

Anupam Rawat

📍 Mumbai, India

✉ anupam.rawat.iitb@gmail.com

🔗 NoviceCoderInfinity

Education

Indian Institute of Technology Bombay (IIT Bombay), Dual Degree (B.Tech + M.Tech)
Electrical Engineering, Minor in Computer Science, CPI 8.16/10

Jul 2022 – Jul 2027

Publications

COSMOS: Soft Mechanism Mixtures with Verifiable Routing for Long-Horizon PDE Forecasting

Dec 2026

Anupam Rawat, Manikandan Padmanaban, Jagabondhu Hazra – *NeurIPS 2026 Workshop on AI for Stochastic Dynamics (STODY)*

Research Experience

IBM Research, AI Research Intern (Advisors - Manikandan Padmanaban, Jagabondhu Hazra)

May 2026 – Jul 2026

- Proposed COSMOS, a soft-routing Mixture-of-Experts neural operator for long-horizon PDE forecasting that keeps four physics specialists active at every step, accepted as a first-author paper at the NeurIPS 2026 STODY workshop.
- Cut long-horizon rollout error $2.4\times$ below FNO on held-out PDE mixtures, also beating GINO and MoE operators.
- Showed through ablations that soft routing beats hard top-1 routing and that the learned router is the main bottleneck.

Conscious Engines (Stackr Labs), Research Engineer Intern, Remote

Aug 2026 – present

- Built draft-and-judge decoding for the SDAR-30B diffusion MoE, where a 1.7B drafter fills each 32-token block and a judge adapter with **26.7M** parameters (0.09% of the model), trained on 128M tokens, decides which drafted tokens the 30B keeps.
- Achieved a $1.28\times$ end-to-end speedup and $1.60\times$ fewer 30B passes, with accuracy within noise on 6 of 7 benchmarks.

VID-CURATOR, Long-Video QA, Targeting WACV 2027 (Guide - Prof. Biplab Banerjee)

May 2026 – Aug 2026

- Designed a budget-bounded probe, retrieve and curate loop for long-video QA that reaches **0.647** accuracy on NEXT-GQA with only 8 sampled frames, beating uniform sampling at 16 frames (**0.493**) while using half the frame budget.
- Trained a LoRA-GRPO frame curator with a correctness-gated reward that helps coverage questions at 2–3 frames.
- Found that temporal localization helps short-span questions but hurts global-coverage ones under a matched VLM.

Open-Vocabulary Multi-Modal Transfer, R&D Project, IIT Bombay (Guide - Prof. Biplab Banerjee)

Jan 2025 – Apr 2025

- Implemented query-key similarity attention on a ViT backbone for zero-shot image-text alignment on unseen classes.
- Improved ranking on unseen remote-sensing classes over contrastive baselines when transferring from web images.

Selected Projects

Amazon ML Challenge 2026, Business Entity Resolution, Team Clauders

Sep 2026

- Built a retrieval and rerank pipeline with LoRA-tuned Qwen3.5-4B judges and LightGBM combiners over 1.7M entities.
- Placed **5th on the public leaderboard** by resolving lookalike business records across the US, India and France.

AlignmentLab, RLVR Reward Controls

Jul 2026 – present

- Ran matched-compute SFT, DPO and GRPO arms on H200s using correctness, random and format-only reward signals.
- Showed that correctness-blind rewards cut GSM8K exact match from **0.861 to 0.50–0.56**, exposing reward misspecification.

PriorityKV, Structure-Aware KV-Cache Eviction

Jul 2026

- Designed role-aware KV-cache eviction that protects system, tool-schema and constraint tokens under tight cache budgets.
- Retained task state in **112/120** long-context decodes versus **1/120** for position-blind eviction on Qwen3-8B and Llama-3.1-8B.

Linguistically Guided Multi-Token Prediction, CS 769 Course Project

Aug 2026 – present

- Testing whether multi-token prediction helps byte-level but hurts subword LMs on matched English and Hindi data.
- Designing boundary-aligned MTP with a variable word-group horizon to separate tokenizer effects from objective effects.

Industry Experience

Wyzr Fintech, Data Science Intern, Remote

May 2025 – Jan 2026

- Raised reconciliation match rate by **33%** with rule-based pre-filters that call the LLM only when deterministic rules fail.
- Added a **250-example** Langfuse eval suite to the PR path, which caught categorization regressions before each release.

Whettle, Technical Lead (Intern), Remote

Jan 2026 – May 2026

- Built a hotel RAG concierge with intent routing, tool calling and human handoff for live guest and staff messaging.
- Built an automated incident resolver that diagnoses live production issues and opens fix PRs that ship straight to staging.

Achievements and Teaching

- KVPY Fellowship (**top 1000** in India, 2021), PRMO qualifier twice, Indian Astronomy Olympiad Stage 2 (**top 255 of 10,000+**).
- Teaching Assistant for Intro to AI/ML at IIT Bombay since Jul 2026, running weekly labs, vivas and grading for **90+** students.

Skills

ML: PyTorch, Transformers, PEFT/LoRA, TRL, OpenRLHF, vLLM, Ray, DDP/torchrun, Qwen2.5-VL, SigLIP

Systems: Python, C++, HPC/LSF, FastAPI, Postgres, Next.js, Langfuse, wandb, Git, Linux