Navyansh Singh

Raipur, Chhattisgarh, India • +91-9630039116 • navyansh24102@iiitnr.edu.in
Google Scholar • huggingface.co/Navy0067

Research Interests

I am interested in which capabilities current language models actually possess, especially on tasks that in humans require understanding and reasoning: what exactly the models do in those cases, and where and why they are limited. I go after specific behaviours where a model departs from what I expected, or does something I cannot cleanly explain, and try to establish what the phenomenon is and why it exists. Three kinds of question recur: open questions about well-documented behaviours, especially where behavioural similarity makes clean separation difficult; construct validity, meaning what limits current benchmarks and whether the way they measure hides a precise issue; and where, exactly, a model's apparent understanding runs out, why, and whether that limit can be removed. I am also moving into mechanistic interpretability, to check what is happening internally against the behaviour.

Education

Indian Institute of Information Technology, Naya Raipur *Expected May 2028*

B.Tech. in Data Science and Artificial Intelligence

Relevant coursework: Machine Learning, AI Foundations, Representational Learning, Data Structures, Probability & Statistics, Linear Algebra, Optimization Techniques

Publications & Preprints

Fallacy Benchmarks Measure Scheme Recognition, Not Fallacy Detection *Aug 2026*

Lead author • Preprint • arxiv.org/abs/2609.18644

- Formulated the measurement claim, designed the two-condition control that isolates scheme content, and led experimental design and most of the implementation.
- Showed that the low false-positive rates fallacy benchmarks report are an artifact of how the "valid"/ "none" class is assembled: two independent annotators found scheme-matched valid arguments in only 3 of 183 sampled items across four benchmarks (CoCoLoFa, Reddit, MAFALDA, Argotario; three-way $\kappa = 0.74$).
- Built a generation pipeline producing two negative conditions differing *only* in argumentation scheme, making the result invariant to generation quality. Classifiers label scheme-matched negatives as the source fallacy type **40.9pp** more often than wrong-scheme negatives, which they instead label with the scheme actually used 85.9% of the time against 0.4% for the source type. FPR rises 16.6% $\rightarrow$ 58.9% (CoCoLoFa) and 5.7% $\rightarrow$ 62.0% (Reddit, trained from scratch); stable across seeds and architectures (ModernBERT, RoBERTa).
- Replicated the dissociation in three zero-shot LLM detectors that never saw the benchmarks (Claude Sonnet 5, Haiku 4.5, Qwen3.6-27B), and showed the raw FPR gap is unstable across prompts and models where the two-condition comparison is not — one detector shows +0.5pp raw but +25.5pp type-specific separation.
- Validated construction with three blind annotators (judge precision 93.1%); FPR on the human-confirmed-valid subset is 56.6%. The effect also reproduces on naturally occurring arguments and attenuates correctly on an independently built negative class (Touché 2026), showing the audit discriminates between construction protocols.
- Released **Scheme Foils** (1,042 scheme-matched and 1,127 wrong-scheme valid arguments, each carrying a human-validated precision estimate) and a five-step benchmark audit that requires no retraining.

Debiasing Logical Fallacy Detection for Real-World Robustness via Counterfactually Augmented Data *July 2026*

Sole author • ACL 2026 Student Research Workshop, pages 363–374 • aclanthology.org/2026.acl-srw.30

- Identified spurious-correlation bias in fallacy classifiers and designed the full pipeline: problem formulation, data generation, fine-tuning, and OOD evaluation.
- Introduced 703 LLM-generated counterfactually augmented (CAD) pairs — minimally differentiated valid and fallacious arguments — to debias models by targeted augmentation; verified with inter-annotator agreement (Cohen's $\kappa = 0.68$).
- Fine-tuned DeBERTa-v3-large on CoCoLoFa augmented with CAD pairs: 58% relative reduction in false-positive rate (64% $\rightarrow$ 26.7%) on a 300-sample Reddit-sourced OOD set.
- Showed LLMs (Llama-3.1-8B, Llama-3.3-70B) degrade sharply under simple human-like prompts (F1 63–72%, FPR 54–74%) despite strong performance under optimized prompts (F1 90–94%). Model, dataset, and interactive demo released on HuggingFace.

Hindsight-Guided Rationale Distillation for Rare Disease Diagnosis *2026*

Co-author • Under review, AACL-SRW

- Ran the qualitative analysis that identified and characterised the ground-truth hallucination artifact, and wrote and framed key sections.
- GT hallucination: label-visible trace generation causes the teacher to embed "ground truth is X" phrasing that SFT copies — found by qualitative sample analysis, confirmed at scale (33.9% of student inferences; 18pp accuracy collapse when the hallucinated label is wrong).
- A filtered variant trained on regex-cleaned traces achieves near-zero contamination (3/2,000) and a significant accuracy advantage over the 8B teacher ($p < 0.001$) at 1.5B parameters.

CONFERENCES

ACL 2026 — 64th Annual Meeting of the Association for Computational Linguistics *July 2026*

Virtual attendee • San Diego, CA