

# Advik Raj Basani

+91 9611724762  
f20221155@goa.bits-pilani.ac.in  
a3v1k.com  
in a3v1k  
floofcat  
poq7xJsAAAAJ

## Education

- 2022–2026 **B.E. Computer Science (Honors)**, *BITS Pilani*, Goa, ([Transcript](#))
- Achieved a **CGPA of 9.3/10.0**; **Rank 1** in the Department, Fall 2024.
  - Recipient of BITS Merit Scholarship, awarded to the top 4% of 1K students, for 3 consecutive semesters in recognition of high academic standing.

## Publications & Preprints

- [Under Review] *Exposing the Illusion of Erasure in Knowledge Editing for LLMs*; **Advik Raj Basani**, Anshuman Chhabra. **Under Review at NeurIPS '26**.
- [Proceedings] *GASP: Efficient Black-Box Generation of Adversarial Suffixes for Jailbreaking LLMs*; **Advik Raj Basani**, Xiao Zhang. **Accepted to NeurIPS '25 and Oral Presentation at Building Trust in LLMs @ ICLR '25**. [Slides] [GitHub]
- [arXiv] *Diversity Boosts AI-Generated Text Detection*; **Advik Raj Basani**, Pin-Yu Chen. **Accepted to TMLR (Transactions of Machine Learning Research), Data in Generative Models @ ICML '25 and Oral Presentation @ CLEF '25**. [Slides] [GitHub] [HuggingFace]
- [Proceedings] *G-GQSA: Exploiting Feature-Based Vulnerabilities and Enhancing Adversarial Resilience in Android Malware Detection*; **Advik Raj Basani**, Hemant Rathore. **Accepted as an Oral Presentation to 22<sup>nd</sup> CCNC '25**.
- [Proceedings] *When Less is More: Achieving Faster Convergence in Distributed Edge Machine Learning*; **Advik Raj Basani**, Siddharth Chaitra Vivek, Advait Krishna, Arnab K. Paul. **Accepted to 31<sup>st</sup> HiPC '24, Best Paper Nominee (top-2.5%)**. [GitHub]

## Research Experience

- Jun. 2026 **UK AI Security Institute**, London, UK, [MATS Research Fellow](#)  
Present Advisor: [Dr. Eric Winsor](#) (Research Scientist)
- Aug. 2024 **IBM Research AI**, Remote, US, Research Intern  
Present Advisor: [Dr. Pin-Yu Chen](#) (Principal Research Scientist)
- Investigating flow-matching approaches for activation steering in LLM, with a focus on safety & hallucinations, in collaboration with [Prof. Nisha Chandramoorthy](#).
  - Developed **DivEye**, a framework capturing surprisal-based diversity to identify statistical fingerprints of AI-generated text; accepted to [DIG-BUGS Workshop at ICML 2025](#).
    - Attained up to 0.99 **AUROC** on various benchmarks and secured 3<sup>rd</sup> on the **RAID** leaderboard (2024); invited for an oral presentation at **CLEF 2025**.
- Jan. 2024 **CISPA Helmholtz Center for Information Security**, Germany, Research Assistant  
Present Advisor: [Prof. Xiao Zhang](#)
- Investigating LLM reverse engineering, extending the work of Carlini et al. ([arxiv.org:2403.06634](#)) to infer architecture & weights from black-box models.
  - Designed **GASP (Generative Adversarial Suffix Prompter)**, an efficient black-box framework for generating coherent adversarial suffixes that expose vulnerabilities in LLM safety mechanisms.
    - Open-sourced code & dataset [AdvSuffixes](#), achieving SoTA performance on proprietary models; accepted at **NeurIPS 2025**.
- Feb. 2026 **Supervised Program for Alignment Research (SPAR)**, Remote, Research Fellow  
May 2026 Advisors: [Dr. Daniel Tan](#) (Center on Long-Term Risk), [Chloe Li](#) (Anthropic)
- Investigated the fragility of deceptive LLM model organisms, demonstrating that finetuning-based auditing methods artificially **inflate safety metrics by triggering catastrophic forgetting of hidden objectives** rather than eliciting genuine confessions.
- Sept. 2025 **PALM Lab**, University of South Florida, Remote, Research Intern  
May 2026 Advisor: [Prof. Anshuman Chhabra](#)
- Investigated the mechanistic failures of knowledge editing in LLMs, proving that **updates superficially suppress rather than erase pre-trained facts** by geometrically displacing memories into highly vulnerable, anisotropic regions.
    - Developed attacks for **reverse engineering** post-hoc edits to recover suppressed knowledge, exposing severe security and redaction risks (submitted to NeurIPS 2026).

Nov. 2023 **Data, Systems & HPC Lab**, Research Assistant

May 2026 Advisor: [Prof. Arnab K. Paul](#)

- Accelerating GPU I/O operations by evaluating data transfer trade-offs between CPU and **GPUDirect**, aimed at improving throughput and reducing latency in training workloads.
- Developed an **open-source framework** for **Distributed Machine Learning** on resource-constrained clusters by prioritizing critical gradient updates to accelerate convergence and reduce communication overhead.
  - Achieved a 13.22× reduction in training time and 62.1% lower communication overhead; work accepted as a **Best Paper Nominee** at the 31<sup>st</sup> **HiPC**.

---

## Work Experience

Oct. 2025 **Trexquant**, *Remote, US*, Global Alpha Researcher – Intern

Nov. 2025 Supervisors: [Saurabh Agarwal](#), [Dr. Yunbo Zhang](#), [Dr. Xin Wang](#)

- Designed and implemented 30+ research-grade alpha factors on **Pysim** for U.S. equity markets, focused on quarterly earnings prediction, statistically validated for robustness.

