Stewart Slocum

https://stewyslocum.com | Google Scholar | sslocum3@mit.edu

Education

Ph.D. Computer Science, *Massachusetts Institute of Technology* **August 2022 - Present**
Advised by Professor Dylan Hadfield-Menell

B.S. Computer Science, Applied Math and Statistics, *Johns Hopkins University* **August 2017 - May 2021**
Advised by Professor Rene Vidal

Research Statement

I'm developing tools to ensure increasingly powerful LLMs and LLM agents remain safe and aligned with human values. My work spans two main areas:

1. Adversarial defenses and evaluations (e.g. Inverse Prompt Engineering, Model Manipulation Attacks Enable Rigorous Evals of LLM Unlearning)
2. Preference learning and alignment (e.g. Diverse Preference Learning, A Bayesian Truth Serum for Scalable Oversight)

I prefer simple methods that scale. I like to use math to distill clear problem formulations and as a source of ideas. However, I think it's important that my research is primarily driven by a close qualitative and quantitative understanding of empirical behavior.

Awards & Service

NSF GRFP Fellowship
Program Committee, Special Alignment Track, AAAI 2025

Publications, Preprints, and Conference Presentations

[1] **Stewart Slocum**, Dylan Hadfield-Menell. "Inverse Prompt Engineering for Task-Specific LLM Safety." *Under review at ICLR, Best Paper Runner-Up at AAAI Workshop on Responsible Language Models* (2024).

[2] **Stewart Slocum**, Asher Parker-Sartori, Dylan Hadfield-Menell. "Diverse Preference Learning for Capabilities and Alignment." *Under review at ICLR* (2024).

[3] Che et al. "Model Manipulation Attacks Enable More Rigorous Evaluations of LLM Unlearning." *Neurips Safe Generative AI Workshop* (2024).

[4] Casper et al. "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback." *Transactions on Machine Learning Research* (2023).

[5] Aditya Chattopadhyay, **Stewart Slocum**, Benjamin D. Haeffele, Rene Vidal, Donald Geman. "Interpretable by Design: Learning Predictors by Composing Interpretable Queries." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[6] Lauren Wheelock, **Stewart Slocum**, Jorma Gorns, Sam Sinai. "Risk-Adjusted Selection for Validation of Sequences in AAV Design Using Composite Sampling." *American Society of Gene & Cell Therapy Conference* (2021).

[7] Sinai et al. "AdaLead: A simple and robust adaptive greedy search algorithm for sequence design." *arXiv preprint arXiv:2010.02141* (2020).

[8] Mulligan et al. "Designing peptides on a quantum computer." *bioRxiv* (2019).

Experience

Algorithmic Alignment Group, MIT *PhD Student* **August 2022 -**

Working on post-training methods for LLM safety, alignment, and robustness.

- Inverse Prompt Engineering – proposes a task-specific (or allow-list) approach to AI safety, in contrast to existing deny-list guardrails. IPE derives robust safety guardrails from prompt engineering data that is already on hand, without using example jailbreaks. IPE achieves state-of-the-art jailbreak robustness against human and automated adversaries.
- Diverse Preference Learning for Capabilities and Alignment – using a social choice theory analysis of RLHF/DPO, we identify the KL-divergence regularizer as a major cause of mode collapse in aligned LLMs. Diverse Preference Learning realigns the LLM output distribution to population-level preferences. Compared to existing diversity-boosting procedures, DPL improves diversity-quality tradeoffs across a range of benchmarks, best-of-N problem solving on hard instances, and gender representation in story-writing.
- Model Manipulation Attacks Enable More Rigorous Evaluations of LLM Unlearning – we establish 1) adversarial attacks elicit hidden harmful capabilities present after LLM unlearning 2) fine-tuning and latent-space attacks can be used to estimate these vulnerabilities with significantly less compute than input-space attacks.
- Training for Test-Time Scaling Project – ongoing project aiming to align training objectives and test-time compute scaling methods. As a complementary direction to "learning-to-think" approaches like OpenAI's o1 or STaR, we propose a novel training objective to optimize LLMs for best-of-N sampling (which remains an important intermediate step in parallel decoding used in tree search). The goal is to train LLMs to take creative, diverse, and complementary problem-solving strategies instead of repetitive ones as they often currently do.
- A Bayesian Truth Serum for Scalable Oversight – ongoing project building a scalable oversight system for LLM alignment using the Bayesian Truth Serum and game-theoretic mechanism design tools.

Vision Lab, JHU *Research Assistant* **October 2020 - July 2022**

- Developed an information-theoretic framework for building interpretable machine learning models. This resulted in the IEEE publication Interpretable by Design: Learning Interpretable Predictors by Composing Interpretable Queries.
- Led a project developing physics-inspired, geometrically aware optimization algorithms for deep learning. Contributions included proposing new algorithms, proving convergence rates, and performing neural network training experiments.

Dyno Therapeutics, *Machine Learning Intern* **June - August 2020 Full-time, - March 2021 Part-time**

Used generative models and reinforcement learning to optimize viral vectors used in gene therapy. Contributed to two novel AAV viral vectors with modified VP1 regions, targeting the nervous system and muscle tissue. This work led to a methods paper AdaLead: A simple and robust adaptive greedy search algorithm for sequence design, popular benchmarking environment FLEXS, and oral presentation at ASGTC 2021 (top 25% of submissions).

NASA Goddard Space Flight Center, *Quantum Computing Research Intern* **Summer 2019**

Developed a quantum annealing solver for protein design on the D-Wave quantum computer. We used our method to create the first quantum-designed peptide. Designing Peptides on a Quantum Computer presented at the American Physical Society's 2022 March Meeting.

NASA Goddard Space Flight Center, *Virtual Reality Software Engineering Intern* **Summer 2017 and 2018**

Built VR visualization tools for scientific data analysis in astophysics and planetary science. See press coverage and journal publication.