Arnav Raj

arnavvraj.compsci@gmail.com • arnav-raj.vercel.app • deadsmash07 • arnavv-raj

Education

Indian Institute of Technology Delhi Delhi, India
Dual Degree (B.Tech + M.Tech), Computer Science & Engineering 2022 – 2027
Coursework: Artificial Intelligence, Machine Learning, Deep Learning, Information Retrieval, AI Responsibility, Numerical Algorithms, Operating Systems, Computer Networks.

Research Interests

RLHF & Reward Modeling · LLM Evaluation & Interpretability · AI Alignment · Reinforcement Learning

Publications

PEBS: Per-rater Empirical-Bayes Shrinkage for RLHF Reward-Model Calibration
Arnav Raj. **Accepted, ICML 2026 Workshop on Pluralistic Alignment**. OpenReview.
Introduced a lightweight statistical layer that calibrates RLHF reward models to each **individual human annotator** rather than collapsing thousands of raters into a single average. Improves alignment to genuine human preferences on standard pluralistic-alignment benchmarks while **leaving the underlying reward model unchanged**.

Retroactive Advantage Correction: Closed-Form V-Trace Bias Correction for Delay-Aware RLHF
Arnav Raj. Under review, **ICML 2026 Workshop on RL from World Feedback (RLxF)**.
Proposed a correction primitive for production RLHF pipelines where **reward signals arrive late**, such as slow code verifiers, large judge ensembles, and queued human review. Allows policy training to continue using delayed feedback instead of discarding it, with a **closed-form proof of unbiasedness** and substantial improvement over standard wait-or-drop strategies.

KG-MuLQA: Multi-hop Question Answering over Knowledge Graphs for Long-Context Evaluation
Nikita, Vidhyakshaya,* Haricharana,* ***Arnav Raj***, et al. (*equal contribution). **Accepted, ACL 2026**. arXiv:2505.12495.
Benchmark for evaluating **multi-hop reasoning** in long-context language models, built from **knowledge graphs over real-world financial documents**. Designed and implemented the end-to-end pipeline for question generation, answer synthesis, and evaluation across frontier LLMs.

Hyperbolic Geometry of Reasoning: Probing LLM Hidden States
Arnav Raj. **Accepted, ICLR 2026 Workshop on Geometry-grounded Representation Learning (GRaM)**.
Interpretability study of how chain-of-thought reasoning models internally encode hierarchical structure. Showed that **hyperbolic geometric probes** recover this structure across the entire network while standard Euclidean probes fail in the deepest layers, evidence that reasoning models compress hierarchy into the representations responsible for the final answer.

Research Experience

Google DeepMind Mar 2026 – Present
STEM Expert, AI Evaluation & SFT *Bengaluru, India (Remote)*

- Authoring supervised fine-tuning and evaluation data for **internal Gemini models** on expert-level machine-learning and deep-learning reasoning tasks; designing structured rubrics for correctness, reasoning depth, and pedagogical clarity that feed directly into Gemini's training-data quality pipeline.

Abundant AI (YC W24) Nov 2025 – Present
ML Engineer, GPU Infrastructure & Research Agents *San Francisco, CA (Remote)*

- Building **GPU-backed evaluation infrastructure** for long-running **research agents**, enabling reinforcement-learning-driven data curation on tasks that require deep-learning compute rather than CPU-only sandboxes; working with the founding research team to make agent-driven workflows productionizable.

Harvard University, Edge Computing Lab May 2024 – Dec 2024
Research Intern *Cambridge, MA (Remote)*

- Built a **LangChain-based benchmarking framework** for LLM-generated **RTL hardware designs** (GPT-4, Llama variants) with automated syntax checking, testbench verification, and power-performance-area analysis with iterative re-prompting for failing designs; compared chain-of-thought, zero-shot, and few-shot prompting on the resulting accuracy and latency tradeoffs.

