

Devising Effective Benchmarks

Nancy Vigil
Independent

Ashish Rai
NYU Courant

The AI safety community has been concerned about the limited landscape for quality evaluations for quite some time. Many have shared their concerns about the ability of current benchmarks and datasets to reliably evaluate AI. The common weaknesses in benchmarks today are their narrow scope, gameability, errors, inaccuracies, and lack of real-world applicability. Although there has been some progress, we still lack any comprehensive standards to understand AI's safety, progression of capabilities, and risks.

1. Problem overview

The US AI safety institute has begun making agreements with top AI labs to begin researching AI safety testing and evaluation^[1]. They have said, "Safety is essential to fueling breakthrough technological innovation." Anthropic has recently started an initiative to fund independent third-party evaluators^[2]. AI researchers, government officials, and the public are calling for more research into evaluations. These initiatives demonstrate the urgent need for better evaluations. Many have noted errors in currently available evaluations and the metrics they claim to measure. The AI safety community has been concerned about the limited landscape for quality evaluations for quite some time. Our solution is to build a startup focused on this issue.

Our startup aims to create the highest quality and most comprehensive benchmark in the field. Through this new benchmark, we will be able to evaluate the safety, capabilities, and risks of superintelligent AI. We strive to be a responsible and safety-centered company focusing on more than just testing school subjects. We delve into the ways that real people interact with a system, with the promise to consider intersectionality every step of the way.

2. Your solution

Our first benchmark will be centered on safety in the existing context of the use of frontier and non-frontier (read: weaker but still in use) LLMs. This will serve as a foundation on which all other safety benchmarks will be built incrementally. For the first safety benchmark, we propose to evaluate for a model's safety in regards to algorithmic biases. Algorithmic biases are one of the ways that AI has come to cause unintended harm to vulnerable populations. With the consultation of experts in the Algorithmic Justice League and other equity-conscious organizations, we will set the standard for evaluating any superintelligent AI.

Benchmark Creation

To uncover algorithmic biases, we plan to create adversarial prompts to evoke underlying biases in the model. For every scenario (from trivial sentence completion to real-world decision making), we generate adversarial prompts to evoke biased

Devising Effective Benchmarks

responses from models. For every adversarial prompt, we create a corresponding bias-aware prompt (including few-shot examples of unbiased answers or a short explanation of bias-awareness). The adversarial prompts are collated to make datasets for defined use cases or domains (Study Helper agent or Healthcare Assistant). These datasets are combined to make a benchmark.

Evaluation

For every model that needs to be benchmarked for safety, we then evaluate the response for biasness, preferably with a strong frontier model like GPT-4.

Remediation

For models which rate lower on the safety benchmarks can be trained with the bias-aware dataset of the corresponding benchmark.

We also plan to employ other ideas such as using mechanistic interpretability to remove learned biases and other unlearning techniques for robust safety performance.

3. Pilot experiment or demo

For this hackathon, we focused on testing AI models for stereotypes, particularly **racial and gender stereotypes**. We compared the performance of different models, namely [LLaMA-3.1-405b-Instruct](#), [Mistral-7b-v0.1](#), [Microsoft Phi-1.5](#) and [GPT-2](#), by analyzing their responses to prompts about the following scenarios:

- LA neighborhood stereotypes
- Racial occupation stereotypes
- Racial stereotypes affecting healthcare treatment of patients

The results show that all models exhibit some level of stereotyping, but they differ in the severity and nature of the stereotypes. Mistral-7b-v0.1, Microsoft Phi-1.5, and GPT-2 consistently produce biased responses with Mistral-7b-v0.1 generating responses with the most negative and harmful stereotypes.

While LLaMA-3.1-405b-Instruct does not necessarily give biased responses, it shows a tendency to overcompensate for negative stereotypes.

There seems to be a correlation between the capability and safety of these models although further experiments are required to support this claim.

We conclude that while there's progress, none of the models achieve the ideal of treating individuals from non-dominant groups without stereotypes. The comprehensive list of prompts and model specific responses can be found at [Benchmark Stereotype Evaluation](#).

Devising Effective Benchmarks

4. Process

Timeframe	What will you do?
Next 3 months	Consult with experts in the field, including the Algorithmic Justice League, HCI specialists, and leaders in current benchmarking to determine the most appropriate metrics for evaluating AI systems
2025	Have a usable version of benchmarking software and datasets. Build more benchmarks in high-risk fields. First customers.
2026	Continue building high-quality benchmarks with guidance from experts in each field to make available to customers.
2027	Build our first iteration of the overarching benchmark which will set the standard for evaluating all superintelligent models.

5. Impact on AI safety & key risks

There is a dire need for better evaluations particularly when determining the safety of AI systems. Until it becomes possible to reverse engineer neural networks completely, the need for third-party out-of-distribution adversarial safety benchmarks would always be prevalent. From our experiments, we infer that the evolution of large language models, especially the largest ones, over the years have increased the safety in these models. However, the non-frontier models are still very much in use due to company affiliations in products (example: use of Gemini in Google Search) or resource constraints (use of small language models like Phi-1.5). There is a **need for models to be safe**, even if they are not the most capable. This is where third-party benchmark solutions like us would be important.

The biggest challenge and risk to our solution is the ability of language models and AI systems in general being deceptively good at hiding unsafe behaviour. Some ways to mitigate this are strong collaborations with subject matter policy experts, red teaming experts, strong adversarial evaluations.

6. Appendix

Sources:

1. [Devising ML Metrics | CAIS](#)
2. [A new initiative for developing third-party model evaluations \ Anthropic](#)
3. [Under the radar? | Ada Lovelace Institute](#)
4. [AI Companies Shouldn't Grade Their Own Homework: Buolamwini](#)

Devising Effective Benchmarks

5. [Bias and Fairness in Large Language Models: A survey](#)
6. [Bias in Language Models: Beyond Trick Tests and Toward RUTED Evaluation](#)
7. [SmartGPT: Major Benchmark Broken - 89.0% on MMLU + Exam's Many Errors](#)
8. [Are We Done with MMLU?](#)
9. [4 shocking AI bias examples | Prolific](#)
10. [Insight - Amazon scraps secret AI recruiting tool that showed bias against women | Reuters](#)
11. [How We Analyzed the COMPAS Recidivism Algorithm — ProPublica](#)
12. [A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.](#)
13. [meta-llama/Meta-Llama-3.1-405B-Instruct · HuggingFace](#)
14. [mistralai/Mistral-7B-v0.1 · HuggingFace](#)
15. [microsoft/phi-1_5 · HuggingFace](#)
16. [openai-community/gpt2 · HuggingFace](#)
17. [esbenkc/DarkGPT: Dark Patterns in Chatbot Design · Github](#)