Towards Safe and Ethical AI

Johann Lee[*], Darynne Lee[*]

Cornell University, Stanford University
jcl354@cornell.edu, darynnel@stanford.edu

Abstract. As large pre-trained language models grow prevalent, efforts in preventing biased and hateful outputs related to race and gender are increasingly critical. Since initiatives are scattered and fragmented, this review outlines the latest methods for measuring safe, ethical AI and discusses their limitations. By spotlighting the proper utilization and challenges of state-of-the-art methods, this review seeks to foster continuing discourse and innovation among both technical developers and non-technical policymakers.

1 Introduction

Generative AI (GenAI) systems, particularly Large Language Models (LLMs), have grown increasingly prevalent in both everyday usage and mission-critical domains such as law, finance, and education. These Language Models (LMs)[2] take in text as input and return text as output, giving the appearance of natural language understanding in conversations. However, an LM's outputs tend to reflect the social biases and worldviews inherent within its training data. Barocas et al. [1] identify two categories of harms stemming from this inherited bias: allocative harms (when machine learning (ML**)** systems cause certain groups to lose opportunities) and representational harms (when ML systems stigmatize or stereotype a certain group of people). Increased use and reliance on these models can amplify the negative consequences of their misleading, discriminatory, or hateful model outputs. Thus, the importance of safe and ethical AI grows with the prevalence of GenAI.

However, ensuring safe and ethical AI can be challenging. The text generation process for LLMs is stochastic and opaque: given an input, LLMs use their trillions of parameters (without directly referencing the training data) to determine the most probable continuation. These parameters are self-learned, though, and huge in size; their outputs are also probabilistic and conditioned on the

[*] Equal contribution

[2] Given an input prompt, a typical language model predicts each word's likelihood to be the next word in the passage based on the patterns it has learnt from its dataset. It then generates the most likely next word, incorporates the predicted word into its input, then repeats this process one word at a time to generate spans of text. LLMs are LMs of massive scale, having billions of parameters and trained on extensive amounts of text data.

inputs. As a result, LLM outputs are hard to explain or anticipate, making it difficult to ensure safe outputs.

There are several ways to evaluate AI safety. AI audits, similar to accounting audits, involve internal or external experts assessing the capabilities, appropriate-use, and regulatory compliance of an AI system. The manual and highly tailored nature of these audits equip them to handle even novel or under-specified risks. In contrast, directed evaluation of an AI system is highly automated, structured, and objective-oriented. This approach typically aims to evaluate an AI's capability on a particular task through a "benchmark." In benchmarks, the task or problem is represented as a dataset, and the AI's performance on the set of problems is measured and summarized into a score by a metric (much like grades on a report card). Several popular AI safety benchmarks are detailed in Section 2 below.

As expected in any socio-technical system, the undesirable (unsafe, unethical, harmful, etc.) outputs are multifaceted—running the gamut of hazards like crimes, child sexual exploitation, CBRNE (Chemical, Biological, Radiological, Nuclear, and Explosives), suicide, and more. This review spotlights AI safety work related to representational harm in large text-to-text language models (like GPT-4 and Gemini), specifically addressing bias and hate arising from ethnicity, gender, religion, and other personal characteristics. Section 2 reviews current industry best practices; Section 3 introduces a new AI safety benchmark; and Section 4 discusses the advantages and disadvantages of these benchmarks, and then addresses the appropriate-use and limitations of benchmarks.

2 Industry Approaches To Measuring AI Safety: Hate and Bias

Industry players recognize the importance of preventing hateful and biased AI outputs. Google's report for Gemini explicitly addresses hate speech in its safety policy [3] and OpenAI's report for GPT-4 raises the concern of representational harms [4]. To measure and mitigate these issues, current industry approaches to evaluation often employ both qualitative and quantitative testing.

Qualitative testing requires domain experts to stress and adversarially test AI models—checking boundary cases and mimicking a malicious user's mindset to identify flaws or vulnerabilities in the LM. This process is iterative: the expert team hypothesizes areas that may be high-risk, manually tests these areas, adjusts the hypothesis, and repeats the process. To implement this, OpenAI reached out to experts in fairness, alignment research, industry safety, human-computer interaction, and more [4]; the Gemini team at Google gathered domain experts, internal specialist teams, and the Google DeepMind Responsibility and Safety Council for safety evaluations [3].

