MAGNUS SAEBO

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

EDUCATION

Columbia University — New York, NY
MS in Computer Science, Thesis Track | GPA: 4.2/4.0 — Sep 2025 - Dec 2026

Cornell University — Ithaca, NY
BA in Computer Science, Mathematics | Major GPA: 3.7/4.0 — Jan 2019 - Dec 2022

SELECTED PUBLICATIONS (* Equal Contribution)

1. **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 Lifelong Agents: Learning, Aligning, Evolving. 2603.03456v1.

2. **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.
arXiv:2601.21649.

3. **Duel-Evolve: Reward-Free Test-Time Scaling via LLM Self-Preferences.**
Sweta Karlekar*, Carolina Zheng*, **Magnus Saebo***, Nicolas Beltran-Velez, Shuyang Yu, John Bowlan, David Blei.
ICLR 2026 Workshop on AI with Recursive Self-Improvement. arXiv:2602.21585.

4. **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.

5. **BlindFL: Segmented Federated Learning with Fully Homomorphic Encryption.**
Evan Gronberg*, Liv d'Aliberti*, **Magnus Saebo***, Aurora Hook.
arXiv:2501.11659.

RESEARCH EXPERIENCE

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 Research Assistant — Aug 2025 - Present
Advanced Research in Software Engineering Lab — Columbia University — *New York, NY*
- Co-leading 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 through a novel data generation and fine-tuning pipeline

Graduate Research Assistant — Jan 2026 - Present
Zhuo Zhang Lab — Columbia University — *New York, NY*
- Developing benchmark for evaluating LLM coding agent refusal accuracy on cybersecurity tasks, leveraging CTF challenges to measure model ability to distinguish malicious from defensive security queries

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

WORK EXPERIENCE

Machine Learning Engineer — Jan 2023 - Aug 2025
Leidos — *Arlington, VA* (Remote)
- Co-authored federated learning framework (arXiv:2501.11659) with provable security guarantees, reducing attack success to near 0% while outperforming similar methods on compute efficiency
- Optimized YOLO segmentation models for FPGA edge deployment using quantization, pruning, and knowledge distillation, decreasing latency by 56% and increasing throughput by 2.25x
- Built active learning pipeline pairing U-Net with BERT-based classifier, prioritizing samples where models disagreed to surface mislabeled data; reduced labeling cost by 29%
- Deployed NLP and CV models for 2 enterprise clients across $20M+ in contracts, including U-Net for anomaly detection with 100:1 class imbalance using tile-based synthetic data generation
- Built production MLOps platform on Kubernetes with automated drift monitoring and retraining across 4 deployment environments using MLflow, Apache AirFlow, and Evidently AI