ACHU MENON

achyutha11@gmail.com | (312) 757-9616 | achyutha11.github.io | Google Scholar | linkedin.com/in/achu-menon

SUMMARY

AI safety researcher working on evaluations of LLM agents and, more recently, of LLMs controlling robots. Founding engineer at Robocurve, previously a MATS research fellow. Papers at ICML 2026 and ICLR 2026 workshops.

EXPERIENCE

Robocurve, Inc., San Francisco, CA **Sept 2026 – Present**
Founding Engineer

- Co-author on the Inspect Robots paper [5]. Gave feedback on library changes and helped develop the paper's main arguments. The library has had nearly 100k downloads since its July release.
- First author on Robocurve's reports evaluating GPT-6 Astra and Fable 5.1 on real robot-arm manipulation tasks [8, 9].
- Co-author on the RoboHarm study [6]. Across seven frontier models, requests refused in text were mostly carried out once the model controlled a robot arm.
- Run Robocurve's open-source benchmarking program. Currently working with eight robotics teams on 50+ new benchmarks, to be released as a joint paper in Q1 2027.

MATS Program, Berkeley, CA **June 2026 – Sept 2026**
Research Fellow, David Lindner stream

- Led a project on how agent harnesses affect safety: 5 frontier models, 5 production harnesses, 5 behavioral domains [10].
- Found that the same model's misalignment rate shifted by up to 67 percentage points depending on the harness.
- Ablated system prompts, tool schemas, and loop structure to find which parts of the scaffold drive the effect.

Supervised Program for Alignment Research (SPAR), Remote **Feb 2026 – May 2026**
Research Fellow, Santiago Aranguri stream

- Stress-tested the realism win rate, a metric for how realistic an evaluation looks. Transcript position alone moved results by up to 43 points.
- Found that win rates and absolute realism scores from the same judge often disagree.

AI Safety Research Collaboration, Remote **Dec 2025 – Feb 2026**
Independent Researcher

- Built an RL-training simulation pipeline to test whether frontier models resist training when given misaligned goals [1].

Supervised Program for Alignment Research (SPAR), Remote **Sept 2025 – Dec 2025**
Research Fellow, Diogo Cruz and Eyon Jang stream

- Built evaluation pipelines to measure goal drift in LLM agents on long-horizon tasks.
- Led writing and submission of first-author paper [2] to the ICLR 2026 Lifelong Agents Workshop.

PUBLICATIONS

Conference & Workshop Papers

[1] E. Jang, D. Falck, J. Braun, N. M. Kirch, **A. Menon**, P. Moodley, S. Emmons, R. S. Zimmermann, D. Lindner. Exploration Hacking: Can LLMs Learn to Resist RL Training? *ICML 2026* (Main Conference).

[2] **A. Menon**, M. Saebo, T. Crosse, S. Gibson, E. Jang, D. Cruz. Inherited Goal Drift: Contextual Pressure Can Undermine Agentic Goals. *ICLR 2026 Workshop on Lifelong Agents.*

[3] M. Saebo, S. Gibson, T. Crosse, **A. Menon**, E. Jang, D. Cruz. Asymmetric Goal Drift in Coding Agents Under Value Conflict. *ICLR 2026 Workshop on Lifelong Agents.*

[4] R. Subramani, R. Arike, O. Terry, **A. Menon**, Z. Jin, F. R. Ward, S. Herd. Continual Learning in LLM Agents: Capabilities, Risks, and Research Directions. *TAIS 2026 Workshop.*

Preprints

[5] C. Leet, **A. Menon**, S. Machcha, S. Zou, A. Patel, A. K. Singh, A. K. Singh, G. Vamsi, J. Ahuja, S. A. Yamani, T. Anand, V. Alle, Z. J. Zhang, T. K. Chan, J. Chooi. Inspect Robots: Evaluating the Capabilities and Safety of Embodied AI. *arXiv:2610.06306*, 2026.

[6] E. Sun, C. Leet, S. Machcha, **A. Menon**, T. K. Chan, Y. Cui, J. Chooi. Refuses in Chat, Complies with Arms: LLM Refusal Does Not Survive Robot Control. *arXiv:2610.14043*, 2026.

[7] S. Gibson, T. Crosse, M. Saebo, **A. Menon**, E. Jang, D. Cruz. Same Facts, Different Updates: Inference Setup Shapes LLM Behavior in Medical Allocation. *arXiv:2608.18108*, 2026.

Technical Reports

[8] **A. Menon**, S. Machcha, S. Zou, T. K. Chan, J. Chooi. GPT-6 Astra on Robotic Manipulation. *Robocurve report*, Sept 2026.

[9] **A. Menon**, S. Machcha, S. Zou, T. K. Chan, J. Chooi. Comparing Fable 5.1 to Fable 5 on Robotic Manipulation. *Robocurve report*, 2026.

Under Review

[10] **A. Menon** et al. More Than Just a Model: Evaluating the Effect of Agent Harnesses on Safety. 2026.

REVIEWING & MENTORSHIP

- **Reviewer**, ICLR 2027.
- **Mentor**, SPAR (Fall 2026): five mentees working as technical program managers on Robocurve's benchmarking program.
- **Areas:** embodied AI safety evaluation, robot manipulation benchmarks, LLM agent evaluations, goal drift, RL training robustness, evaluation realism and LLM-as-judge methods.

EDUCATION

University of California, San Diego **June 2026**
M.S., Computer Science (AI specialization) · GPA 4.00/4.00

University of Chicago **June 2022**
B.S., Molecular Engineering (Chemistry Track) · GPA 3.88/4.00, Magna cum Laude