Tashon Braganca

AI Engineer | LLMs | RAG | Agents | Available November 2026
Bengaluru, India | +91 93222 48585 | tashon.braganca@gmail.com
linkedin.com/in/tashon-braganca-982784197 | github.com/TashonBraganca | tashonbraganca.vercel.app

SUMMARY

AI engineer with 14 months of shipping LLM and vision-language systems across 3 companies. Shipped 2 production LLM products as the sole engineer, raised RAG accuracy from 58% to 91%, cut LLM cost by 83%, and took Gemma and MiniCPM-V decoding from 30 to 130 tokens/s.

EXPERIENCE

Propheus Bengaluru, India
Machine Learning Engineer Intern May 2026 - Present

- Deployed a cable-tree contact detector on NVIDIA Jetson Orin across 30 km of Bengaluru roads, at **96% precision** and 73% recall on 500 blind-checked crops. SAM3 masks feed a ConvNeXt-Tiny classifier I trained on 497 images I labelled.
- Beat SAM3 on cable-contact detection, **0.757 vs 0.655 AUC**, after training and benchmarking RF-DETR, YOLO and MM-Grounding-DINO on 249 hand-labelled frames.
- Fine-tuned MiniCPM-V with LoRA to read shop signboards, lifting exact-match accuracy from 43.8% to 61.5% on 96 hard crops, a **17.7-point gain**.
- Cut wire segmentation from 64.4 s to 0.45 s per frame, **145x faster**, by distilling a 12-pass SAM3 ensemble into a 27.4M-parameter segmenter that matched its coverage.
- Raised Gemma and MiniCPM-V decoding from 30 to **130 tokens/s** with CUDA graphs, multi-token prediction and 4-bit quantization, and made 4 small VLMs 3 to 8x faster with llama.cpp.

Nelo Bengaluru, India
AI Engineer Intern Jan 2026 - May 2026

- Raised RAG answer accuracy from **58% to 91%** on a 40-question test set by tuning chunking, overlap and retrieval prompts over LlamaIndex and Pinecone.
- Replaced a manual support workflow with a **3-agent LangGraph system** for search, escalation and fulfilment, live on 5 WhatsApp flows via Meta Cloud API and Zoho CRM.

Penzo Technologies Remote
AI Engineer, sole engineer on Graho Sep 2025 - Dec 2025

- Shipped Graho, a paid AI report product, **alone in 8 weeks** with Next.js, FastAPI and an LLM layer on GCP Cloud Run.
- Cut LLM cost per conversation by **83%**, Rs 0.18 to Rs 0.03, with prompt templates, token budgets and Redis caching. Reports cost Rs 0.50 to generate and sell for Rs 10.

PROJECTS

Forge AI Studio | github.com/TashonBraganca/Forge Apr 2026

- Local LLM fine-tuning studio that goes from git clone to a running LoRA job in **under 2 minutes**, driving LLaMA-Factory and Unsloth. Fine-tuned 3 models on consumer GPUs, including two 2B models.

MediLens | github.com/TashonBraganca/Medicio-Ai Sep 2025

- Fully offline clinical triage. LLaVA reads medical images, Tesseract reads lab reports, and the app returns urgency scores at **75% accuracy** on 100+ cases, running on a laptop CPU.

CandidRank | candidrank.cc Jan 2026 - Present

- Solo-built resume screening SaaS with pgvector ranking against JDs, GPT-4o structured scoring and OCR fallback. Credit billing, auth and **137 production deploys** in 90 days.

SKILLS

ML & Computer Vision: Python, PyTorch, TensorFlow, OpenCV, Scikit-learn, NumPy, Pandas, Deep Learning, Hugging Face, YOLO, RF-DETR, SAM3, ConvNeXt, Object Detection, Segmentation, Image Processing, Data Annotation, Knowledge Distillation, Model Evaluation

Generative AI: LLMs, RAG, LangGraph, LangChain, LlamaIndex, Pinecone, pgvector, Multi-Agent Systems, MiniCPM-V, LLaVA, LoRA, QLoRA, Unsloth, vLLM, llama.cpp

MLOps & Infra: NVIDIA Jetson, Edge Deployment, Docker, FastAPI, GitHub Actions CI/CD, Weights & Biases, GCP Cloud Run, AWS, PostgreSQL, Redis, Next.js, TypeScript, SQL

EDUCATION

MVJ College of Engineering, B.E. Data Science, CGPA 8.5 2026
Supervised Machine Learning, Stanford and DeepLearning.AI (2025) | SYNTHACK Hackathon, 3rd Place (2025)