Selected Projects

LLM Code-Agent Evaluation Suite *Python, Docker, LangChain, OpenAI API*
End-to-end evaluation harness for LLM coding agents inspired by **SWE-bench and MLE-bench**. Implemented **Docker-sandboxed** task execution, automated unit-test validation, and code-quality scoring across multi-step coding scenarios with iterative self-correction. Used to systematically benchmark frontier coding models (**GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro**) and surface where each model fails and how reliably it recovers when given another attempt.

Neural-Augmented Retrieval Engine *Elasticsearch, FAISS, hnswlib, SentenceTransformers*
Production-style hybrid search system over a corpus of more than one million documents, combining classical **BM25 lexical matching** with **HNSW-based dense vector retrieval**. Added LLM-driven query rewriting, relevance feedback, and a **cross-encoder re-ranking** stage, served through a FastAPI backend with Redis caching to support low-latency interactive queries against the full index.

Reinforcement-Learning Agent for Code Optimization *PyTorch, OpenAI Gym, Ray RLlib*
Reinforcement-learning agent that learns to rewrite programs for better runtime and memory profile. Trained a **Proximal Policy Optimization (PPO)** policy under a custom reward combining execution time, memory footprint, and correctness signals, with curriculum-style task scheduling; the agent produced consistent performance improvements over unoptimized baselines across a range of algorithmic problems.

Heterogeneous GNN for User Personality Prediction *PyTorch, PyTorch Geometric, NetworkX*
Graph neural network for predicting **Big-Five-style personality traits** from user-product interaction graphs. Designed a heterogeneous **GraphSAGE** architecture with attention-based aggregation over a bipartite user-product graph, incorporated temporal interaction features, and trained multi-task heads jointly across the trait dimensions, outperforming non-graph tabular baselines.

Technical Skills

Languages: Python, C/C++, SQL, JavaScript, Bash.
ML / AI: PyTorch, PyTorch Geometric, Hugging Face Transformers, TransformerLens, LangChain, scikit-learn, CUDA.
Infrastructure: Docker, Kubernetes, Linux, Git, GitHub Actions, FastAPI, Flask, Weights & Biases.
Retrieval / Evaluation: Elasticsearch, FAISS, hnswlib, SentenceTransformers.

Honors & Awards

- **NK Securities Scholar**: merit scholarship awarded annually to the top 30 undergraduate students at IIT Delhi for sustained academic and technical excellence across the institute.
- **Smart India Hackathon**: National Top-5 finalist in both the 2023 and 2024 editions of India's largest student innovation competition, organised by the Government of India with over 200,000 participating teams nationwide.
- **JEE Advanced 2022**: All-India Rank **1,158** out of more than one million candidates (top 0.1%) in the entrance examination for admission to the Indian Institutes of Technology.
- **KVPY SX Fellowship 2021**: National science fellowship jointly awarded by the Department of Science & Technology, Government of India, and the Indian Institute of Science, Bangalore, for exceptional aptitude in the basic sciences.
- **IMC Prosperity Trading Challenge 2025**: Global Rank **8** in Round 1 of an international algorithmic-trading competition hosted by IMC Trading, a leading global market-making firm.
- **Codeforces Expert** (rating 1700+): competitive programming, with regular participation in rated contests across algorithms, data structures, and combinatorial problem-solving.

Leadership & Service

AI Safety Club, IIT Delhi 2025 – Present
Founding Member & Technical Lead

- Co-founded a student research group focused on AI alignment, interpretability, and evaluation; led reading groups on mechanistic interpretability (TransformerLens) and completed the BlueDot Impact AI safety programme and ARENA (Alignment Research Engineer Accelerator) interpretability curriculum.

STEM AI Hackathon 2026 (AI-Collab Hack) Jan 2026 – Present
Technical Consultant, IIT Delhi, Imperial College London, Microsoft Garage

- Mentoring 20+ student teams building AI agents for STEM education at a joint hackathon between IIT Delhi, Imperial College London, and Microsoft Garage.

Senior Editor, Tech Ambit (Pan-IIT Magazine) 2023 – 2025
Led a 15-member editorial team spanning 23 IITs; curated and edited 30+ technical articles on AI and systems research.