SAISAB SADHU

+91 9933387874 saisab21@iiserb.ac.in sadhusaisab@gmail.com
LinkedIn GitHub Scholar

EDUCATION

Indian Institute of Science Education and Research Bhopal (IISER Bhopal) 12/2021 - 04/2026
MS & B.Tech (Integrated) in Data Science and Engineering CPI : 7.31/10

INDUSTRY EXPERIENCE

Lexsi Labs March 2026 – Present
Research Intern, Mechanistic Interpretability *Mumbai, India*

- Investigating whether standard machine unlearning methods produce genuine circuit disruption or behavioral suppression, using EAP-IG attribution shifts over Gemma Scope SAE features pre- and post-unlearning on TOFU forget10.
- Operationalizing the hypothesis that unlearning, like instruction-tuning, suppresses behavioral expression of internal knowledge states without disrupting underlying circuits; validated via relearning attack recovery speed.

MIQ Digital Jan 2026 – March 2026
Senior Analyst Intern, DnA Team *Bengaluru, India*

- Analyzed large-scale user datasets on Databricks over AWS S3; applied statistical methods and ML models for campaign optimization and attribution modeling while completing MS thesis remotely.

RESEARCH EXPERIENCE

Bio Medical Data Science Lab, IISER Bhopal Jan 2025 – Present
Graduate & Undergraduate Researcher *Bhopal, India*
PI: Dr. Tanmay Basu — Collaborators: Dr. Dwaipayan Roy (IISERK), Dr. Biswajit Patra (IISERB)

- **MS Thesis (Ongoing):** Investigating mechanisms to overcome imperfect context retrieval and resolve knowledge conflicts, both **parametric** and **external** in RAG Frameworks. Developed a dialectical engine that operationalizes formal cross-examination to dynamically assess source credibility based on logical resilience, achieving 77% improvement on FaithEval and 28% on RAMDocs.
- **BS Thesis:** Engineered a hybrid extractive–abstractive summarization pipeline achieving strong performance (53.13 ROUGE-1) on CNN/DailyMail; developed a ModernBERT-based Siamese extractive stage using a scaled adaptive margin triplet loss for optimal candidate ranking.
- **Biomedical NLP (Ongoing):** Developing an end-to-end deep learning framework for automated PICO (Population, Intervention, Comparison, Outcome) extraction from full-text clinical papers in collaboration with ICMR Bhopal to curate annotated datasets of full research articles to accelerate evidence synthesis for systematic reviews.

School of Public Policy, IIT Delhi May 2024 – July 2024
Research Intern *New Delhi, India*
Guide: Dr. Nandana Sengupta; Co-Guide: Dr. Ravinder Kaur, Dr. Sangeeta Kohli

- Analyzed 2000+ faculty profiles (IRINS) to identify a 12% gender differential in negative marking impact; evaluated the socio-economic viability and impact of the 20% supernumerary quota for women at IITs.

PUBLICATIONS

When RAG Disagrees: Detecting Latent Epistemic Conflict via Logit Interactions
Accepted at the 49th International ACM SIGIR Conference on Research and Development in IR *SIGIR 2026*

- Identified a mechanistic law ($p < 10^{-42}$) predicting internal epistemic conflict in 70B models; discovered the Alignment Paradox where instruction-tuning decouples internal tension from textual behavior.

- Developed a 25ms "Mechanistic Auditor" that identifies latent sycophancy with 76% F1, significantly outperforming model self-reporting and achieving a 600× speedup over state-of-the-art probing.

DARE: A Dialectical Framework for Adversarial and Evidence-Aware RAG
Accepted at the 48th European Conference of Information Retrieval *ECIR 2026*

- Resolved factual conflicts in RAG via a dialectical cross-examination process; achieved SOTA gains of **77%** on FaithEval and **28%** on RAMDocs.
- Introduced dynamic credibility assessment, a mechanism that infers source reliability from logical resilience to adversarial challenges rather than static weighting.

Structured Adversarial Synthesis: A Multi-Agent Framework for Generating Persuasive Financial Analysis from Earning Call Transcripts
Published at Proceedings of The 10th Workshop on FinNLP (EMNLP 2025) *EMNLP Workshop 2025*

- Designed a hierarchical agentic framework modeling investment committee debates to synthesize persuasive financial analysis; demonstrated a **68.75% win rate** over cooperative baselines.

