# Arnav Raj

+91 8434779969 | arnavvraj.compsci@gmail.com | arnav-raj.vercel.app
github.com/deadsmash07 | linkedin.com/in/arnavv-raj

## Education

**Indian Institute of Technology Delhi** Delhi, India
*Dual Degree (B.Tech + M.Tech) in Computer Science & Engineering* 2022 – 2027

**Relevant Coursework:** Machine Learning, Deep Learning, Information Retrieval, Special Topics in ML, Logic for CS, Theory of Computation, Operating Systems, Computer Networks, Database Management Systems

## Research Interests

AI Safety · LLM Evaluation & Interpretability · Model Observability · Reinforcement Learning · Retrieval-Augmented Generation · ML Systems

## Publications

**Where Lies Go to Hide: Detecting Hallucinations in LLMs via Internal Representations**
*Arnav Raj*
Submitted to ICLR 2025 Workshop. Investigated the internal representation of truth and falsehood in LLMs by analyzing the geometry of activation spaces. Developed unsupervised probing framework mapping reasoning tokens to hyperbolic space, achieving 87.5% detection accuracy (0.937 AUROC) on hallucinations as geometric outliers.

**KG-MuLQA: Multi-hop Question Answering over Knowledge Graphs for Long-Context Evaluation**
*Nikita Tatarinov, B Vidhyakshaya Kannan,* Haricharana Srinivasa,** ***Arnav Raj****, et al. (*equal contribution)*
Submitted to ACL 2025 (ARR). Framework for systematic question generation and structured long-context reasoning evaluation. 20,139+ multi-hop QA pairs across 170 documents. arXiv:2505.12495

## Research Experience

**AI Training Data Research Intern (RLHF / RL)** Nov 2024 – Present
*Abundant AI* *San Francisco, CA (Remote)*

- Designed advanced ML and data science tasks that expose systematic failure in state-of-the-art LLMs
- Contributed to datasets powering 3 of the top 6 global AI labs and multiple Fortune 500 enterprises
- Curated high-difficulty tasks and environments used to train LLMs via reinforcement learning

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

- Built LangChain benchmarking framework for RTL code generation across GPT-4 and Llama models
- Implemented end-to-end validation pipeline: syntax checking → testbench validation → PPA analysis with automated re-prompting for failing designs
- Compared Chain-of-Thought, zero-shot, and few-shot prompting strategies across graded design complexity

- Tracked accuracy and latency metrics across different prompt engineering approaches

**Research Intern** May 2024 – Jun 2025
*Georgia Institute of Technology – FSI Lab* *Atlanta, GA (Remote)*

- Co-developed KG-MuLQA framework for generating multi-hop knowledge-graph questions (ACL 2025 ARR submission)
- Created dataset of 20,139 long-context multi-hop QA pairs for structured reasoning evaluation
- Built scalable LLM benchmarking pipeline with auto-chunking, batched generation, and multi-chunk answer synthesis across 170 credit agreements
- Designed evaluation infrastructure for long-context understanding in financial documents

## Selected Projects

**Transformer-Based Hangman Solver** *PyTorch, NLTK, Transformers*
Built curriculum learning based GRU model achieving 67% success rate on unseen data using language heuristics and multi-strategy guess selection.

**Advanced Financial Analytics Platform** *C++, Python, Flask, Plotly*
Implemented low-latency trading strategies (market-making, mean-reversion, statistical arbitrage). Achieved Sharpe Ratio 2.1 and 25% reduction in max drawdown. Real-time PnL visualization dashboard.

**Neural-Augmented Retrieval Engine** *Elasticsearch, Faiss, hnswlib, SentenceTransformers*
Hybrid search system combining inverted indexing with dense vector search for semantic matching. FastAPI backend with Docker deployment.

**Graph Neural Network for User Personality Prediction** *PyTorch, PyTorch Geometric*
Designed GNN on bipartite user-product interaction graph. Achieved Weighted F1-score of 0.91, outperforming baselines by 15%.

**Context-Aware Spelling Correction** *Python, NLTK, NLP*
Implemented noisy-channel model with N-gram language models and multiple smoothing techniques. 88% accuracy on context-dependent spelling errors.

**Monte Carlo Tree Search Agent for Havannah** *Python, MCTS*
Built MCTS-based AI with UCB exploration and domain-specific heuristics. 81% win rate over RAVE-only baselines.

**SDN-Based Intelligent Network Controller** *Python, Ryu, OpenFlow, Mininet*
Implemented OpenFlow controller with proactive L2 learning, shortest-path routing, and loop prevention for complex topologies.

**OS Kernel Enhancements in xv6** *C, x86 Assembly, QEMU*
Extended xv6 pedagogical OS with virtual memory improvements, scheduler optimizations, and new system calls.

## Technical Skills

**Languages:** Python, C/C++, SQL, Verilog, Bash, Dart, Assembly (RISC-V), Prolog, Lean
**ML/AI:** PyTorch, PyTorch Geometric, Transformers, LangChain, NLTK, Scikit-learn, LightGBM, CUDA
**Systems:** Docker, Kubernetes, Git/GitHub, CI/CD (GitHub Actions), Linux/UNIX, Flask, FastAPI
**Research:** Elasticsearch, Faiss, hnswlib, SentenceTransformers, OpenCV, Mininet, QEMU

## Honors & Awards

- **IMC Prosperity Trading Challenge:** Global Rank 8 (Round 1) – 2025
- **Smart India Hackathon:** 2× National Top 5 Finalist – 2023, 2024
- **JEE Advanced:** All India Rank 1158 (Top 0.1% in over 1,000,000+ candidates) – 2022
- **KVPY SX Fellowship:** Awarded by Government of India & IISc Bangalore – 2021
- **National Science Olympiads:** Top 250 Astronomy (NSEA), Top 300 Chemistry (NSEC) – 2021
- **Codeforces:** Expert (1700+ rating)
- **Best Mess Secretary:** Awarded as Best secretary for my leadership work in college.

## Leadership & Service

**Founding Member & Technical Lead** 2025 – Present

*AI Safety Club, IIT Delhi*

- Co-founded club focused on AI alignment, interpretability, and evaluation research
- Led reading groups on mechanistic interpretability, model evaluation, and safety
- Completed structured alignment training (BlueDot Impact) and ARENA curriculum modules

**Senior Editor** 2023 – 2025

*Tech Ambit (Pan-IIT Magazine)*

- Led 15-member editorial team across 23 IITs publishing articles on emerging technologies
- Curated and edited 30+ technical articles, increasing readership by 40%

**Mess Secretary** Jun 2024 – May 2025

*Zanskar Hostel, IIT Delhi*

- Elected by 400+ residents; managed 13-member team improving hygiene and food quality
- Led digitalization of entire mess operations and transparency.