

Arush Sharma

+91 92058 06901 | arushsharma536@gmail.com | New Delhi, India
linkedin.com/in/arush-sharma-44a796288 | github.com/Arush777



Education

IIT (ISM) Dhanbad

B.Tech, Electronics and Communication Engineering
(CGPA: 8.19/10.00)

Expected April 2027

Dhanbad, India

Technical Experience

Conscious Engines (Stackr Labs India Pvt. Ltd.) | Research Engineer Intern | backed by a16z

Aug 2026 – Present

- Draft-aligned decoding for **SDAR-30B-A3B** (diffusion MoE) using a dense 1.7B SDAR model which acts like a drafter which pre-fills each 32-token block and a self-distilled LoRA judge (**26M** params, attention layers) commits or re-masks each token.
- Later adding a **35M** LoRA (attention + MLP) that aligns drafter output with 30B MoE model. Results of the pipeline Vs base model across several benchmarks, reducing the SDAR-30B-A3B passes per block **12.8 to 7.6**, **1.35 $\times$** increase in token/s, with comparable accuracy.

IBM Research | Research Intern, AI for Quantum | Dr. Anupama Ray

May 2026 – Aug 2026

- Built a CliffSearch-style multi-agent algorithm-discovery loop in which open- and closed-model agents proposed, mutated, critiqued, and evaluated analytic QAOA schedule families across iterative generations.
- Evolved the linear-ramp baseline into the fixed power-law schedule, then carried it from plain QAOA into LR-QAOA/RQAOA experiments and combined it with sign-hedged beam recursion for simulator, MPS, and IBM hardware evaluation.

Aeroqube Inc. | Agentic AI Intern | [Offer letter](#)

Jun 2025 – Nov 2025

- Fine-tuned **MedGemma 4B** for a medical-tourism assistant and deployed a Dockerized AWS pipeline with RAG, LangSmith traces/evaluations, and repeatable prompt-model comparisons.
- Leveraged Gemini API to generate health itineraries for our customers, and using vercel to create a frontend and utilizing langraph for RAG to ground it with context that we got from **MedGemma 4B**.

Research Papers

From Points to Distributions: Probabilistic Latent

Autoregression for Medical Image Representation Learning | [paper](#)

Under review at WACV 2027

- Deterministic medical SSL collapses ambiguous anatomy into overconfident point embeddings; PACL instead predicts per-patch Gaussian latents and transfers uncertainty into classification and segmentation.
- As first author, formulated the mean/variance and anti-collapse losses, reaching **0.987** CXR AUROC, **0.938** BUSI AUROC, and **0.604** IoU (+2%/+2.1%/+1.3 pts over prior SOTA) under Dr. Soumen Basu (Meta) and Dr. Manoj Kumar (IIT Roorkee).

BrainRMA: ROI-Preserving Multiscale Attention for Brain Network Analysis | [paper](#)

Under review at AAAI 2027

- Brain-graph models either miss long-range dependencies or erase ROI identity through pooling; BrainRMA keeps ROI tokens while fusing global attention with reliability-weighted ROI-to-cluster context.
- As second author, ran the ablation suite and co-wrote methodology/experiments, reaching **82.35%** ABIDE and **93.75%** OSF accuracy (+3.35/+6.25 pts over SOTA) under Dr. Soumen Basu (Meta) and Dr. Manoj Kumar (IIT Roorkee).

Sign-hedged Beam QAOA: Scaling Quantum Optimization Algorithms

Under preparation for publication

- Developed sign-hedged beam RQAOA by pairing a CliffSearch-evolved fixed power-law schedule with width-B sign hedging to prevent irreversible correlator-sign errors. Resulting research is under preparation for publication.
- Improved SK ± 1 ($n=20$) AR **2.1 $\times$** to **0.998** in simulation and **1.000** on *IBM Marrakesh*; scaled MaxCut to $n=60$ at **0.984** (+17.6% vs. QAOA) and portfolio optimization to **0.983** on 20 assets (+26% vs. RQAOA) and **1.000** on 16-asset hardware. Won the **Best Paper Award** at the **IBM Research Internship Symposium 2026**.

Projects

PriorityKV: Structure-Aware KV-Cache Eviction | PyTorch, Transformers, H200, BF16/INT4 | [GitHub](#)

- Designed a role-aware eviction policy that protects system, tool-schema, constraint, sink, and recent tokens, then falls back toward attention when protected context exceeds the keep budget. Across **15,360 generations** on Qwen3-8B and Llama-3.1-8B, it retained task state in **112/120** cases versus **1/120** for position-blind eviction; H200 BF16/INT4 profiling cut KV payload **28%** and peak CUDA memory **13%**, while exposing an **11–12%** latency overhead.

AMLC 2026: Business Entity Resolution | multilingual-e5, bge-reranker-v2-m3, Qwen3.5-4B LoRA, LightGBM | [GitHub](#)

- Built a cost-tiered matching cascade over **1.7M** business entities (US/India/France) under macro $F_{0.5}$: a fine-tuned **multilingual-e5** bi-encoder blocks to top-20 candidates (**99.5%** recall), then four cross-encoders feed a **LightGBM** matcher that confidently resolves **93%** of cases; scored **0.99151** on the public leaderboard, **rank 5 of 80,000** teams.
- Escalated only the **6%** slice the cross-encoders left uncertain to an ensemble of **six Qwen3.5-4B LoRA judges** fused by a cross-fitted LightGBM combiner with edit-signature features (house-number offsets) that separate typos from planted lookalikes, lifting held-out $F_{0.5}$ **0.9906 to 0.9928**; full pipeline **5.09B** params under the 8B limit.

Inter IIT 14.0: Arista Intelligent RRM | GNN, Conservative Q-Learning, SDR, Linux networking | [GitHub](#)

- Architected a three-loop RRM system that converts SDR measurements into interference graphs and uses a GNN state encoder with offline CQL to tune channel, power, bandwidth, and OBSS-PD on Linux virtual APs. At convergence, raised stable edge-client throughput $\sim$ **18%** and composite QoE **+0.007** (40% throughput, 30% retries, 30% RSSI), with **zero configuration churn** and **P95 retries <8%**; ranked **12th/23 IITs**.

Technical Skills

Programming: Python, C/C++, SQL, JavaScript/TypeScript

ML: PyTorch, Transformers, TRL/PEFT, OpenRLHF, vLLM, PyTorch Geometric, scikit-learn

Research & Systems: SFT, DPO, GRPO/RLVR, LoRA/QLoRA, FlashAttention/Triton, DDP/FSDP, KV-cache optimization, quantization, ablations, OOD evaluation, CUDA, LSF, Docker, AWS

Achievements: **Best Research Paper Award Winner** (IBM Research India) | **Hackfest 2nd Prize** (AtomMail)

AIR 5181, JEE Advanced (top 0.47% of 1.11M+) | Selected for **Amazon ML Summer School** (top 3,000/134,000).