Efficient Benchmarking for Agent Evaluations

Summary

What: Apply the "Fluid Benchmarking" methodology (Item Response Theory + Adaptive Testing) to expensive, agentic AI safety benchmarks (e.g., OS-HARM, Agent-SafetyBench).

Motivation: Agentic evaluations are crucial for assessing catastrophic risks, but they are currently very expensive, noisy, and rely on simple metrics (like success rates) that saturate quickly. The recent Fluid Benchmarking paper (Hofmann et al., 2025) showed massive efficiency and validity gains on static benchmarks by adapting the evaluation to the model's ability level. **My guess is that applying this to agentic safety evals is low-hanging fruit.** We want to see if we can get a clearer, more stable signal on loss-of-control risk while drastically reducing evaluation costs.

The non-summary

Motivation and Background

We are increasingly evaluating models in complex, multi-step agentic environments (like simulated operating systems or web environments). These evals are great for realism, but they have serious limitations:

1. **They are expensive and slow.** Running agents in sandboxed environments takes significant compute and time (e.g., OS-HARM reports ~$53 and 5 hours for 150 tasks on a small model). This severely limits how often we can run these checks during training.
2. **The metrics are noisy and limited.** Most benchmarks just report aggregate success or safety rates. This treats all tasks equally, ignoring that some safety failures are trivial and others are critical. It also makes it hard to distinguish progress when benchmarks saturate.

The "Fluid Benchmarking" paper offers a compelling alternative. They showed that applying psychometric techniques, specifically Item Response Theory (IRT) and Computerized Adaptive Testing (CAT), can dramatically improve standard benchmarks. They achieved higher validity and lower variance on MMLU using **50× fewer items**.

They do this by:

- **Using IRT:** Estimating the difficulty and discrimination of each item, and scoring models based on a latent "ability" rather than raw accuracy.

- **Adaptive Testing:** Dynamically selecting the next item that is most informative for the model's current estimated ability.

While IRT has been applied to some static safety benchmarks (e.g., AIR-Bench), it has not been systematically applied to complex agentic environments focusing on loss-of-control (unintended model misbehavior, dangerous side-effects). This project aims to bridge that gap.

Project Directions

The project will utilize the existing `allenai/fluid-benchmarking` codebase as a starting point. The plan is deliberately open-ended, focusing on exploring the following interconnected research directions rather than a rigid sequence of steps.

1. Adapting IRT Modeling for Agentic Tasks

The first major hurdle is making IRT work for agents. This involves gathering evaluation data across many models on 1-3 agentic safety benchmarks (e.g., OS-HARM, Agent-SafetyBench) and figuring out how to model it.

Key questions:

- **Defining "Items" and Scoring:** How do we treat a multi-step agentic trace? This is non-trivial. We will likely need to move beyond binary (safe/unsafe) to ordinal scales (e.g., {safe+successful, safe+failed, minor infraction, major unsafe action}).
- **IRT Model Choice:** Are standard models sufficient, or do we need more complex ones? For example, we might explore Graded Response Models for ordinal scores, or Multidimensional IRT (MIRT) to separate "task completion ability" from "safety compliance ability."

2. Analyzing Benchmark Quality (Psychometric Health)

Once we have fitted IRT models, we can analyze the benchmarks themselves through a psychometric lens. I'm interested in using this to understand where current evals are falling short.

We aim to identify:

- **Difficulty Gaps and Saturation:** Do the benchmarks have enough difficult items to distinguish between frontier models? Or are they saturated at the high end?
- **Low Discrimination Items:** Which scenarios fail to differentiate between models of different safety levels? (E.g., tasks where outcomes are random or where LLM-judge noise dominates).
- **Redundancy:** Are there clusters of scenarios measuring the exact same underlying safety failure mode?

3. Implementing Adaptive Testing and Measuring Validity

The core of the Fluid approach is adaptive testing. We will implement this and measure the impact on efficiency and validity.

- **Adaptive Testing Implementation:** Can we implement dynamic item selection (CAT) to significantly reduce the number of scenarios needed to estimate a model's latent "safety ability" with high confidence? How much variance does this reduce compared to random sampling?
- **Cross-Benchmark Predictive Validity:** We can use the IRT-estimated latent "safety abilities" to study predictive validity more rigorously. Does a high safety ability on OS-HARM predict safety ability on Agent-SafetyBench? Do agentic abilities correlate with static safety benchmarks? (My intuition is the correlation might be weak, which would be an interesting result).