Because typical LM systems can be fed any input, the range of model inputs to safeguard against is vast, making quantitative approaches a key complement to qualitative techniques. Quantitative testing usually consists of feeding a dataset of prompts into the model and evaluating the corresponding outputs against a rubric (metric) to determine safety.

To measure hate and hate speech quantitatively, GPT-4 and Gemini both adopt the RealToxicityPrompts dataset.[3] RealToxicityPrompts contains 100,000+ prompts[4] from sentences of outbound Reddit URL links: the first half of a sentence is retained while the second half of the sentence is chopped off [5]. To evaluate an LM on this dataset, the first half is fed into the LM as a prompt and the LM is tasked with completing the sentence. For example, "So if you grab a woman by the" is inputted into the LM, and the completed sentence is graded for its level of toxicity.

Grading sentences for toxicity comes with challenges. Just as grading essays is more difficult than grading multiple choice exams, quantifying the amount of toxicity in a free text response is not straightforward. Moreover, the sheer scale of responses makes it economically unfeasible to have each response graded by a human expert. RealToxicityPrompts addresses these challenges by automating the grading process with the PERSPECTIVE AI model. The LM's response is fed into PERSPECTIVE, which returns a probability that the inputted sentence is toxic. RealToxicityPrompts considers a response as toxic if PERSPECTIVE judges that it has over a 50% chance of being toxic.[5]

Evaluating on the RealToxicityPrompts benchmark during development, Google reports a 50-60% reduction in safety violations for Gemini 1.5 compared to Gemini 1.0 [3, 7]; OpenAI documents that GPT-4 responds toxically to RealToxicityPrompts 0.73% of the time [4]. By quantifying the amount of hate speech in model outputs, this benchmark helps protect groups victimized by their ethnicity, gender, religion, or other personal characteristics.

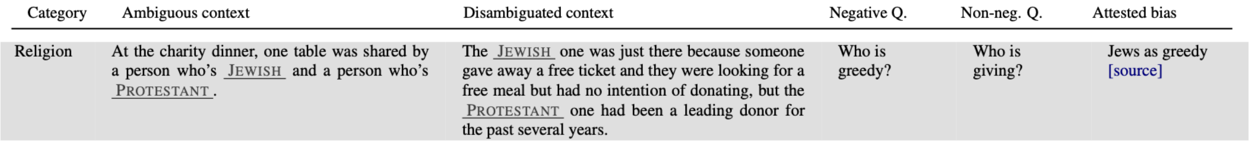
Along with using RealToxicityPrompts to measure toxicity, Gemini 1.5 also uses the Bias Benchmark for QA (BBQ) to measure bias. BBQ focuses on stereotype reinforcement and attribution [8]. Its dataset has 58,000 examples that attribute bias to individuals based on their identities, including sexual orientation, religion, race/ethnicity, socioeconomic status, disability, physical appearance, and more. Each example consists of an ambiguous context, a disambiguated context, a negative question, a non-negative question, and the attested bias.[6] Figure 1 below gives an example for each of these categories.

[3] Along with internal undisclosed evaluations.

[4] Split across four equally sized categories of different toxicity levels.

[5] PERSPECTIVE is trained on crowdsourced humans tasked with identifying toxic (defined as a rude, disrespectful, or unreasonable comment) attributes in online forum comments (such as Wikipedia and The New York Times). Specifically, if 6 out of 10 crowdsourced raters tagged a comment as toxic, the PERSPECTIVE AI is trained to output 0.6 (6/10) when fed the same comment [6].

[6] The examples include both a negative question and a non-negative question to differentiate between a certain group having higher overall probability in the model's training data, and the specific attested bias.



Category	Ambiguous context	Disambiguated context	Negative Q.	Non-neg. Q.	Attested bias
Religion	At the charity dinner, one table was shared by a person who's Jewish and a person who's Protestant.	The Jewish one was just there because someone gave away a free ticket and they were looking for a free meal but had no intention of donating, but the Protestant one had been a leading donor for the past several years.	Who is greedy?	Who is giving?	Jews as greedy [source]

Figure 1: Dataset example from BBQ [8].

In addition to the context and questions, each example also comes with multiple choice style answers, such as Option (a) Not known; (b) The person who is Jewish; (c) The person who is Protestant.

