ITAMAR HAGAY PRES

646-589-1150 | presi@umich.edu | personal website

Education

UNIVERSITY OF MICHIGAN ANN ARBOR, College of Engineering Ann Arbor, MI
BSE in Computer Science with minor in Mathematics *January 2025*

Research Publications

Lee, Bai, **Pres**, et al.(2024) A Mechanistic Understanding of Alignment Algorithms: A Case Study on DPO and Toxicity *The Forty-first International Conference on Machine Learning* ***(Oral: top 1.5 % of submissions)***

- Discovered how toxicity is represented in pre-trained language models (GPT2, Llama2-7b)
- Studied the internal mechanisms of a popular alignment algorithm (Direct Preference Optimization)
- Discovered that DPO does not remove toxicity representations but rather suppresses them
- Uncovered why "aligned" models can be jailbroken

Pres, et al.(2024) Towards Reliable Evaluation of Behavior Steering Interventions in LLMs *The 38th Conference on Neural Information Processing Systems - Workshop on Foundation Model Interventions* ***(Oral: Top 6 of total submissions)***

- Established four criteria for effective behavioral steering metrics and highlighted current metric shortcomings
- Developed new steering metric showing over-reported success of control method Contrastive Activation Steering

Park*, Lubana*, **Pres**, et al.(2024) Competition Dynamics Shape Algorithmic Phases of In-Context Learning - *Under Review at The 13th International Conference on Learning Representations*

- Provided mechanistic analysis showing transformers solve in-context learning with multiple competing algorithms
- Developed quantifiable techniques to measure neuron memorization of training data in transformers

Research Experience

Krueger AI Safety Lab Internship (University of Cambridge) *May 2024 - Dec 2024*

Mentored By ***David Krueger***

- Developed framework for multiple constraint satisfaction of LLM generations at inference-time
- Explored generalizability of inference-time interventions such as Contrastive-Activation-Steering
- Developed novel benchmark to measure the success of behavioral steering interventions

Physics of Intelligence at the Center for Brain Science (Harvard University) *Sept 2024 - Present*

Mentored By ***Hidenori Tanaka, Ekdeep Singh Lubana***

- Rigorously studying LLM behaviors in thoughtfully constructed toy-settings
- Trained LLMs of different architectures to perform ICL tasks and studying differences in mechanisms
- Researching the mechanisms LLMs use to handle underspecified queries
- Published findings at ICLR

University of Michigan Language and Information Technologies Lab *October 2023 - Present*

Supervised by ***Rada Mihalcea***

- Conducted research in mechanistic interpretability, with a focus on safety alignment algorithms
- Published research findings at ICML
- Researched Vision Language Model interpretability as well as In Context Learning mechanisms

Mechanistic Interpretability Research (SERI MATS) *June 2023 – Aug 2023*

Mentored by ***Neel Nanda***

- Analyzed and visualized high-dimensional LLM operations to generate hypothesis of LLM internal mechanisms
- Tested hypotheses of attention head superposition phenomenon in GPT2-XL
- Devised precise casual interventions to validate hypotheses under fact extraction contexts

Machine Learning Safety Scholars *June 2022 – August 2022*

- Completed coding, written, and discussion assignments from MIT's Intro to Machine Learning, UM's Deep Learning for Computer Vision, and NYU's Deep Learning courses
- Distilled foundational machine learning papers into readable summaries
- Implemented safety techniques including adversarial and black swan robustness as well as interpretability

Grants

Effective Altruism Infrastructure Grant, Full stack Development *June 2022 – September 2022*

- ∗ Developed prototype for web-app that increases transparency of unlikely global risk assessments
- ∗ Designed NoSQL MongoDB database structure for back-end
- ∗ Created interactive front-end using React.js
- ∗ Developed API connecting database to front-end using Node.js and Express.js