SoK: Benchmark Datasets for Evaluating AI Safety — Gaps Between Guidelines and Practice

Nami Ashizawa[1] and Osamu Saisho[1]

NTT, inc., Tokyo, Japan {nami.ashizawa,osamu.saisho}@ntt.com

Abstract. This SoK paper systematizes safety-evaluation perspectives from eight major international AI guidelines and examines how existing benchmark datasets support them. Governments and research bodies have published AI safety guidelines that outline key perspectives and standards of security and safety for responsibly developing and using AI to address potential societal risks brought by the rapid AI deployment. However, while these guidelines define what perspectives should be evaluated, they often fall short of specifying how such evaluations should be conducted, particularly which input datasets should be used to evaluate AI model outputs. This paper unifies eleven perspectives, operationalize them into search queries, and retrieve candidates from GitHub and Hugging Face. From 726 initial hits, 505 met screening criteria, and 93 qualified as safety benchmarks. Then, we provide a mapping of benchmarks to the perspectives and identify the systematic gaps for practice. Our findings reveal systematic imbalances: perspectives such as *Control of Toxic Output* and *Fairness and Inclusion* are comparatively well covered, whereas *Data Quality* and *Verifiability* remain largely unsupported by existing datasets. On the basis of these results, we propose recommendations for benchmark design and for guideline structures to better align resources with safety guidance and enable more comprehensive output-level evaluation. Totally, this study offers a foundation for aligning "what to evaluate" with "how to evaluate," thereby strengthening the security and safety of AI systems.

Keywords: Gaps between benchmark datasets and evaluation practice · AI safety · Dataset surveys for AI evaluation · Evaluation perspectives on AI security and safety · Large language models.

1 Introduction

Artificial intelligence (AI) systems are increasingly being deployed throughout society. It raises critical risks for AI security and safety, and these risks are already being observed in real-world applications. From a security perspective, unsafe AI behavior may even facilitate adversarial exploitation, such as data exfiltration, poisoning, or prompt-injection attacks. In response, governments and organizations worldwide are working to define and mitigate risks such as information leakage, the generation of harmful output, and the amplification of social biases. They are also releasing guidelines to provide perspectives and

standards for the security and safety necessary to develop, deploy, and use AI. However, these guidelines vary in scope and abstraction, providing limited operational guidance on how evaluations should be conducted and which tools should be used. For example, [23] introduces AI safety evaluation perspectives and the types of evaluation resources for these perspectives, but it does not specify how each resource should be applied to evaluate a given perspective in a concrete protocol. Likewise, [20] surveys AI safety evaluation perspectives and lists existing evaluation resources; however, it does not indicate which resources correspond to the perspectives defined in the guideline. As a result, readers cannot determine which set of resources collectively covers the full range of evaluation perspectives.

AI safety evaluation can target multiple components of a system, such as model architecture, training datasets, input data, output data, and the end-to-end development process. Among these, output is particularly critical because it can directly affect vulnerable end-users who are targeted due to a lack of knowledge and can be weaponized by adversaries. For example, outputs that contain harmful biases may lead to unfair discrimination or psychological harm [23]. In another example, output-level harm may allow malicious actors to mount systematic attacks by exploiting model biases or manipulating generation through jailbreak prompts [21, 24]. Therefore, it is crucial to assess whether the output is safe for use by untrained or malicious end-users from the perspectives of fairness, trustworthiness, and robustness against adversarial misuse.

Various methodological approaches, including case studies [53], benchmarks [63], red-teaming [58], and auditing [34], have been proposed to evaluate the safety of such AI outputs in practice [24]. All of these rely on providing carefully designed input datasets to AI systems and analyzing the resulting outputs. Therefore, benchmark datasets serve as foundational resources to define what perspectives are consistently tested across AI models and studies. Yet, most existing datasets were not originally designed for safety evaluation, and their alignment with safety guidelines remains unclear. This misalignment creates systematic blind spots: risks identified in guidelines cannot be reliably assessed in practice, leaving potential security vulnerabilities untested.

