






SATHWIK REDDY

 [GitHub](#)  [Website](#)  [LinkedIn](#)  chintala.reddy@research.iiit.ac.in  +91 9347959704

EDUCATION

International Institute of Information Technology, Hyderabad
B.Tech in Computer Science

Expected: June 2027

SKILLS

Languages: C, C++, Python, SQL, HTML/CSS

Libraries: NumPy, Pandas, PyTorch, TensorFlow, Keras, Transformers, scikit-learn, NLTK, spaCy

WORK EXPERIENCE

Undergraduate Researcher, Precog, IIIT Hyderabad
Advisor: Prof. Ponnuram Kumaraguru

June 2025 – Present

- Co-authored “PRIVACYBENCH,” a novel conversational benchmark for evaluating privacy vulnerabilities in personalized AI assistants, submitted to ACL ARR 2025.
- Investigated “Hidden Poison” camouflaged data poisoning attacks (Jun-Jul 2025, with Prof. Amartya Sanyal); integrated research codebases (‘Witches’ Brew’), analyzed model activations using clustering, and identified critical reproducibility challenges in the foundational papers.
- Contributed to the development of SARAL, an AI system that converts arXiv papers into educational videos.

Summer Intern, IIT Dharwad
Advisor: Dr. Konjengbam Anand

May–June 2024

- Developed a retrieval-augmented generation (RAG) chatbot for legal question answering specific to Indian law.
- Scraped Supreme Court case data and cleaned it for use in downstream classification tasks.
- Built a multi-label classifier to categorize Indian legal cases into Bailable/Non-Bailable, Cognizable/Non-Cognizable, and Initial Trial Court based on case judgments.

PUBLICATIONS

- Srija Mukhopadhyay, **Sathwik Reddy**, Shruthi Muthukumar, Jisun An, and Ponnuram Kumaraguru.
“PRIVACYBENCH: A Conversational Benchmark for Evaluating Privacy in Personalized AI.”
ACL ARR 2025 October Submission (Under Review).

PROJECTS

Custom 154M Parameter Multilingual SLM from Scratch

Sep 2025

- Designed and pretrained a 154M parameter decoder-only language model from scratch on a 3-billion-token multilingual corpus (English, Telugu, Bodo).
- Implemented a modern transformer architecture (Qwen-based) with Grouped-Query Attention (GQA), SwiGLU activation, RoPE, and RMSNorm for computational efficiency.
- Fine-tuned the model on downstream tasks (Logical Reasoning, Text-to-SQL), identifying and mitigating "catastrophic forgetting" and diagnosing the "Perplexity Paradox" to select the optimal checkpoint based on qualitative evaluation over raw metrics.

Privacy-Aware Conversational AI Evaluation (EACL 2026 Submission)

Jul 2025 - Oct 2025

- Designed and implemented a multi-turn conversational evaluation framework to benchmark privacy vulnerabilities in personalized RAG-based AI assistants.
- Employed an LLM-based prober, using direct and indirect strategies, to systematically test an assistant’s ability to preserve ground-truth user secrets during realistic dialogues.
- Quantified that baseline assistants leak secrets in up to 26.56% of interactions and analyzed the effectiveness of prompt-based defenses, identifying inappropriate data retrieval (IRR > 63%) as the core architectural flaw.

SARAL: AI System for Automated Educational Videos

May–June 2025

- Contributed to SARAL, a full-stack AI system that automates converting LaTeX/arXiv papers into engaging educational videos and slides.
- Built a robust pipeline using Gemini API (script generation), Sarvam API (narration), and MoviePy (video synthesis).
- Adopted by **2000+ researchers and students** to simplify paper consumption.

CERTIFICATIONS AND TECHNICAL ACHIEVEMENTS

- **Problem Setter, INOAI Summer Camp (2025)**, created two AI challenges for India's International AI Olympiad selection camp
- Participated in International Advanced Summer School on Natural Language Processing (IASNLP) 2024, conducted at IIIT Hyderabad from 21 June 2024 to 6 July 2024.
- Deep Learning Specialization by Coursera.
- Delivered a talk on Neural Networks and Deep Learning at IIIT Kottayam with 50+ audience.
- Wrote blogs on various topics in Deep Learning and Machine Learning