Stanford

Nhi Ngoc Truong

- Research Intern, Program-Koyejo, O.
- Staff, Program-Koyejo, O.

Bio

BIO

Nhi is an undergraduate researcher in the Stanford Trustworthy AI Research (STAIR) Lab, advised by Professor Sanmi Koyejo. She studies how we measure the reliability of AI systems, with a focus on the validity of AI benchmarks, meaning whether an evaluation really captures what it claims to and when its results can be trusted. Her current work looks at the assumptions built into widely used language model benchmarks and how to test them more carefully. Her earlier research on language models has appeared at leading natural language processing venues, including co-first-authored work at NAACL 2024 and an ICLR 2024 workshop. Outside of research, she plays in the symphonic band, badminton, and is taking up rock climbing.