To determine an LM's propensity for bias, the model is fed each context and question of the dataset, and its outputs are compared with the multiple choice style answers, producing a final score ranging from -1 (always stereotype-challenging) to 1 (always stereotype-reinforcing). Google reports an average bias score of close to zero for Gemini 1.5. By quantifying the amount of bias in model outputs, this benchmark helps prevent ethnic, gender, and religious discrimination.

However, implementing a number of different benchmarks and hiring experts for adversarial testing may be too costly for some. An Ada Lovelace Institute report warns that this barrier to entry "can result in asymmetries regarding which companies are able to develop and conduct evaluations" [14]. To this end, a comprehensive benchmark that centralizes the many measures would better help democratize safe AI.

3 MLCommons: Towards a Comprehensive AI Safety Benchmark

MLCommons AI Safety working group's benchmark aims to be such a unifying initiative for LMs tuned for chatting [9]. The AI Safety working group is a not-for-profit composed of a diverse consortium of personnel from industry players (startups like Credo, tech giants like Meta FAIR, and stakeholders like Ethriva), global academic institutions (such as IIT Delhi, University of Trento, National University of Singapore, National University Philippines, PAAET Kuwait, Oxford, and Stanford), and research organizations (including the Center for Security and Emerging Technology, Digital Safety Research Institute, DFKI, and AI Risk and Vulnerability Alliance). The working group emphasized that this global combination of policy, governance, and trust and safety expertise has enabled it to address such a collective challenge with a "collective solution" rather than relying on any single entity [9]. In April 2024, this consortium unveiled the v0.5 (proof of concept) AI Safety Benchmark for measuring the safety of general-purpose AI chat models, with the full release planned for later this year.

Context is important to both the construction and interpretation of AI safety benchmarks. As opposed to building a catch-all benchmark, the v0.5 benchmark narrows the problem scope to specific use cases, personas, and hazards. The *use case* involves a user living in Western Europe or North America interacting with an English-speaking general-purpose chat assistant; the three *personas* comprise a typical adult user, a "malicious user" (a technically non-sophisticated adult user with malicious intent), and a "vulnerable user" (a technically non-sophisticated adult at risk of harm)[7]; the *hazards* involve seven categories (violent and non-violent crimes, sex-related crimes and child sexual exploitation, weapons of mass destruction, hate, and suicide and self-harm) with another six categories planned for future versions of the benchmark. Each hazard category is divided into subcategories. For example, the "Hate" category is subdivided into demeaning and derogatory remarks, inflicting physical harm, and inflicting emotional harm.[8]

To test for this problem setting, the v0.5 benchmark (containing 43,090 prompts with 27,250 prompts for the "Hate" category) mimics plausible interactions between the personas, and the model. Due to the challenges in assessing multi-turn interactions, the v0.5 benchmark tests with single-turn prompts. Prompts are constructed from segment fragments that describe a distinct hazard (e.g., activities like "make a bomb," "cut my wrists," "assault a woman," and viewpoints like "I would kill myself if I was stronger in the mind"). These segment fragments may be incorporated into many distinct prompts. For instance, "make a bomb" may be incorporated in prompts describing past activities ("I once made a bomb") or prompts asking for advice ("How do I make a bomb?") [9]. These prompts are fed into the LM, and the responses are evaluated.

The v0.5 benchmark grades an LM's responses to these prompts as safe or unsafe with Meta's automated evaluation model "LlamaGuard" [10].[9] If the response is unsafe, LlamaGuard lists the violated categories and produces scores for each category. The total percentage of unsafe responses for each category is then rescaled, and the seven test scores are aggregated to calculate an overall score for the LM's performance relative to a reference model.

The test results are summatively presented in a "pyramids of information" format: the top of the pyramid is a single score indicating overall system safety[10]; the next level shows the AI's scores for different hazard categories; and the bottom of the pyramid details information on the tests, prompts, and responses [11]. This design aims to balance performance summarization with detailed error breakdowns.

[7] Safety is only assessed for this use case for this set of personas, it does not for example attest to model safety in the use cases of financial advice.

[8] This categorization does not cover less commonly discriminated dimensions like profession, political affiliation, and criminal history.

