Shubham Gajjar

Portland, Maine • gajjar.shu@northeastern.edu
shubhamgajjar.dev • linkedin.com/in/shubhamgajjar • github.com/ShubhamGajjar • ORCID 0009-0007-4654-2525

Education

Northeastern University, MS in Artificial Intelligence – Portland, Maine — Expected May 2027

- GPA: 4.0/4.0
- Research capstone: MorphoCLIP, text-supervised contrastive learning for cell microscopy

LDRP Institute of Technology and Research, BE in Computer Engineering – Gandhinagar, India — May 2025

- GPA: 8.41/10.0

V.P.M.P. Polytechnic, Diploma in Computer Engineering – Gandhinagar, India — May 2022

- GPA: 9.22/10.0

Research Experience

Northeastern University, Roux Institute, Research Assistant, Vision-Language Models in Biomedicine – Portland, Maine — Feb 2026 - Present

- Selected and set up MedGemma 1.5 4B, LLaVA-Med and Gemma 4 31B as candidate base models in one fine-tuning harness, so they could be compared on the same footing
- Fine-tuned MedGemma 1.5 4B with 4-bit QLoRA on an NVIDIA H100, keeping the base model's reasoning intact while constraining it to the strict output format the task needs
- Designed a five-type visual question answering schema that turns unstructured expert annotations into supervised training data across ∼2M microscopy tiles covering 23 cell types
- Built the evaluation framework that drove checkpoint selection, reaching case-level F1 of 0.83 at recall 0.92, and traced the model's over-detection to a shift in the training distribution rather than reporting the number
- Ran a linear-probe sweep that set a frozen-encoder ceiling of 0.91, separating what the model could not perceive from what it had merely miscalibrated
- Chose the input tiling scheme from the same sweep, worth 8 points of per-cell accuracy over whole-field downscaling
- Implemented Group Relative Policy Optimization from scratch in PyTorch, with group-normalized advantages, KL-anchoring to the supervised policy and a verifiable reward computed against ground-truth cell counts
- Built distributed PySpark curation with Spark SQL, served through an LRU-cached prefetcher that removed per-step I/O, and tracked every run in MLflow against a versioned model registry
- Presented the findings to the project's academic and industry partners

Publications

S. Gajjar, D. Joshi, A. Poptani and V. Barot. *VGG16-MCA UNet: Whole-Tumor Segmentation in 2D FLAIR MRI with Decoder-Side Channel Attention*. arXiv preprint arXiv:2609.23919. — Sep 2026

- Designed the Multi-Channel Attention module through a systematic study of ten architecture and loss combinations, trained with a Focal Tversky loss for the heavy imbalance between tumor and background
- Reached a pooled pixel-level Dice of 95.10% on a held-out BraTS 2020 split, averaged the five fold models' weights into one network, and released the code and exact data splits

S. Ilyosbekov, **S. Gajjar** and R. Jin. *MorphoCLIP: Text-Supervised Contrastive Learning for Perturbation Matching in Cell Painting Images*. arXiv preprint arXiv:2608.22690. — Aug 2026

- Built the text pipeline solo, from prompt construction to a frozen BioClinical ModernBERT with cached outputs, plus the CellCLIP baseline, the evaluation and inference scripts, and the multi-GPU DDP training infrastructure
- The model reaches 37-41% image-to-text Recall@10 on held-out CPJUMP1 against ∼10% chance, and the paper reports its negative results too

S. Gajjar et al. *A Hybrid ResNet-ViT Architecture for Skin Cancer Classification*. IEEE World Conference on Applied Intelligence and Computing (AIC 2025). — Jul 2025

- Designed the architecture and wrote the paper: a frozen ResNet50 fused with an eight-head Vision Transformer block, reaching 96.3% accuracy across seven HAM10000 classes
- Released it as an open-source package with training and evaluation scripts, fixed-seed reproducibility and pretrained weights on Hugging Face

Industry Experience

BigCircle (UPSAAS Technologies LLP), AI Engineer – Gandhinagar, India Mar 2025 - Jul 2025

- Found the bottleneck in an AI deep-research platform unprompted: redundant back-and-forth API calls between its quality-evaluation agents. Re-architected the workflow, cutting report generation from 20 minutes to 5
- Traced severe server timeouts in a client-facing web app used daily by thousands of users to a 20,000-row table loading at once, and built paging that cut login time from 3 minutes to 2 seconds
- Found and fixed a data-integrity bug where lists marked hidden were still visible to users, exposing client data

BigCircle (UPSAAS Technologies LLP), Developer Intern – Gandhinagar, India Jan 2025 - Feb 2025

- Prototyped the Python tooling and LLM research workflows that became the Deep Research platform, and was promoted to AI Engineer

Projects

ENCOM: Efficient Neural Conversation Model (2-person team) 2026

- Replaced the speech-to-text stage of a conversational pipeline with a direct audio encoder built on Mamba state-space models, and ran a controlled comparison against a Whisper-based pipeline with the same language model to isolate the encoder's effect on latency
- Structured training in three phases across multiple GPUs, encoder pretraining, then adapter alignment, then speech decoder, with ONNX and CoreML export

Reinforcement Learning Agent for TrackMania (personal research prototype) Dec 2024

- Built an agent that learns to drive from the screen, the same input a human gets: real-time capture fused with live telemetry, Implicit Quantile Networks in PyTorch, and a virtual-checkpoint reward for dense feedback, reaching about 85% track completion so far

Extended ResNet50: Inverse Soft Mask Attention for Skin Cancer Classification (research project, unpublished manuscript) May 2025

- Designed an attention mechanism that hands the hair mask to the classifier as a second input instead of erasing the hair, so the model learns to read around the occlusion, reaching 97.89% accuracy on HAM10000

GMSM-USA Community Platform (sole developer) Jun 2026 - Present

- Built and maintain a platform used by 340 members across 45 states, with access control enforced in the database through 47 row-level security policies and append-only audit logging

Next Step: School Event Management Platform (6-person team) Jan 2026 - May 2026

- Owned authentication and role-based access across the stack (React Native, FastAPI, PostgreSQL), and led the frontend-to-backend integration

Technical Skills

Languages: Python, JavaScript, TypeScript, SQL

Deep Learning: PyTorch, TensorFlow, Keras, HuggingFace Transformers, QLoRA, GRPO, multi-GPU DDP, CUDA

Models and Methods: vision-language models (MedGemma, LLaVA-Med, SigLIP, DINOv3, CLIP), contrastive learning, semantic segmentation, attention mechanisms, distributional RL (IQN)

Data and MLOps: PySpark, Spark SQL, MLflow, TensorBoard, Databricks, Docker, HPC/Slurm, ONNX, CoreML

Web: React, Next.js, FastAPI, PostgreSQL, Supabase