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 Sep 2025 - Dec 2026

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

SELECTED PUBLICATIONS (* Equal Contribution)

1. **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 ICML 2026. arXiv:2601.21649.

2. **Asymmetric Goal Drift in Coding Agents Under Value Conflict.**
   Magnus Saebo, Spencer Gibson, Tyler Crosse, Achutha Menon, Eyon Jang, Diogo Cruz.
   Submitted to ICLR 2026 Workshop on Lifelong Agents (LLA).

3. **Duel-Evolve: Pairwise Preference Black-Box Optimization of LLM Responses.**
   Sweta Karlekar*, Carolina Zheng*, **Magnus Saebo***, Shuyang Yu, Nicolas Beltran-Velez, John Bowlan, David Blei.
   Submitted to ICLR 2026 Workshop on AI with Recursive Self-Improvement (RSI).

4. **Inherited Goal Drift: Contextual Pressure Can Undermine Agentic Goals.**
   Achutha Menon, **Magnus Saebo**, Tyler Crosse, Spencer Gibson, Eyon Jang, Diogo Cruz.
   Submitted to ICLR 2026 Workshop on Lifelong Agents (LLA).

RESEARCH EXPERIENCE

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 open-weight models up to 8x larger across multiple SWE tasks

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
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

Research Assistant Jan 2020 - Aug 2023
Peter McMahon Lab — Cornell University *Ithaca, NY*

- Developed model compression and modular scaling techniques for training physical neural networks, outperforming prior PNN methods with 3x fewer parameters on image classification

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

SKILLS

Research Interests	Agentic AI, AI for Code, AI Safety/Alignment, Inference-Time Scaling, Model Compression
ML Frameworks	PyTorch, TensorFlow, Hugging Face Transformers, Anthropic/OpenAI API, LangChain
Infrastructure	AWS, GCP, Docker, Kubernetes, MLflow, Airflow, Prometheus/Grafana, Slurm, SGLang, vLLM