MAGNUS SAEBO

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

EDUCATION

Columbia University — New York, NY
MS, Computer Science, Thesis Track | GPA: 4.18/4.0 — Sep 2025 - Dec 2026

Cornell University — Ithaca, NY
BA, Double Major in Computer Science and Mathematics — Jan 2019 - Dec 2022

RESEARCH EXPERIENCE

Research Fellow — June 2026 - Present
MATS 10.0 → 10.1 Extension (Mentor: Prof. Peter Henderson, Princeton University) — *Berkeley, CA*

- Leading the **Legal Gradual Disempowerment Observatory**, investigating how US law would change if frontier models decided ∼5k post-cutoff state and federal appeals, aggregating AI-human differences in values, reasoning, and outcomes
- Engineered an **investigator agent** that leverages small decision models to parse AI-human case data to generate validated hypotheses with FDR-corrected statistics at ∼$10/run; found Grok would meaningfully narrows workers' rights
- Co-developed framework for extracting legal values and **constructing legal character profiles**, mapping frontier models and human judges onto shared value axes; e.g. Grok follows rules more rigidly than humans while Sol is more adaptive
- Built pipeline **aggregating differences** found between human and AI judges at case level, amassing 10k+ quote-grounded, adversarially checked differences; revealed models decide questions courts left open at 4x a model-vs-model null

Graduate Research Assistant — Jan 2026 - Present
Zhuo Zhang Lab — Columbia University — *New York, NY*

- Led **CyberDualEval (NeurIPS '26)**, an Inspect benchmark of 425 real-world vulnerabilities across 28 CWE types, probing whether models correctly discern safe from risky cybersecurity tasks across various user contexts
- Introduced **Task Weaponization Score**, an expert-annotated metric for task risk, validated by showing that the success of exploits generated from a model's task output strongly correlates with the task's TWS score
- Found frontier models both over-refuse and over-comply, with safeguards unduly tracking surface cues over misuse risk

Graduate Research Assistant — 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 for training on SRE tasks; these world models are optimized against red-teaming agents to be **robust against reward hacking** when training policies
- Co-led training of **SWE-Spot** (arXiv:2601.21649), a **family of 4B-parameter repo-expert coding agents** that outperform fine-tuned models up to 8x larger across multiple SWE tasks, maximizing data privacy with marginal skill loss

Research Fellow — Aug 2025 - Jan 2026
Supervised Program for Alignment Research (SPAR) — *New York, NY* (Remote)

- Built **evaluation framework for measuring goal drift** in coding agents under value conflict; found that adversarial codebase comments can exploit model value hierarchies to override system prompt constraints for small frontier-lab models

Graduate Research Assistant — Aug 2025 - Jan 2026
David Blei Lab — Columbia University — *New York, NY*

- Co-led **Duel-Evolve** (2602.21585), test-time search guided by an LLM's own pairwise judgments; +20 pts on MathBench

WORK EXPERIENCE

Machine Learning Engineer — Jan 2023 - Aug 2025
Leidos — *Arlington, VA* (Remote)

- **Co-authored federated learning framework** based on homomorphic encryption, **BlindFL** (arXiv:2501.11659), with provable security guarantees, reducing attack success to near 0% while outperforming similar methods on compute efficiency
- Worked with **enterprise clients to design and deploy NLP and CV models** under tight compute and latency constraints, including U-Net anomaly detection under 100:1 class imbalance
- **Built production MLOps platform on Kubernetes** with automated drift monitoring and retraining across 4 deployment environments using MLflow, Apache Airflow, and Evidently AI

SELECTED PUBLICATIONS

1. **CyberDualEval: Measuring Dual-Use Cyber Risks in Frontier Language Models.**
   Magnus Saebo, Francesco Piccoli, Maxwell Watson, Pranjali Thakur, Yangruibo Ding, Zhuo Zhang.
   NeurIPS 2026, Evaluations and Datasets Track. Dataset.

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. arXiv:2603.03456. ***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. arXiv:2603.03258.