While safety guidelines specify **what** should be evaluated, they provide limited guidance on **how** to evaluate it. Although benchmark datasets are sometimes used in practice, many were created for other purposes and do not explicitly align with guideline perspectives, resulting in uncertainty and inconsistency in safety and security evaluations. Although several works [37, 42, 66, 71, 87] have investigated the threats to AI safety and the comprehensive quality of countermeasures, they are not associated with the perspectives of AI safety evaluation proposed in worldwide guidelines [19–26]. To address this gap, this work adopts a Systematization of Knowledge (SoK) approach: we extract and unify common evaluation perspectives from the eight major AI safety guidelines, systematically collect existing benchmark datasets from GitHub[1] and Hugging Face[2], and critically analyze dataset documents to assess their alignment with the unified safety

[1] `https://github.com/`
[2] `https://huggingface.co/`

evaluation perspectives. We then identify the missing elements from a practical perspective and recommend future initiatives needed in AI safety.

Specifically, we operationalize the unified perspectives into search phrases, apply them to benchmark repositories, and manually examine dataset documents and associated papers. This process clarifies which perspectives are explicitly supported and which remain challenged or unsupported. Our results reveal systematic imbalances in how safety evaluation perspectives are supported by benchmark datasets: perspectives such as `Control of Toxic Output` and `Fairness and Inclusion` are relatively well covered, whereas perspectives like `Data Quality` and `Verifiability` remain almost entirely unaddressed. These blind spots highlight systemic limitations in current evaluation practice. By analyzing both coverage and co-occurrence patterns, we provide a structured view of where evaluation is well supported and where critical gaps persist. We also discuss the potential reasons for the gaps and future directions to improve benchmark dataset availability across the unified perspectives in AI safety.

The contributions of this paper are summarized as follows:

- Systematization of AI safety guidelines: We extract and unify evaluation perspectives from eight major international AI safety guidelines and provide a consolidated framework for AI safety evaluation.
- Dataset mapping to the unified perspectives: We systematically collect benchmark datasets from GitHub and Hugging Face and analyze their documents to determine alignment with the unified evaluation perspectives.
- Gap identification: We reveal systematic imbalances in coverage showing that perspectives such as `Control of Toxic Output` and `Fairness and Inclusion` are well supported, while `Data Quality` and `Verifiability` remain largely unaddressed.
- Critical interpretation: We discuss potential reasons for these gaps, clarifying why certain perspectives are overlooked, and how this creates blind spots in AI security and safety evaluation.
- Recommendations for future AI safety: We outline future directions for designing benchmark datasets that better align with guidelines and support comprehensive AI safety evaluation.

2 Guidelines and Unified Evaluation Perspectives

2.1 Overview of Worldwide Guidelines

As AI technologies continue to advance and become more widely adopted, the development of AI services is accelerating. In response, leading institutions in several countries have published documents outlining their approaches and methods for evaluating AI safety.

The National Institute of Standards and Technology (NIST) has released several AI evaluation guidelines in the United States. In particular, the *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* [19] presents a risk management framework designed to promote and support the responsible

design, development, deployment, and use of AI systems. In addition, the *AI 600-1: Generative Artificial Intelligence Profile* [21] provides guidance on managing risks specific to generative AI. In Singapore, the AI Verify Foundation has published the *CATALOGUING LLM EVALUATIONS* [20], which discusses taxonomy, challenges, and recommended evaluation and testing approaches for large language models (LLMs). The foundation has also released the *Model AI Governance Framework for Generative AI* [25], which aims to establish a governance framework for generative AI systems. In the United Kingdom, the AI Safety Institute has issued the *International Scientific Report on the Safety of Advanced AI: interim report* [24], which compiles current findings on the capabilities and risks of advanced AI. The *Guide to Evaluation Perspectives on AI Safety* [23] was recently published in Japan, reflecting a growing effort to establish consensus around AI safety evaluation. The European Commission has published *The General-Purpose AI Code of Practice* [26], which aims to promote the adoption of human-centric and trustworthy AI. The *AI Safety Governance Framework* [22] was released in China, promoting consensus and coordinated efforts on AI safety governance among governments, international organizations, companies, research institutes, civil society organizations, and individuals, with the aim of effectively preventing and mitigating AI safety risks.

