Satvik Golechha

zsatvik@gmail.com 7vik.io Scholar LinkedIn GitHub

Focus areas: Frontier AI alignment/security, interpretability, and reinforcement learning.

Research Experience

UK AI Security Institute (AISI)
Aug 2025 - Present

Research Scientist — *London, UK*

- Frontier alignment and security research including interpretability, sandbagging, and auditing of frontier systems with a focus on model internals and model organisms.

Center for Human-Compatible AI at UC Berkeley
2025 (May-Aug)

Research Intern — *Berkeley, CA*

- Reinforcement learning for efficient multi-turn exploration w/ LLM-agents.

ML Alignment & Theory Scholars
2024 – 2025

Research Scholar — *Berkeley, CA*

- Mentor: Adrià Garriga-Alonso (FAR AI): Benchmarked frontier agentic deception and activation monitoring using deception probes; sparsity proofs for absorption in SAEs.
- Mentor: Nandi Schoots (Oxford): Studied the geometry of hierarchical concepts in LLMs and modular training of neural networks to improve interpretability.
- Training program: Neel Nanda (Google DeepMind): Explored challenges in mechanistically interpreting model representations for harmful behavior.

Microsoft Research
2023 – 2025

Pre-doctoral Researcher — *Bengaluru, India*

- Theory of representation learning & hierarchical distance measures. (w/ Neeraj Kayal)
- Designed invariability and robustness measures for optimizing in-context examples in LLMs. (w/ Amit Sharma and Amit Deshpande)
- Built and open-sourced a multimodal, multi-lingual expert-in-the-loop chat systems for high-stakes, nationwide, healthcare deployments. (w/ Mohit Jain)

Wadhwani AI
2021 – 2023

Associate Research Scientist — *Mumbai, India*

- Formulated AI problems in healthcare and worked with the Govt. of India to train and deploy scalable, robust, and interpretable models for public health interventions.

Independent
2023 – 2025

External collaborations — *Remote*

- **Anthropic**: Collaborated on auditing LLMs for hidden objectives.
- **SPAR 2025**: Mentored a project on zero-knowledge auditing for undesired behaviors.
- Various independent research projects (see publications on next page).

Selected Work

(please see Scholar for a full up-to-date list and reach out to know more about my current work!)

Auditing Games for Sandbagging
Jordan Taylor, Sid Black, Dillon Bowen, Thomas Read, Satvik Golechha,
Alex Z-M., Oliver M., Connor K., Kola A., Jacob M., Sam Marks, Chris Cundy, Joseph Bloom
2025, AISI [code, demo, paper]

Among Us: A Sandbox for Measuring and Detecting Agentic Deception
Satvik Golechha, Adrià Garriga-Alonso
NeurIPS 2025 (Spotlight) [code, poster, blog]

Intricacies of Feature Geometry in Large Language Models
Satvik Golechha, Lucius Bushnaq, Euan Ong, Neeraj Kayal, Nandi Schoots
ICLR 2025 (poster); best blog award at ICLR [blog, code, award]

A is for Absorption: Studying Feature Splitting and Absorption in SAEs
David Chanin, James W., Tomáš D., Hardik B., Satvik Golechha, Joseph Bloom
NeurIPS 2025 (Oral) [code]

Auditing Language Models for Hidden Objectives
Samuel Marks, Johannes Treutlein, ..., Satvik Golechha, ..., Evan Hubinger
2025, Anthropic (external collaboration) [paper, Anthropic, blog]

NICE: To Optimize In-Context Examples or Not?
Pragya Srivastava, Satvik Golechha*, Amit Deshpande, Amit Sharma*
ACL 2024 (main) [paper, code]

Progress Measures for Grokking on Real-world Tasks
Satvik Golechha
ICML 2024: Workshop on High-dimensional Learning Dynamics [paper, code]

Predicting Treatment Adherence of Tuberculosis Patients at Scale
Mihir Kulkarni, Satvik Golechha*, Rishi Raj*, Jithin K Sreedharan*, Alpan Raval*
NeurIPS 2022 [paper, media]

CataractBot: An LLM-Powered Expert-in-the-Loop Chat System
Pragnya Ramjee, Bhuvan Sachdeva*, Satvik Golechha*, Mohit Jain*
UbiComp 2025 [paper, code]

Education

Birla Institute of Technology and Science, Pilani *Aug 2017 – June 2021*
Bachelor of Engineering in Computer Science *Pilani, India*