Potential Challenges and Backup Plans

1. **Challenge:** Difficulty obtaining enough diverse model evaluations on agentic benchmarks to fit stable IRT models (the "cold start" problem). Running agentic evals is expensive.
    - **Backup:** Focus on 1-2 benchmarks where data is available. If necessary, allocate time/budget to generating new evaluation traces for a diverse set of smaller open-source models (e.g., 1B-7B parameters).
2. **Challenge:** Defining "items" and scoring in agentic environments is harder than in QA. Ambiguity in complex traces can introduce noise into the IRT model.
    - **Backup:** Start with benchmarks that have clearer, rule-based safety checks. Simplify scoring to binary (Safe/Unsafe) initially, and gradually introduce more complex ordinal scoring if time permits.
3. **Challenge:** IRT assumptions (like unidimensionality) may not hold well for complex agentic safety, which covers diverse risks.
    - **Backup:** Focus the analysis on specific sub-scales or risk categories (e.g., "tendency for unintended side-effects") where unidimensionality is more plausible. Exploring MIRT is a stretch goal.

Scope and Ambition

Least ambitious version: Successfully fit IRT models to 1-2 agentic safety benchmarks using existing data and binary scoring. Provide a detailed psychometric analysis of these benchmarks (identifying noisy items, difficulty gaps) and demonstrate potential efficiency gains using adaptive testing simulations.

Most ambitious version: Develop a robust, generalized framework for Fluid Benchmarking across multiple agentic environments using sophisticated IRT models (MIRT or Graded Response Models). Implement and validate adaptive testing, demonstrating significant cost/time

reductions and improved reliability. Conduct a cross-benchmark predictive validity study, establishing which benchmarks are most informative for loss-of-control risk, and release an adaptive testing toolkit.

Output

We aim to produce a research paper detailing our methodology and findings, suitable for submission to an AI safety or ML conference/workshop (e.g., NeurIPS, ICML). Additionally, we will create a public GitHub repository containing:

1. Extensions to the `fluid-benchmarking` codebase for agentic evaluations.
2. The trained IRT parameters (the "psychometric map") for the analyzed safety benchmarks.
3. Analysis notebooks and documentation.

We will also write a blog post summarizing key findings (e.g., LessWrong or the Alignment Forum).

Theory of change

This project aims to improve our ability to measure and mitigate loss-of-control risks in advanced AI systems by improving the tools we use for evaluation:

1. **More Efficient Safety Evaluations:** By drastically reducing the cost and time required for agentic evaluations, we enable more frequent testing throughout the development lifecycle (e.g., during pre-training) and broader analysis across the model ecosystem. This accelerates the iteration speed of safety research.
2. **Higher Quality Benchmarks:** Applying psychometric rigor helps identify flaws, noise, and saturation in safety benchmarks. This ensures that our metrics are actually measuring what we care about (loss-of-control risk), rather than artifacts of the evaluation setup or just general capability.
3. **Better Understanding of Safety Capabilities:** Moving from raw accuracy to latent "safety abilities" provides a clearer, more stable signal of a model's propensity for loss-of-control, facilitating better tracking of genuine safety progress.

Risks and downsides (externalities)

1. **Over-reliance on Metrics/Oversimplification:** IRT often assumes a single underlying ability. Safety is likely multidimensional. There is a risk that IRT-based abilities, while better than raw accuracy, are still treated as the ground truth, leading us to oversimplify the risk landscape. We will mitigate this by carefully analyzing model fit and clearly stating the limitations of the latent trait approach.

2. **Benchmark Gaming:** If adaptive testing methodologies become standard, models might eventually learn to game the adaptive procedure itself (e.g., intentionally failing early items to be routed to easier tasks). This seems unlikely with current models but is worth monitoring.

Acknowledgements

This project is a direct follow-up to Fluid Language Model Benchmarking (Hofmann et al., 2025) and builds on the open-source codebase provided by the Allen Institute for AI. It is also informed by related work in applying IRT to LLM evaluation (e.g., Reeval, PSN-IRT) and the extensive work done by the creators of agentic safety benchmarks (e.g., OS-HARM, Agent-SafetyBench).

Roles: Team members may specialize based on strengths, potentially splitting focus between (1) adapting the Fluid codebase and running IRT analyses (psychometrics/ML/statistics focus) and (2) integrating with agentic benchmark environments and running evaluations (software engineering/evals focus).We may also split team members independently across different benchmarks, depending on their workstyles.

Skill requirements
Required:

- Strong Python programming skills (PyTorch or similar ML frameworks).
- Experience with large language models and their evaluation.
- Familiarity with basic statistics and data analysis.

Recommended:

- Experience running agentic evaluations or working with complex simulation environments (e.g., Docker, VMs, simulators).
- Familiarity with AI safety concepts, particularly agent alignment and loss-of-control risks.
- Good scientific writing and communication skills.

Nice to have:

- Familiarity with psychometrics or Item Response Theory (IRT).
- Experience with large-scale data processing and managing compute resources.
- Experience with open-source collaboration and Git version control.