Sahaj Kishorkumar Mistry

+91 63533 73133 | sahajmistry005@gmail.com | Website | LinkedIn | GitHub | Google Scholar

Education

Indian Institute of Technology Delhi **2026 – 2028**
Master of Technology in Machine Intelligence and Data Science *New Delhi, India*

Indian Institute of Technology Jammu **2020 – 2024**
Bachelor of Technology in Computer Science and Engineering; CGPA: 8.91/10 *Jammu, India*

Research Experience

IIT Delhi (Advisor: Prof. Sayan Ranu) **Jun 2026 – Sep 2026**
Graduate Researcher – Lossless Speculative Acceleration of Multi-Agent LLM Systems *Submitted to ICLR 2027*

- Formalized lossless speculative execution for multi-agent LLM systems, with a latency model splitting the speculation budget between *which* agents run next and *what* they receive.
- Cast next-agent prediction as **link prediction** over an agent graph mined from past traces using a lightweight **GRU**; boosted Speculator LLM message accuracy with **retrieval-augmented in-context learning**.
- On 4 multi-agent systems and 6 Qwen/Gemma/DeepSeek model pairs, cut wall-clock time by **17–38%** with identical outputs, beating SOTA by **6–28%** at lower token and FLOP cost.

University of Pennsylvania (Advisor: Prof. Sanath Chandra Guntuku) **May 2023 – Aug 2023**
Researcher – Brain Tumor Segmentation *Published at MICCAI 2023*

- Proposed a dual-encoder **nnU-Net** with cross-attention and a novel **ODConv3D** (3D dynamic convolution) layer.
- Ranked **1st** in the BraTS-Africa Challenge (validation) with Dice of 0.835 / 0.848 / 0.756 (ET / TC / WT).

KAUST (Remote) **May 2023 – Jul 2023**
Research Intern – 360° Panoramic Depth Estimation

- Adapted ZoeDepth to 360° panoramas with ODConv, **LoRA** and a perceptual loss; **0.250 RMSE**, beating prior SOTA.

Work Experience

Samsung R&D Institute **Sep 2025 – Jul 2026**
R&D Engineer *Noida, India*

- Built on-device image captioning (**ViT + GPT-2**) on Snapdragon 8 Gen 3 for visually impaired users.
- **INT8 quantization** cut memory **4**× (1 GB → 250 MB) and latency **3.3**× (500 → 150 ms) at <2% accuracy drop.

Miko **Jun 2024 – Sep 2025**
Applied Scientist *Mumbai, India*

- Built an LLM story-generation pipeline with a **content-safety** guardrail module, serving **500+ rpm** peak.
- Developed a face-swapping model using controlled image-guided **diffusion inpainting** for personalized stories.

Competitions

Amazon ML Challenge 2026 – Business Entity Resolution | **Top 10 of 32,000+ teams** **Sep 2026 – Oct 2026**

- Built in 3 days a pipeline linking **9.97M** noisy multilingual records to 1.73M entities (**0.9909 F0.5**, public LB).
- Fine-tuned an **e5 bi-encoder** + **FAISS HNSW** for blocking (99.4% recall at 1.28 candidates/record), then **cross-encoders**, two-stage **LightGBM** and an **LLM judge**; **self-trained** on pseudo-labels for an unlabeled country.

Publications

ICLR 2027 (under review) – ProSpect: Proactive Speculation for Lossless Acceleration of Multi-Agent Systems

MICCAI 2023 – Multiscale Encoder and ODConv Enrichment in nnU-Net for Brain Tumor Segmentation

IEEE ICASSP 2023 – Drone-vs-Bird: Drone Detection Using YOLOv7 with CSRT Tracker *(2nd place)*

CVPRW 2023, 2024 – NTIRE Challenges: Stereo Image SR; HR Depth of Specular & Transparent Surfaces *(7th place)*

Skills & Service

Research Areas: LLM Agents & Multi-Agent Systems, Graph Learning, Retrieval, Computer Vision, Efficient ML
Tools: Python, C/C++, PyTorch, Hugging Face Transformers, FAISS, LightGBM, LangGraph, LoRA, INT8 Quantization
Service: Reviewer, MICCAI BrainLesion Workshop 2023; Speaker, *"Demystifying Neural Networks"* (DeepLearning.AI)