

# George Ingebretsen

✉ [george.ingebretsen@gmail.com](mailto:george.ingebretsen@gmail.com) | 🌐 [Website](#) | 🔗 [LinkedIn](#) | 🎓 [Google Scholar](#) | ☎ 425-420-6248

## Education

**University of California, Berkeley**

Fall '22 – Spring '25

*B.S. in Electrical Engineering and Computer Science*

## Experience

**Center for AI Safety**

Spring '25 – Now

*Special Projects Associate and Technical Executive Assistant*

- Public Engagement Team – Directed branding, tone development, and external agency management for an ambitious public engagement initiative; led aspects of early-stage strategy and stakeholder outreach
- Co-authored [A Definition of AGI](#): coordinated outreach and feedback process across 20+ authors
- Organized CAIS “AI for Humanity” forum and panel (Yoshua Bengio, Max Tegmark, Helen Toner, Dwarkesh Patel)
- Executive Assistant to the Director – Managing Dan Hendrycks’ calendar, email triage, and task prioritization; drafting correspondence with government officials, lab leadership, and academic collaborators

**Center for Human-Compatible AI (CHAI)**

Winter '24 – Spring '25

*Research Intern – Advised by Micah Carroll (now OpenAI)*

- Investigating motivated reasoning in LLM chains of thought: after RL training that incentivizes harmful behaviors, models produce CoT that rationalizes harmful outputs by appealing to safety training. Fine-tuned Llama-70B and analyzed CoT traces to characterize when this rationalization occurs
- Reproduced [Alignment Faking in Large Language Models](#) (Anthropic) on open-source models (Mistral-24B, Llama-70B, Nemotron-70B), measuring compliance gaps across models

**Adversarial Robustness Research**

Summer '23 – Fall '24

*Research Collaborator – Advised by Kellin Pelrine (McGill, Mila, Far AI)*

- ACL 2025 Findings and two NeurIPS 2024 workshop papers (co-first author) on vulnerabilities in frontier models
- Created [MultiBench](#), a 1,100-prompt multi-modal safety benchmark that measures safety generalization across structurally different but semantically equivalent attacks (e.g., single-image vs. multi-image, text vs. cipher-encoded)
- Showed robustness to one attack structure does not imply robustness to other structures

**Singular Learning Theory Research**

Summer '24 – Fall '24

*Research Collaborator – Advised by Lucius Bushnaq (Apollo Research)*

- Developed a Hessian-rank method for bounding the local learning coefficient (LLC) using 2nd-order Taylor expansion of behavioral loss; validated on toy models (transformers, MLPs) using Lanczos algorithm in Jax
- Found 98.8% of parameter directions are “free” in a modular addition transformer; matrix factorization bounds matched analytical predictions

**AI Interpretability Research**

Winter '23 – Summer '24

*Research Collaborator – Advised by Arun Jose, Alex Turner (Google DeepMind)*

- Pre-trained a GPT-2 transformer, then adversarially trained against its own linear probes to test probe robustness
- Models temporarily evade probes while maintaining task performance, but freshly-trained probes recover features given sufficient training steps in all tested settings, implying that linear representations remained

## Publications

**[A Definition of AGI](#)**

arXiv Preprint ([arXiv](#))

D. Hendrycks, D. Song, ..., **G. Ingebretsen**, ..., M. Tegmark, G. Marcus, E. Schmidt, Y. Bengio (33 authors)

**[Emerging Vulnerabilities in Frontier Models: Multi-Turn Jailbreak Attacks](#)**

arXiv Preprint ([arXiv](#))

T. Gibbs\*, E. Kosak-Hine\*, **G. Ingebretsen\***, J. Zhang, J. Broomfield, S. Pieri, R. Iranmanesh, R. Rabbany, K. Pelrine

**[Decompose, Recompose, and Conquer: Multi-modal LLMs are Vulnerable to Compositional Adversarial Attacks in Multi-Image Queries](#)**

NeurIPS 2024 RBFM, NeurIPS 2024 Red Teaming GenAI ([OpenReview](#))

J. Broomfield, **G. Ingebretsen**, R. Iranmanesh, S. Pieri, R. Rabbany, K. Pelrine

**[The Structural Safety Generalization Problem](#)**

ACL 2025 Findings ([ACL Anthology](#))

J. Broomfield\*, T. Gibbs\*, E. Kosak-Hine\*, **G. Ingebretsen\***, T. Nasir, J. Zhang, R. Iranmanesh, S. Pieri, R. Rabbany, K. Pelrine

## Technical Skills

Python, Java, JavaScript,  $\LaTeX$ , Git, PyTorch, Jax, SLURM, WandB, NumPy, Pandas, React, Node.js