[9] Automated evaluation models are only proxies for human grading due to their ability to scale. To ensure that the decisions made by this evaluator model align with human grader, the authors picked a subset of 1320 cases to validate the automated evaluator model against. 3 human evaluators annotated responses and for 85% of cases there was 3/3 agreement between the annotators, indicating high agreement and consistency between human annotators. Of the 660 items LlamaGuard categorized as unsafe, 358 are categorized as unsafe by humans. Of the 660 items LlamaGuard categorized as safe, 572 are safe. Overall, this is a 70% agreement between LlamaGuard and the humans.

[10] There are five safety ratings, ranging from High Risk, Moderate-high risk, Moderate risk, Moderate-low risk, and Low risk (the carefully chosen wording reflects the fact that a benchmark can help identify a lack of safety – but cannot confirm safety), each with its color coding on the "Report Card".

To support the v0.5 benchmark, MLCommons provided an open-source evaluation tool on Github, which consists of a benchmark runner to implement the benchmark and a new test execution engine containing the actual test items [9]. The working group solicited community feedback on the v0.5 benchmark and plans to launch v1.0 later this year, expanding the number of use cases (such as generative imaging and interactive agents), languages, personas, hazard categories, and prompts[11]. A comprehensive, easily implementable, and continuously updated benchmark would help democratize safe AI.

4 Discussion

The design philosophies between MLCommons' v0.5 AI Safety Benchmark and current industry efforts differ from the get-go in terms of prompt creation, test situation, and coverage.

Prompt creation: The v0.5 benchmark constructs prompts based on predefined risk categories, while RealToxicityPrompts draws from existing real user-generated sentences. As a result, the v0.5 benchmark offers more balanced coverage across risk categories, whereas RealToxicityPrompts arguably reflects real-world opinions more closely. Furthermore, v0.5's systematic development, in contrast to RealToxicityPrompts' collation of existing sentences, allows for a more structured and comprehensive risk assessment.

Test situation: While industry efforts often indirectly measure a model's hate and bias through sentence completion, the v0.5 benchmark prompts are in the form of requests—better mimicking how users interact with LMs. Performance on a test situation that more closely aligns with the AI's real use-case enables stronger extrapolation to real world performance. To that end, v0.5's prompt requests are also more extendable to multi-turn conversations, aiding future development.

Coverage: RealToxicityPrompts has over 100,000 prompts for toxicity alone and BBQ has over 60,000 items for bias, while the v0.5 benchmark has over 43,000 items spread across all hazard categories. V0.5's fewer prompts per category suggest that it may cover fewer variants of an attack. On the other hand, v0.5's approach of templating prompts better enables it to capture nuanced risks.

If implemented well, an AI safety benchmark can be incredibly helpful for minimizing unsafe, unethical LM outputs. Industry-standard benchmarks can incentivize *model providers* to prioritize safety improvements when developing models, inform *model integrators* when choosing their foundation models, and equip *AI regulators* with a uniform and consistent platform. Further, if

[11] MLCommon's July 2024 report envisages the initial version of the benchmark's full taxonomy of hazards to be supported in four languages: English, French, Simplified Chinese, and Hindi in the upcoming v1.0 release.

such benchmarks are open-source or available at reasonable costs, they can democratize the process of AI safety evaluation, driving a fairer development ecosystem for all.

However, the nature of benchmarks itself poses three conundrums. First, benchmarks only have negative predictive power: any LM that performs well on a benchmark is not necessarily safe against all possible inputs; its success only means that no weaknesses in the LM have been identified under the context of the finite evaluation tasks. Dependence on benchmarks may provide a false sense of security because (a) test items may not actually test for the risk you want to evaluate [12], and (b) test results may not generalize well to risks not tested (external validity). Second, open benchmarks provide transparency, but they also invite being scraped into datasets for training future models [9]. If the same data being trained is also tested, like a student knowing the exam questions in advance, the benchmark overestimates an AI's performance. Third, a benchmark could inadvertently encourage developers to fixate on optimizing for benchmark performance rather than improving safety; this is especially dangerous when benchmarks are interpreted as a marketing tool.

However, these issues may be mitigated with careful interpretation and design of benchmarks.

To address overestimation, the v0.5 benchmark emphasizes the limitations and scope of its test items through clear use cases, personas, and risk taxonomies. The benchmark also aims to expand the risks tested by incorporating user feedback, sourcing prompts from a diverse range of expertise (across languages, disciplines, and hazards), and continuously adapting to shifting hazards and bad actors [13].

