Sathwik Reddy

chintala.reddy@research.iiit.ac.in | +91 93479 59704 | Website | GitHub | Google Scholar

Education

International Institute of Information Technology (IIIT), Hyderabad June 2028
B.Tech + M.S. by Research in Computer Science

Publications and Preprints

1. Srija Mukhopadhyay, **Sathwik Reddy**, Shruthi Muthukumar, Jisun An, Ponnurangam Kumaraguru. **PRIVACYBENCH: A Conversational Benchmark for Evaluating Privacy in Personalized AI**. Accepted at **IJCNN 2026**; **nominated for Best Paper Award**, 2025. | arXiv:2512.24848
2. George Rahul*, Ishan Kavathekar*, Maitreya Chitale*, Manas Mittal*, **Sathwik Reddy***. **Cogito Ergo LLM: Evaluating Situational Awareness in Language Models**. Preprint, 2026. *(*Equal contribution; authors listed alphabetically.)* | Paper

Experience

Undergraduate Researcher, Precog Lab, IIIT Hyderabad May 2025 – Present
Advisor: Prof. Ponnurangam Kumaraguru

- Co-developed *PRIVACYBENCH*, a benchmark for evaluating privacy leakage in personalized AI assistants in multi-turn conversational settings, in collaboration with Prof. Jisun An (Indiana University); accepted at IJCNN 2026.
- Investigating efficient reward design for RL on vision–language models with Prof. Chirag Agarwal (Univ. of Virginia); building the rollout, probe-training, and GRPO fine-tuning pipeline for Qwen2.5-VL and LLaVA, with rewards read from the frozen base model's internal states instead of an LLM judge.

Research Intern, TurboML June 2026 – July 2026
Mentor: Arjit Jain

- Running training and reasoning-behavior experiments on recursive reasoning architectures, including Hierarchical Reasoning Models (HRM) and Tiny Recursive Models (TRM), which reason through repeated latent updates rather than long explicit chains of thought. Detailed methods and results are covered by a non-disclosure agreement.

Projects

SABER: Benchmarking Situational Awareness in Language Models 2026

- Co-developed SABER, a benchmark measuring situational awareness in LLMs along three axes: self-awareness (tool-use reasoning), awareness of other agents (role inference in multi-agent settings), and strategic use of context (instrumental goals such as self-preservation and deception).
- Designed conditions that separate performative alignment from baseline behavior; found that hiding evaluation cues barely shifts behavior, whereas explicit long-horizon framing sharply raises instrumental actions (self-preservation 21% $\rightarrow$ 45% for GPT-5-mini, the strongest model tested).
- Showed a consistent self-awareness gap on tool-reasoning tasks (80% vs. $\leq$2% optimal-tool selection, proprietary vs. small open-source), and uniformly weak multi-agent role inference that varying decoding temperature did not fix, pointing to a reasoning limitation rather than a sampling artifact. | Report

Direct Feedback Alignment for Efficient LLM Fine-tuning 2026

- Tested Direct Feedback Alignment as a backprop-free alternative for LoRA fine-tuning of Llama-3.2-1B on math datasets, with Prof. Chirag Agarwal; found random feedback matrices fail to approximate gradients in the high-dimensional weight space of LLMs, and reported the negative result. | Code

Training a Multilingual Small Language Model from Scratch 2025

- Trained a 154M-parameter Qwen3-style decoder-only LM from scratch on 3B+ tokens (English, Telugu, Bodo); studied perplexity-vs-downstream gaps and catastrophic forgetting during fine-tuning.
- Designed a custom architecture (Grouped-Query Attention, SwiGLU, RoPE, RMSNorm) with a 50K-vocab BPE tokenizer and cut training time $\sim$20x by switching from DDP to HuggingFace Accelerate. | Code

Skills

ML/NLP: LLMs, reasoning evaluation, RAG, multimodal systems, finetuning, agents
Frameworks/Tools: PyTorch, Hugging Face Transformers, FastAPI, LangChain, NumPy, scikit-learn, Git, Linux
Languages: Python, C++, SQL

Activities

- Selected participant, Agyeya AI Interpretability & Safety Workshop, IISc Bangalore (inaugural cohort), Sept 2026.
- Reviewer, TrustNLP (ACL 2026) and EIML (ICML 2026); sub-reviewer, ACL ARR (Jan 2026).
- Problem Setter, Indian National AI Olympiad (INAIO) 2025 and 2026.
- Attendee, IndoML 2025; participant, IASNLP 2024.
- Wrote blogs on ML and Math topics. | Link