](#) [Kaggle](#)

- **Pipeline:** MedASR (105M Conformer) $\rightarrow$ MedGemma 1.5 4B + 4.2M-param LoRA $\rightarrow$ structured SOAP note + 6 clinical inference tools; LoRA fine-tuned on 712 GPT-4o-curated samples for \$1.28; 4-bit NF4, fully offline (no patient data egress)
- **Impact:** 90/100 quality score, 0% hallucinated findings, 100% section completeness (S/O/A/P), 46% shorter notes vs. base MedGemma; ~ 11 min/encounter saved ($\sim 1,100$ additional patient visits per physician annually)

EXPERIENCE

Systems Development Intern – Cognavi India

Mar 2026 – Present

- **CRM & product:** Maintain and extend features on the company CRM; data cleaning and quality control keeping operational records reliable across the team
- **Data pipelines:** Built Python data-extraction and enrichment pipelines feeding operational databases; automated previously manual record-keeping workflows
- **LLM integration:** Integrated external APIs (Anthropic, Tavily) into internal LLM workflows; internal platform tooling for operational stability

TECHNICAL SKILLS

- **LLM Fine-tuning:** LoRA, EWC, PEFT, curriculum learning, quantization, gradient checkpointing
- **Frameworks:** PyTorch, Transformers, bitsandbytes, FastAPI
- **Data Engineering:** Data cleaning, enrichment, extraction & ETL pipelines, API integration, SQL
- **Deployment & Tooling:** HuggingFace Hub, Kaggle, Gradio, Streamlit, Cactus SDK, Git, Python

EDUCATION

B.Tech, Computer Science and Engineering

ABES Engineering College, Ghaziabad (AKTU)