

# MAGNUS SAEBO

(914) 519-7079 ◇ m.saebo@columbia.edu ◇ linkedin.com/in/msaebo ◇ magnus-saebo.com

## EDUCATION

### Columbia University

MS in Computer Science, Thesis Track | GPA: 4.18/4.0

New York, NY

Sep 2025 - Dec 2026

### Cornell University

BA in Computer Science, Mathematics

Ithaca, NY

Jan 2019 - Dec 2022

## RESEARCH EXPERIENCE

### Research Fellow

June 2026 - Present

MATS (Mentor: Prof. Peter Henderson, Princeton University)

Berkeley, CA

- Leading the **Legal Gradual Disempowerment Observatory**, a framework that investigates how the US legal system would change with various AI models as judges at multiple court levels
- **Developed methodology for extracting behavioral differences between human and AI judges at scale**, revealing suspicious legal propensities for several frontier models and surfacing high-level insights on general model character
- Ran self-preference bias experiments investigating impact of human versus AI litigants on AI judges

### Graduate Researcher

Jan 2026 - Present

Zhuo Zhang Lab — Columbia University

New York, NY

- Developed an Inspect benchmark **CyberDualEval** and a novel weaponizability score to measure frontier model ability to correctly discern safe from risky dual-use cybersecurity requests given various user contexts with industry expert labels

### Graduate Researcher

Aug 2025 - Present

Advanced Research in Software Engineering Lab — Columbia University

New York, NY

- Leading work building **small code world models** to simulate RL environments optimized against red teaming agents to be **robust against reward-hacking** while out-performing real environments on training policies
- Co-lead development of **SWE-Spot**, a family of 4B-parameter repo-expert coding agents that outperform fine-tuned models up to 8x larger across multiple SWE tasks, for maximizing data privacy with marginal capability sacrifice

### Research Fellow

Aug 2025 - Jan 2026

Supervised Program for Alignment Research (SPAR)

New York, NY (Remote)

- Built OpenCode-based evaluation framework for measuring goal drift in coding agents; demonstrated that adversarial codebase comments can exploit model value hierarchies to override system prompt constraints in frontier models

### Graduate Researcher

Aug 2025 - Jan 2026

David Blei Lab — Columbia University

New York, NY

- Developed **Duel-Evolve**, an inference-time evolutionary optimizer that uses LLM pairwise self-preferences instead of scalar rewards, improving over comparable methods by 20% on MathBench and 13% on LiveCodeBench

## SELECTED PUBLICATIONS (\* Equal Contribution)

1. **CyberDualEval: Measuring Dual-Use Cyber Situational Awareness in Frontier Language Models.**  
Magnus Saebo, Francesco Piccoli, Maxwell Watson, Pranjali Thakur, Yangruibo Ding, Zhuo Zhang.  
*Under review at NeurIPS 2026.*
2. **Asymmetric Goal Drift in Coding Agents Under Value Conflict.**  
Magnus Saebo, Spencer Gibson, Tyler Crosse, Achu Menon, Eyon Jang, Diogo Cruz.  
*ICLR 2026 Workshop on Agents in the Wild: Safety, Security, and Beyond.* [2603.03456v1](#). **Spotlight, top ~10% accepted.**
3. **Inherited Goal Drift: Contextual Pressure Can Undermine Agentic Goals.**  
Achu Menon, Magnus Saebo, Tyler Crosse, Spencer Gibson, Eyon Jang, Diogo Cruz.  
*ICLR 2026 Workshop on Lifelong Agents: Learning, Aligning, Evolving.* [arXiv:2603.03258](#).
4. **SWE-Spot: Building Small Repo-Experts with Repository-Centric Learning.**  
Jinjun Peng\*, Magnus Saebo\*, Tianjun Zhong, Yi-Jie Cheng, Junfeng Yang, Baishakhi Ray, Simin Chen, Yangruibo Ding.  
*Under review at NeurIPS 2026.* [arXiv:2601.21649](#).
5. **BlindFL: Segmented Federated Learning with Fully Homomorphic Encryption.**  
Evan Gronberg\*, Liv d'Aliberti\*, Magnus Saebo\*, Aurora Hook.  
[arXiv:2501.11659](#).

## WORK EXPERIENCE

### Machine Learning Engineer

Jan 2023 - Aug 2025

Leidos

Arlington, VA (Remote)

- Co-authored federated learning framework, **BlindFL**, with provable security guarantees, reducing attack success to near 0% while outperforming similar methods on compute efficiency