Arnesh Banerjee

+91-8902037524 | arnesh.banerjee.ds27@heritageit.edu.in | website | arnesh24 | ArneshBanerjee

Education

Heritage Institute of Technology, Kolkata GPA: 8.876/10
B.Tech in Computer Science and Engineering (Data Science) *Aug. 2023 – June 2027*

Research Interests

Deep Learning · Large Language Models · AI Safety & Alignment · Multi-Agent Reinforcement Learning (MARL) · RL Environments & Simulation · Applied ML

Research Experience

IIT Kharagpur (GRISHMA Summer Internship 2026) May 2026 – Present
Advisor: Dr. Sourangshu Bhattacharya *On Site*

- Working on IgLM, India's first population-specific genomic foundational model (StripedHyena2 architecture), under the GRISHMA Summer Internship Program; within this effort, reached pooled XGBoost ROC-AUC **0.933** on single-sample oral-cavity cancer detection from RNA-seq (58,147 genes, 1,206 samples across 1,160 patients) by training an L1-embedded classifier family under patient-grouped 5-fold stratified cross-validation with predictions pooled out-of-fold.
- Verified the signal is biological — not tissue-of-origin — by designing a layered TCGA negative class (non-oral HNSC, solid-tissue normal, 8 unrelated cancers) and reporting per-stratum sub-AUCs of 0.800, 0.900, and 0.987; confirmed robustness to class skew via a 1.5:1 balanced re-run preserving the headline (XGBoost 0.928).
- Also ported the Cedars-Sinai Molecular-Twin (MTPilot) L1-embedded model family into a shared module reused across the IgLM detection and downstream survival-prediction pipelines.

Jadavpur University November 2025 – May 2026
Advisor: Prof. Debotosh Bhattacharjee *Remote*

- Designed a hybrid CNN–Transformer segmentation model (ResNet-34 encoder + ASPP + Transformer bottleneck + SE-gated skip connections) with a differentiable Chan–Vese level-set loss, reaching 0.9716 Dice and 0.9463 IoU on the DMR-IR dataset (357 thermograms, 119 patients) under patient-stratified 5-fold cross-validation.
- Surfaced an annotation-quality ceiling in weakly supervised thermography by benchmarking against four state-of-the-art baselines (Attention U-Net, UNet++, DeepLabV3+, TransUNet) on five metrics (Dice, IoU, HD95, ASSD, BF1) and showing all models converge to statistically indistinguishable Dice (≈ 0.97, $p > 0.05$, paired Wilcoxon with 1000-resample bootstrap CIs).
- Also built a robustness battery (label-noise injection at 10−−30%, augmentation regimes, 25−−100% training subsets) and an explainability suite (Grad-CAM, attention maps, Monte-Carlo dropout uncertainty) for clinician-facing decision support.

New Jersey Institute of Technology June 2025 – November 2025
Advisor: Dr. Arnob Ghosh *On Site / Virtual*

- Built a 2,500-pair safe/unsafe prompt–response dataset spanning jailbreak strategies, indirect requests, role-play, multi-step instructions, and ethical/unethical educational queries; assigned absolute binary harm labels (replacing the Bradley–Terry pairwise scheme) and fine-tuned the final six layers of LLaMA-2-7B-chat-hf with a dense classification head as the CS-RLHF cost model.
- Validated semantic grounding of the cost model on the held-out test split and the external XS-Test benchmark, reaching $\approx 92\%$ alignment with human safety judgments and XS-Test scores of 0.91–0.96 (matching human verdict 0.89–0.92), versus 0.07–0.32 for the Safe-RLHF baseline cost model.
- Demonstrated the trained policy is **8**× more efficient at flagging unsafe responses than Safe-RLHF and is preferred by humans in $\approx 60\%$ of head-to-head comparisons (+70 Elo) by sampling 1,000 prompts from the curated dataset for policy evaluation; co-authored the resulting COLM 2026 submission (arXiv:2510.03520).

Heritage Institute of Technology October 2024 – March 2025
Advisor: Ms. Arpita Talukdar *On Site*

- Improved WPBC accuracy to **93.67**% (SVM + RFE) and WDBC to **97.77**% (LogReg + RFE) by adding RFE/SFS feature selection, SMOTE class balancing, and GridSearchCV hyperparameter tuning over a dual-stage diagnosis-and-recurrence ML framework benchmarking five classifiers (RF, SVM, Logistic Regression, MLP, XGBoost) under stratified 10-fold cross-validation.
- Also identified clinically relevant nuclear features through a comparative analysis across model–feature-selection combinations, reported with bootstrap confidence intervals.

Publications

- Ayushi Bhattacharjee, **Arnesh Banerjee**, Arpita Talukdar. 2026. Recursive and Wrapper-Based Feature Selection for Breast Cancer Diagnosis and Prognosis. In the proceedings of the 4th Analytics Global Conference (AGC 2026), March 2026. **Oral.**

Pre-Prints

- Kartik Pandit, Sourav Ganguly, **Arnesh Banerjee**, Shaahin Angizi, Arnob Ghosh. 2025. Certifiable Safe RLHF: Semantic Grounding and Fixed Penalty Constraint Optimization for Safer LLM Alignment. *Under review, COLM 2026*; arXiv:2510.03520.
- **Arnesh Banerjee**, Debotosh Bhattacharjee. An Intelligent Weakly Supervised Framework for Breast Thermography Segmentation Using Hybrid CNN–Transformer Networks. *In preparation for Expert Systems with Applications.*

Independent & Ongoing Research

- **Arnesh Banerjee**. Co-evolutionary Multi Agent RL for Autonomous Drones. In collaboration with the AI for Defence Lab, ULiège, Belgium.
- Kartik Pandit, Sourav Ganguly, **Arnesh Banerjee**, Ayushi Bhattacharjee, Avirup Chakraborty, Arnob Ghosh. Analyzing Historical Revisionism in LLMs in the Context of Indian History. *Advisor: Dr. Arnob Ghosh.*
- **Arnesh Banerjee**, Ayushi Bhattacharjee, Subhajit Datta. Understanding the Limitations of LLMs in Mathematical Reasoning. *Advisor: Prof. Subhajit Datta.* Part of B.Tech degree coursework.

Achievements

- **Department Third**: 4th Semester, SGPA 9.46, B.Tech CSE(DS), Heritage Institute of Technology, Kolkata.
- **Institutional Innovation Council (IIC)**: Selected as one of 10 IIC members across all years to represent the Department of CSE(DS), HIT Kolkata.
- **Summer Research Internship Offers 2026**: Selected for IIT Kharagpur, 3× IIT Patna, IIT Dhanbad, and the IIM Ahmedabad AI Venture Summer Internship.
- **WBJEE 2023**: Top 5.3% in West Bengal.

Technical Skills

Languages: Python, R, C, Java, SQL, LaTeX
Deep Learning & LLMs: PyTorch, Hugging Face Transformers, TRL, PEFT/LoRA, Keras, JAX
Classical ML: Scikit-Learn, XGBoost, CatBoost, H2O
Reinforcement Learning: Gymnasium, OpenAI Gym, Stable-Baselines3; DQN, Double DQN, PPO, SAC, RLHF
Data & Visualization: Pandas, NumPy, SciPy, Matplotlib, Seaborn, Plotly
Frameworks & Tools: Flask, FastAPI, Shiny, Docker, Git, Linux
Cloud: AWS
Relevant Courses: Machine Learning, Data Mining, Data Warehousing, Linear Algebra, Probability & Statistics, Calculus, Discrete Mathematics, Data Structures, Algorithms, Operating Systems, Database Management Systems