Pritish Saha

6290609505 | pritish171@gmail.com | Linkedin | Github | Scholar

Education

Indian Institute of Technology, Kharagpur — **Kharagpur, West Bengal**
Dual Degree in Manufacturing Science and Engineering — *Nov'22 – Jun'27*

Internships

Research Fellow (MARS 4.0) | *Cambridge AI Safety Hub* | *Cambridge, UK (Hybrid)* | *[Link]* — ***Dec'25 - Present***
Under Prof. Fernando Rosas, studying Bayesian mixed-state geometry and multi-timescale latent inference in transformers
- Devised hierarchical HMM/transducer pipelines with slow hidden drivers, fast dynamics, and exact 6-state Bayesian filters
- Probed 4L causal transformers; linearly decoded full 6D Bayesian posterior from residual stream (R^2=0.982, MSE=0.0043)
- Quantified layerwise timescales, shuffle/untrained/temporal controls, and driver–transducer subspace separation (z=-4.69)

Research Intern | *Complex Networks Research Lab* | *IIT Kharagpur (On-site)* | *[Link] [LoR]* — ***Jul'25 - Present***
Under Prof. Pawan Goyal, developed LaViDA, a reward-gated latent auxiliary loss augmenting GRPO for a math LLM pipeline
- Revealed RL/SFT reach identical ceilings via opposite dynamics: distribution-shaping (n=8) vs mode-narrowing (greedy)
- Built VRAM-**optimized GRPO** loops (LoRA-r64, vLLM, flash-attn) scaling 16 rollouts on a single H100 at 36s/step
- Designed **Oracle-augmented** MSE alignment, achieving 79.57% n=8 mean correctness (**+4.70pp** over GRPO, p=.007)

Research Intern | *RAAPID INC* | *Remote* | *[Link] [LoR]* — ***Mar'25 - Present***
Under Prof. Amitava Das, architected GRIT from scratch, evaluated medical LLM safety, and actively implementing FROST
- Authored fused Triton kernels for covariance accumulation and GPU-side **Cholesky** inversions, bypassing GMEM writes
- Architected a **dual-stream async** CUDA pipeline for natural-gradient updates, masking latency across **60+ modules**
- Benchmarked MedGemma via CARES-18K (**86% ASR**) and engineered FROST token-level decoding for antidistillation

Publications

Memory Transformers | *ICLR 2026 NFAM Workshop* | *[Code]* | *[Paper]* — ***Mar'26 - Present***
Co-authored "Mixture of Chapters: Scaling Learnt Memory in Transformers" accepted at ICLR 2026 NFAM Workshop
- Proposed sparse learnable memory banks and cross-attention layers for latent storage and retrieval beyond RAG systems
- Developed MoE-inspired **chapter routing** to partition memory banks, enabling scaling to **262K** learned memory tokens
- Outperformed iso-FLOP vanilla transformers on pretraining and knowledge benchmarks, with lower forgetting during IFT

GRIT: Geometry-aware PEFT | *arXiv Preprint* | *[Code]* | *[Paper]* — ***Mar'25 - Present***
Sole first author of "GRIT: Geometry-aware PEFT via rank-space K-FAC, Fisher reprojection & dynamic rank adaptation"
- Constructed a **second-order** PEFT framework utilizing rank-space **natural gradients** to accelerate LLM **convergence**
- Calibrated adaptive **Fisher-spectrum** thresholding to **dynamically** adjust LoRA ranks and **optimize** parameter usage
- Yielded competitive performance to LoRA & QLoRA on generative & NLU tasks, reducing trainable params by 25-80%

Competitions

Runner-up | *General Championship Data Analytics, IIT Kharagpur* | *[Link]* — ***Mar'26***
Captained the team to build a full-stack GenAI analytics dashboard for Frammer AI, delivering insights via NLQ and KPI labs
- Orchestrated a NLQ system to drive dashboard insights & KPI analysis (LangGraph, self-healing SQL, 90% EX on BIRD)
- Augmented evaluation capabilities using Gaussian-anchored synthetic data generation to test frameworks on a star schema

Participant | *Amazon ML Challenge 2025* | *[Link]* — ***Oct'25***
Secured 40.8 SMAPE by stacking Qwen2.5-VL-3B SFT & LightGBM on CLIP/text features via a price-space meta-learner
- Eliminated I/O bottlenecks via **offline tensorization** & **WebDataset**, accelerating 4-bit QLoRA throughput by ~**75%**
- Enforced alignment via special-token masking & chat templates, using **Pseudo-Huber** loss with monotonic constraints

National Finalist | *Decision Science Track - The American Express Campus Challenge* | *[Link]* — ***Jul'25***
Achieved 0.59 MAP score on the final, unseen evaluation set with a 3-stage GBDT-Transformer ensemble for offer ranking
- Engineered **3k+** features with a **parallelized** pipeline, creating leakage-free profiles with advanced temporal metrics
- Trained a Transformer on **GBDT residuals** using a **listwise** ranking loss to correct the ensemble's systematic errors

Technical Skills

- **Languages/ Tools :** Python, C/C++, CUDA, Linux, Docker, Git, WebDataset, FastAPI, ChromaDB, LangGraph
- **Libraries :** PyTorch, JAX, Triton, Penzai, Transformers, FlashAttention-2, PEFT/LoRA, bitsandbytes, TRL

Key Courses Taken

- **University:** Safety Fundamentals of Generative AI | Operations Research | Probability & Statistics | Linear Algebra
- **MOOCs:** Stanford CS229 (ML) | Stanford CS230 (DL) | LLM Agents MOOC | Algozenith | Summer Analytics

Positions Of Responsibility

Senior AI Developer, KodeinKGP, IIT Kharagpur — ***Aug'23 - May'24***
- Conducted AI Team selections, led workshops, wrote Medium articles & facilitated the National Science Week hackathon

Achievements

Codeforces **Pupil** (Handle: **pritish171**) | Demonstrated Academic Excellence with **95.60%** in Grade 10 (ICSE) and **90.80%** in Grade 12 (CBSE) | Secured merit positions in JEE Main, JEE Advanced, and WBJEE

Extra-Curricular Activities

Represented Patel Hall in Football and Water Polo Teams for Interhall Sports General Championship | Karate black belt with participation in multiple tournaments | NSS volunteer, contributed to improving living conditions in nearby villages