

Seunghyun Yoo

West Lafayette, IN | (+1) 765-476-9977 | yoo236@purdue.edu
carteryoo.github.io | github.com/CarterYoo | [LinkedIn](#) | x.com/Seunghyun31614

Research Interests

Mechanistic Interpretability, Technical AI Safety, and Multi-Agent Systems

Education

Purdue University

B.S. in Artificial Intelligence; Minor in Mathematics

Aug. 2024 – Expected Dec. 2027

West Lafayette, IN

Publications

- **Seunghyun Yoo.** *White Box Evidence Packages for Policy Audit Reports.* Second Workshop on Technical AI Governance Research (TAIGR), ICML 2026. Sole author. [PAPER](#) | [POSTER](#)
- Meliksah Yorulmazlar, Justice Sefas, **Seunghyun Yoo**, and Lulu Shao. *Feature Evolution in Diffusion Language Models.* Non-Autoregressive Language Models Workshop, COLM 2026. Poster. [PAPER](#) | [POSTER](#) | [LIVE DEMO](#)

Research Experience

Research Fellow

Supervised Program for Alignment Research; Mentors: Yuxiao Li and Di Wu

Sep. 2026 – Present

Remote

- Studying whether reading the internal states of interacting agents reveals influence that transcripts miss, in *Who's Steering Whom? Interpretable Influence and Equilibria in Human-Agent Systems.*
- Designed and ran pre-registered steer-and-read pilots on Qwen2.5 models: agent A is steered along a contrastive activation direction while agent B's residual-stream projection is read turn by turn.
- On 300 GSM8K problems, a probe on B's pre-answer activations predicted when B would adopt A's wrong answer (AUROC 0.73), outperforming an LLM judge reading the same transcript (0.56).

Undergraduate Researcher

Governance and Responsible AI Lab (GRAIL)

Jun. 2025 – Present

Purdue University

- Designed a [policy audit evaluation](#) across 60 AGORA cases, 600 reports, and ten evidence conditions, showing that audit reports can cite irrelevant internal model evidence while still reading as plausible; authored the resulting TAIGR paper.
- Built a [policy representation analysis](#) in Gemma 2 2B, from corpus preparation through SAE extraction, statistical testing, and causal interventions; identified 35 layer-24 features stable across five permutation seeds in 2,438 records.
- Developed [AGORA-RAGTruth](#), a policy hallucination benchmark with manual span annotations and governance tags, with mechanistic analyses separating retrieved-context use from parametric knowledge.
- Annotated and validated [AGORA policy documents](#) and redesigned the [Explore-AGORA graph](#) schema linking policy passages to risks, strategies, and institutional metadata.

Research Fellow

Algoverse AI Safety Fellowship; Supervisor: Justice Sefas

Feb. 2026 – Jun. 2026

Remote

- Analyzed SAE features in Dream-v0-Base-7B across six layers and 64 denoising steps, covering activation statistics, feature interpretation, and token-commitment visualizations.
- Identified recurring feature patterns: tracking masked tokens, activating near token commitment, and responding selectively to unmasked tokens.
- Compared token unmasking across math, natural-language, and code reasoning prompts; the resulting paper, *Feature Evolution in Diffusion Language Models*, was accepted to the Non-Autoregressive Language Models Workshop at COLM 2026.

Hackathon Projects

Ethogram [LIVE DEMO](#) | [CODE](#)

Oct. 2026

AI Village × Grove Research AI Swarm Dynamics Hackathon

- Built a pipeline that turns multi-agent logs into a plain-language behavior dictionary, a storyline of phases and turning points, and an interactive behavior atlas, with every claim linked to the events behind it.
- Validated behaviors with two independent LLM judges and blind detection by a different model; two-judge agreement reached $\kappa = 0.89$ and 0.86 on records of 20,932 and 67,969 events.

MARGIN Arena: Agents Under Resource Limits [PAPER](#) | [CODE](#)

Sep. 2026

AI Incident Response Sprint; Apart Research

Remote

- Designed and published on PyPI a Python environment for studying agent behavior under five independently configurable limits: tokens, time, verification calls, tool calls, and sessions.
- Integrated Inspect and ControlArena tasks with original scoring, a tamper-evident run journal, and optional activation capture for interpretability; validated resource accounting and stopping behavior with 50 software tests.
- Placed in the top 25% of more than 230 submissions.

Denoising Trajectories as Early-Warning Signals for Diffusion Language Model Control [PAPER](#) | [CODE](#)

Mar. 2026

AI Control Hackathon; Apart Research / Redwood Research

Remote

- Built SAE-based monitoring and steering for LLaDA-8B-Instruct, with executable shell benchmarks and tests under prompt and payload variation.
- Trajectory monitoring detected attacks at mean denoising step 9.60, versus 63.00 for output-only checks; early detection held under variation (10.53 versus 63.00).

Teaching

AI Study Leader

LikeLion US / OZ Lab

Sep. 2025 – Present

West Lafayette, IN

- Teach weekly sessions on LLM From Scratch, MiniArena, and Interpretability for Purdue's Korean student community.

Selected Posters and Presentations

- *Co-CoT: A Prompt-Based Framework for Collaborative Chain-of-Thought Reasoning*. RISE AI Conference, Oct. 2025. Poster and presentation; first author.
- *Factually Grounded and Human-Aligned RAG Systems for AI Governance and Policy*. Purdue URC, 2025. Poster.
- Internal GRAIL presentations: *Explore-AGORA 2* (Feb. 2026) and *DPO + Explore-AGORA* (Nov. 2025).

Honors and Recognition

- Full Member, Sigma Xi, The Scientific Research Honor Society, Jul. 2026.
- Travel Grant, Grove Research, 2026 (\$1,500).
- Travel Grant, TAIGR Workshop at ICML 2026 (\$1,000).
- Travel Grant, RISE AI Conference, University of Notre Dame, Oct. 2025.
- Purdue PLaCE Student Spotlight, 2024.

Technical Skills

Programming and ML Python, PyTorch, Hugging Face Transformers, Git

Language models RAG, diffusion models, model evaluation

Interpretability Sparse autoencoders, activation patching, probing, TransformerLens, NNSight, NDIF

Activities

AI Safety Purdue and ML@Purdue, general member; Purdue eSports, Overwatch Junior Varsity.