SANG T. TRUONG

ai.stanford.edu/~sttruong – sttruong@cs.stanford.edu

EDUCATION **Stanford University** *Ph.D. '27 & M.S. '24 in Computer Science, M.S. '25 in Statistics*
Advisors: Sanmi Koyejo & Nick Haber @ SAIL, StatsML, GSE
Stanford Data Science Scholar, Stanford Human-Centered AI Fellow

DePauw University *B.A. '21 in Computer Science, Economics, & Computational Chemistry*
Advisors: Jeff Gropp, Humberto Barreto, Todd Neller @ Management Fellows Program

RESEARCH I develop foundational and applied methods for AI Measurement Sciences, drawing on tools from probabilistic machine learning and measurement theory. My research supports the scientific development of safe AI systems that serve people across diverse backgrounds and needs. Below are my selected papers:

1. **Truong** et al. Guardians of the Measurement: Information Design for Robust AI Evaluation *Preprint'26*

2. **Truong** et al. Reliable & Efficient Evaluation via Factor Model. *Preprint'26*

3. **Truong** et al. Generalized Neural Scaling Laws via Measurement Theory. *Preprint'26*

4. **Truong** et al. Reliable and Efficient Amortized Model-based Evaluation. *ICML'25*
   - *Media cover:* Stanford Engineering, Stanford's Center for Research in Foundation Models
   - *Invited talk:* Stanford Data Science '25, Stanford HAI+Education Summit '25, NCME Seminar '25, Virtue AI Seminar '24, Guest Lecture at Stanford CS120: AI Safety '24, Stanford Psychometrics Seminar '24
   - *Implementation:* Code: sangttruong/reeval ☑, Data: stair-lab/reeval ☑, HELM's Evaluation Toolkit ☑

5. **Truong** et al. Fantastic Bugs and Where to Find Them in AI Benchmarks. *NeurIPS'25*
   - *Media cover:* Stanford HAI ☑
   - *Implementation:* Code: sangttruong/fantastic-bugs ☑, Data: stair-lab/fantastic-bugs ☑

6. **Truong**, Haupt, Koyejo. Machine Learning from Human Preferences. *The Living Textbook '25*
   - *Corresponding Course:* Stanford CS329H: Machine Learning from Human Preferences ☑(Fall'23-25)
   - *Invited talk*: Stanford CORES Symposium'25
   - *Implementation:* Source: stair-lab/mlhp ☑

7. **Truong** et al. Finetuning and Comprehensive Evaluation of Vietnamese Large Language Models. *NAACL'24.*
   - *Media cover:* The New York Times, Stanford Human-centered AI, and Stanford AI Lab Blog Post
   - *Invited talk:* Asia Tech x Singapore Summit'25, Stanford Data Science '24, FPT AI Research Vietnam '24
   - *Implementation:* HELM Leaderboard: crfm.stanford.edu/helm/villm ☑, Code: stair-lab/villm ☑, Data: stair-lab/villm ☑

EXPERIENCE
1. Microsoft Research Asian in Beijing – *Research Intern '26, science of AI evaluation*

2. Microsoft Research in New York City – *Research Intern '25, science of AI evaluation*

3. Virtue AI – *Research Intern '24, reliable and efficient amortized model-based evaluation*

4. Emergence AI – *Research Intern '24, efficient fine-tuning and evaluation safety large language models*

5. Google – *Software Engineering Intern '21, developed hyperSpec software for spectral analysis*

6. Community Health Network – *Data Science Intern '20, data-driven decision support*

7. Cummins – *Data Science Intern '19, data-driven decision support*

SELECTED AWARDS *Paper Awards:* NeurIPS'23 Best Paper Award, National Security Agency's Best Paper Award'24, ICLR'25 Workshop Oral, ICLR'25 Workshop Spotlight, NeurIPS'25 Spotlight

Grants*: Stanford HAI Seed Grant '24 ($40,000), Google-HAI Grant '24 ($90,000), Microsoft Accelerate Foundation Models Research Grant '23 ($80,000)

Fellowships: Emerson Consequential Scholarship '24, Cardinal Ventures DeepTech Fellowship '24, Stanford School of Engineering Fellowship '21

Early Awards: The Ferid Murad Medal '21 [Press], Robert Thomas Award in Computer Science '21, Bruce Long, Frank Carlton, and Randal Wilson Awards in Economics '19 '20 '21, John Ricketts Award in Chemistry '21, American Chemical Society Award '18, Silver Medal in Biochemistry @ Intel ISEF '16, Bronze Medals in Chemistry @ National Olympiad '15 '16

SERVICES & LEADERSHIP *Workshops Organized:* Convening on AI Construct Validity '25 (sponsored by Schmidt Science and MacArthur Foundation), NAACL'25/AAAI'26 Language Models for Underserved Communities Workshop series (founder), ICLR'24/ICML'24 Data-centric Machine Learning Research Workshop, Stanford Computing & Society group '22-'25 (lead)

Reviewing: NeurIPS '22-'25, ICLR '23 '25, ICML '22, AAAI '22, Stanford's Generative AI for the Future of Learning '23

Mentorship: Collaborated with and mentored undergraduate and graduate students from Stanford, UoT, UC Berkeley, and Oxford, many of whom have gone on to pursue research careers in graduate schools (UCLA, NUS, Max Planck) or industry