Structure-Aware Chunking for Abstractive Summarization of Long Legal Documents
Published in JustNLP Workshop at IJCAI-AACL 2025 *IJCAI-AACL 2025 Workshop*

- Proposed a rhetorically-informed pipeline for ultra-long legal documents; identified the "Coherence Gap" trade-off between local phrase accuracy and global narrative flow.

Hierarchical Pedagogical Oversight: A Multi-Agent Adversarial Framework for Reliable AI Tutoring
Accepted for Presentation at AAAI 2026 EGSAI Community Activity *AAAI EGSAI 2026*

- Operationalized adversarial oversight for reliable AI tutoring; an 8B-parameter model structured via HPO outperformed GPT-4o by **3.3%** in Macro F1 on the MRBench dataset.

TECHNICAL SKILLS

Languages	Python, C, R, SQL, Matlab, Bash
ML/DL Frameworks	PyTorch, TensorFlow, Keras, Scikit-learn, Hugging Face (Transformers, PEFT)
NLP & LLMs	NLTK, Gensim, spaCy, stanza, LangChain, LangGraph
GenAI & Research	AutoGen, DPO, LoRA/QLoRA, RlHF, G-Eval, RAG architectures
Data & Automation	PySpark, pandas, NumPy, BeautifulSoup, Selenium
Dev Tools	Git, Docker, Kubernetes, SLURM, Databricks, Weights & Biases

ACHIEVEMENTS

- **AAAI 2026 EGSAI Selection:** Selected to present *"Hierarchical Pedagogical Oversight: A Multi-Agent Adversarial Framework for Reliable AI Tutoring"* at AAAI 2026; one of 51 works chosen for global submissions.
- **Top Performer, FinNLP @ EMNLP 2025:** Ranked first globally on the official 'Win Rate vs Analyst Report' metric; system reports were preferred over professional human analysts.
- **Recipient, Student Innovation Grant (Rs. 2 Lakhs):** Awarded by IICE (Funded by DST, GOI) to develop an AI fintech platform; demonstrated 68% profit increase in backtesting.
- **CARE Conference Travel Grant:** Awarded full registration waiver and travel support (IIT Guwahati) for poster presentation at the Collaborative for Academic Research Excellence Conference.

WORKSHOPS & CONFERENCES

- **FinNLP Workshop at EMNLP 2025:** Virtually presented *"Structured Adversarial Synthesis"* and participated in shared task discussions at the 10th FinNLP Workshop. *EMNLP(2025)*
- **CARE Conference (Data Science & AI):** Presented a poster on Multi-Agent Adversarial RAG at the First Mehta Family Foundation CARE Conference hosted by IIT Guwahati. *Guwahati, India (2025)*
- **Climate Change AI Summer School (Online):** Engaged with leading researchers on ML applications for climate science; participated in hands-on workshops on climate modeling. *Pittsburgh, US (2024)*

- **7th Summer School on AI (Focus on CV & ML):** Selected for an intensive program organized by the Centre for Visual Information Technology (CVIT) at IIIT Hyderabad. *Hyderabad, India (2024)*

POSITION OF RESPONSIBILITY

- **Secretary, Student Development Council - IISERB:** Oct 2023 – Aug 2024
  Organized institute-wide initiatives, workshops, and panel discussions to promote student growth.
- **Founder & Lead, Entrepreneurship Cell - IISERB:** Dec 2023 – Aug 2024
  Fostered a startup culture by hosting networking sessions and speaker events for aspiring entrepreneurs.
- **Institute Placement Coordinator, Center for Career Development:** Aug 2022 – Oct 2023
  Coordinated campus placements and industry outreach; achieved a notable increase in recruiter engagement.

REFERENCES

Dr. Tanmay Basu	Assistant Professor & HOD, Dept. of Data Science and Engineering, IISER Bhopal
Dr. Dwaipayan Roy	Assistant Professor, Dept. of Computational and Data Sciences, IISER Kolkata
Dr. Biswajit Patra	Assistant Professor & HOD, Dept. of Economic Sciences, IISER Bhopal