Between Rigor and Reality:
How AI Safety Benchmark Developers Understand Benchmark Use and Maintenance

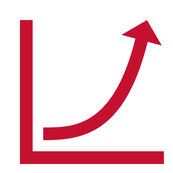
Presented by Megan Li

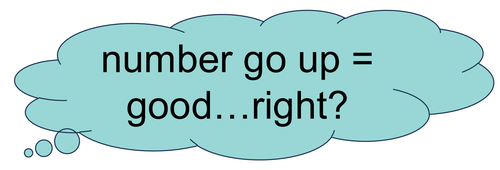
Work in collaboration with Ningjing Tang, Chris Stewart, Hoda Heidari, and Hong Shen

the motivation.

Safety benchmarks are thought of as the primary tool for measuring and ensuring the safety of genAI systems, yet growing evidence suggests that they fail to meet the needs of real-world safety evaluation.

a combination of a curated dataset and scoring metric to evaluate model behavior in risk-relevant scenarios





known challenges of developing and maintaining safety benchmarks.

During development

- Establishing construct validity
- Achieving satisfactory inter-rater reliability

Post-publication/deployment

- Concept shift
- Distribution shift
- Data leakage, saturation, Goodhart's Law

But how do benchmark developers understand and experience these challenges?

RQ1: What needs do benchmark developers perceive safety benchmarks are *intended* to meet? How do they compare to the needs they perceive safety benchmarks to meet *in practice*?

RQ2: How do they currently interpret and/or address challenges that contribute to gaps between safety benchmarks and real-world needs?

method.

We conducted five semi-structured interviews with practitioners who reported having experience developing or maintaining safety benchmarks for genAI, focusing on (1) how they assess whether a benchmark remains aligned with its intended construct and (2) what triggers benchmark updates.

ID	Professional Context	Focus
P1	Industry	Evaluation Methodology
P2	Academia	Mental Health
P3	Industry*	Bias, Toxicity
P4	Industry	Human well-being
P5	Industry*	Pluralism

findings.

There was a striking gap between how participants *intended* their benchmarks to be used and how they perceived their benchmarks were used *in reality*. Two challenges recurred across interviews:

1) Achieving and interpreting inter-rater reliability

Some participants viewed the difficulty of achieving inter-rater reliability as a barrier to overcome, while others viewed the difficulty of achieving inter-rater reliability as a reflection of the inherent pluralism of safety.

"reliability…is a necessary but not sufficient condition for validity."

"safety is definitely a subjective thing"

2) Acknowledging threats to external validity

Participants acknowledged well-established threats to external validity, but described lacking the resources, incentives, and/or signals to address them.

maintaining a safety benchmark involves *"playing catch up with [online] communities."*

"the only moment in which [a benchmark prompt] is valid is when you have not shared that prompt with anyone"

takeaways.

The reality of safety evaluation may require a bricolage approach.

Our findings suggest that the limitations of public safety benchmarks stem in part from their disconnection from the real-world feedback loops that keep evaluation relevant. Thus, we propose a *bricolage* approach to safety evaluation that may begin with a rigorous theory, but is iteratively refined bottom-up in concrete deployment contexts.

A shared agenda between industry and academia to improve the practice of safety evaluation.

- Methodological transparency on internal safety evaluations
- Closing the feedback loop of benchmark development
- Collaborative standard sharing on benchmark updates and retirement

Check out our paper!