May. 2025 **Oracle**, *India*, Member of Technical Staff – Intern

Aug. 2025 Supervisors: [Anurag Sinha](#), [Harish Dalmia](#) (GenAI Team)

- Designed an end-to-end feature importance and factor analysis framework for Oracle's internal **AutoML pipeline**, offering configurable modes (fast vs. comprehensive) with model-agnostic support and scalable interpretability analysis.
- Integrated an extensive evaluation system, improving existing frameworks by 11×, for sensitive consumer use-cases such as employee attrition and financial market analysis.

May. 2024 **Centre for Development of Advanced Computing**, *Kolkata, India*, Research Intern

Aug. 2024 Supervisor: [Bibekananda Kundu](#) (Research Scientist)

- Conducted R&D on fine-tuning LLMs to be human-centric via reinforcement learning, integrating emotional intelligence, empathy, and task-oriented conversational abilities.

---

## Selected Projects

[\[Repo\]](#) 2025 **[Re]-Teaching Differentially Private Prompt Tuning for LLMs**

- Reimplemented & verified all experiments from **Flocks of Stochastic Parrots: Differentially Private Prompt Learning for LLMs** ([arXiv:2305.15594](#)), a study on applying differential privacy to prompt tuning for LLMs, and proposed new techniques that improved benchmark performance by ~2%. [\[Slides\]](#)

[\[PR\]](#) 2024 **FaustNet: Enabling ML in Faust**

GSoc Contribution

Mentors: [Thomas Rushton](#), [Dr. Stéphane Letz](#), [Dr. Yann Orlarey](#) (INRIA & GRAME)

- Developed an automatic differentiation library for the functional, audio domain-specific language **Faust** during [Google Summer of Code](#), enabling audio engineers to integrate neural networks & other ML techniques.

---

## Teaching Assistantships

CS F363 **Compiler Construction**, *Spring, 2025*

Lead TA; Instructor: [Dr. Santonu Sarkar](#)

Tasks: *Assignments, Labs & Grading*

~300 students

CS F242 **Microprocessors & Interfacing**, *Spring, 2025*

Course Mentor (CM); Instructor: [Dr. Gargi Alavani](#), [Dr. Manideepa Mukherjee](#)

Tasks: *Invigilation & Grading*

~300 students

CS F446 **Data Storage Technologies & Networks**, *Spring, 2025*

Course Mentor (CM); Instructor: [Dr. Arnab K. Paul](#)

Tasks: *Grading*

~65 students

BCSZC315 **Multicore & GPGPU Programming**, *Summer, 2025*

Lead TA; Instructor: [Dr. Gargi Alavani](#), [Dr. Kunal Korgaonkar](#)

Tasks: *Creation of Teaching Material & Grading*

~40 students

CS F242 **Microprocessors & Interfacing**, *Spring, 2024*

Lead TA; Instructor: [Dr. Gargi Alavani](#)

Tasks: *Tutorials, Lab Creation & Autograder*

~300 students

CS F422 **Parallel Computing**, *Fall, 2024*

Course Mentor (CM); Instructor: [Dr. Gargi Alavani](#)

Tasks: *Tutorials, Assignments & Docker-based CUDA simulator*

~50 students

CS F111 **Computer Programming**, *Spring, 2023*

Course Mentor (CM); Instructor: [Dr. Arnab K. Paul](#)

Tasks: *Plagiarism Detection & Autograder*

~1000 students

---

## Relevant Coursework

BITS Computer Networks\*, Data Structures and Algorithms, Operating Systems\*, Time Series Analysis and Forecasting\*, Deep Learning\*, Machine Learning, Compiler Construction\*, Computer Architecture

\* - Top 5 Student

Coursera Specialization in Google Data Analytics [\[Certificate\]](#), Advanced Learning Algorithms [\[Certificate\]](#), Supervised Machine Learning: Regression and Classification [\[Certificate\]](#)

## Accomplishments

- Grants Recipient of the IEEE TCPP Grant and DDF Grant, NTSE & KVPY Scholarships.
- 2025 Organizer of the [CISPA European Cybersecurity & AI Hackathon Championship](#), Vienna.
  - 2025 Reviewer for TMLR, ICLR & ACL.
  - 2025 Scored 326 (170Q, 156V) / 340 in GRE & 112 (27R, 28L, 30S, 27W) / 120 in TOEFL.
- 2024-25 2× winner of [Hackenza](#), a hackathon organized by ASCII, BITS Goa.
- 2023 Coordinator & Lead, Google Developer Student Club, BITS Goa.
- Coordinated hackathons, seminars, and events across four verticals; oversaw the **AI/ML vertical** and club operations.
- 2023 Core Member for [BITSKrieg](#), ranked 1<sup>st</sup> for performance in CTFs across India.
- 2022 Developer for **Twitch Rivals: Medieval Mayhem**, **BisectHosting's GameMaster** & several other Minecraft tournaments and events. [[Playlist](#), [2.5M+ views](#)]

## Technical Proficiency

- Languages Proficient [Python, Java, C++, ~~TeX~~], Intermediate [Faust, Julia, Rust] & more
- Libraries / PyTorch, Transformers, JAX, Flax, HuggingFace, GitHub, GitLab, Anaconda, SpringBoot, Slurm, Docker,  
Softwares vLLM, Kafka, ZeroMQ, SciKit-Learn, GraphQL, Gemini & OpenAI APIs