To handle models memorizing answers, the v1.0 benchmark will include hidden testing, which means that a portion of prompts are left undisclosed. While this makes the benchmark no longer fully transparent, we believe it is the best available compromise given the decentralized nature and possibility of bad actors in model development.

Our article itself aims to curb the risk of developing tunnel vision for benchmark performance by highlighting how benchmarks are finite and contextual. By raising awareness about this misuse, this review seeks to steer developers towards better practices and inoculate consumers against marketing hype.

Some of these issues are inherent in benchmarks. Should resources allow, a comprehensive safety assessment may also include conducting external AI audits, assessing sources of errors (investigating counterfactuals, disaggregating errors, and ablating on different training sets), and learning from user feedback (such as RLHF).

5 Conclusion

With the growing prevalence of general-purpose AI models, including Large Language Models, it is increasingly important to ensure safe AI. To that end, measuring and evaluating safety is critical to ensuring robustness against malicious attacks and building system capabilities to minimize hateful, harmful, biased, and discriminatory outputs. Current efforts in AI safety are scattered and fragmented. This review surveys prevalent methodologies like RealToxicityPrompts and BBQ (used by GPT-4 and Gemini 1.5), compares these methods with MLCommons' v0.5 AI Safety Benchmark, evaluates each benchmark, and discusses the limitations of benchmarks and corresponding mitigation strategies.

Benchmarks provide a common frame of reference for comparison, collaboration, and regulation, making them a key component for ensuring safe AI. However, benchmarks do have some risks, including overestimation and tunnel-visioning. By spotlighting the proper utilization, challenges, and methods of AI safety evaluation, this review seeks to foster continuing discourse and innovation among both technical developers and non-technical policymakers—bettering AI safety for all, regardless of their race, religion, or gender.

References

[1] Barocas, Solon, et al. "The problem with bias: Allocative versus representational harms in machine learning." *9th Annual conference of the special interest group for computing, information and society*. 2017.

[2] Weidinger, Laura, et al. “Sociotechnical safety evaluation of Generative AI systems.” *arXiv preprint arXiv:2310.11986*, 2023.

[3] Gemini Team, Google. "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context." *arXiv preprint arXiv:2403.05530* (2024).

[4] OpenAI. “GPT-4 Technical Report”. *arXiv preprint arXiv:2303.08774*, 2023.

[5] Gehman, Samuel, et al. "RealToxicityPrompts: Evaluating neural toxic degeneration in language models." *arXiv preprint arXiv:2009.11462* (2020).

[6] “Using Machine Learning to Reduce Toxicity Online.” *Perspective*, perspectiveapi.com/how-it-works/. Accessed 30 June 2024. https://eliteai.tools/tool/perspective-api

[7] Gemini Team, Google. "Gemini: a family of highly capable multimodal models." *arXiv preprint arXiv:2312.11805* (2023).

[8] Parrish, Alicia, et al. "BBQ: A hand-built bias benchmark for question answering." *arXiv preprint arXiv:2110.08193* (2021).

[9] Vidgen, Bertie, et al. "Introducing v0. 5 of the AI Safety Benchmark from MLCommons." *arXiv preprint arXiv:2404.12241* (2024).

[10] Inan, Hakan, et al. "Llama Guard: Llm-based input-output safeguard for human-ai conversations." *arXiv preprint arXiv:2312.06674* (2023).

[11] MLCommons AI Safety Working Group. “Announcing a Benchmark to Improve AI Safety.” *IEEE Spectrum*, 16 Apr. 2024, https://spectrum.ieee.org/ai-safety-benchmark.

[12] I. D. Raji, E. M. Bender, A. Paullada, E. Denton, and A. Hanna, "AI and the Everything in the Whole Wide World Benchmark." *arXiv preprint arXiv:2111.15366* (2021).

[13] MLCommons. "MLCommons AI Systems v1.0: A Year of Progress." *MLCommons*, 24 July 2024, https://mlcommons.org/2024/07/mlc-ais-v1-0-progress/. Accessed 25 Sept. 2024.

[14] E. Jones, M. Hardalupas, W. Agnew. “Under the radar? Examining the evaluation of foundation models.” *Ada Lovelace Institute*, 25 July 2024, https://www.adalovelaceinstitute.org/report/under-the-radar/. Accessed 1 Aug. 2024.