While these documents address AI safety evaluation, they differ in specific evaluation perspectives, granularity, and recommended methodologies. This study first visualizes the commonalities in evaluation perspectives across these guidelines and to reveal their correspondence with existing benchmark datasets. In this study, we focus on eight major guidelines issued by leading public institutions in the United States, Singapore, the United Kingdom, Japan, the European Union, and China. We define the evaluation perspectives on AI safety referring to these documents because they are officially endorsed, internationally visible, and explicitly address AI safety evaluation.

2.2 Coverage Comparison of Evaluation Perspectives

In this study, we compared the definitions and descriptions of evaluation perspectives in each document. By analyzing overlaps in terminology, expression, and content, we consolidate similar perspectives into eleven unified categories. Among these guidelines, the recently published *Guide to Evaluation Perspectives on AI Safety* [23] integrates and builds upon multiple prior guidelines. Therefore, we used its categorization to define our eleven perspectives and mapped their coverage against the other guidelines in Table 1. Our categorization is primarily based on [23], but we also include `Preventing Facilitation of Weapons Acquisition`, which is not mentioned in [23] but is explicitly emphasized in other guidelines [20–22, 24, 26]. These categories provide a consistent framework for comparing safety evaluation practices internationally and highlight common ground that is essential for both safety and security assessment. To investigate how benchmark datasets align with these perspectives, we identify characteristic phrases that define each evaluation perspective (Tables 2, 3, 4, 5, and 6). The following section describes how such phrases are extracted and formulated.

Table 1. Coverage of evaluation perspectives in AI safety guidelines

Evaluation Perspectives on AI Safety	[19]	[20]	[21]	[22]	[23]	[24]	[25]	[26]
Control of Toxic Output		✓	✓	✓	✓	✓		
Prevention of Misinformation, Disinformation and Manipulation	✓		✓	✓	✓	✓	✓	✓
Fairness and Inclusion	✓	✓	✓	✓	✓	✓	✓	✓
Misuse and Unintended Use			✓		✓	✓		
Privacy Protection	✓		✓	✓	✓	✓	✓	✓
Ensuring Security	✓	✓	✓	✓	✓		✓	✓
Explainability	✓	✓		✓	✓	✓	✓	✓
Robustness	✓	✓		✓	✓	✓		✓
Data Quality	✓	✓	✓	✓	✓	✓	✓	✓
Verifiability			✓		✓	✓		
Preventing Facilitation of Weapons Acquisition		✓	✓	✓		✓		✓

3 Benchmark Dataset Collection

3.1 Identification of Evaluation Phrase

We first define appropriate search keywords to investigate benchmark datasets related to the AI safety evaluation perspectives defined in Section 2.2. In this work, we use vocabulary extracted from the eight referenced guidelines [19–26] that indicate each evaluation perspective. This section describes the extraction method and vocabulary in detail to ensure reproducibility.

The search keywords should be strongly associated with each evaluation perspective. To select appropriate keywords, we focus on sentences in the guidelines that describe the desired state of AI systems from each perspective. Specifically, we identify two types of descriptions: those evaluating whether the AI is *protected from certain risks*, and those assessing whether the AI *maintains a desired state*. The former type is summarized in Tables 2 and 3, while the latter is summarized in Tables 4, 5, and 6.

Bold phrases in the tables denote search keywords closely associated with each evaluation perspective. These keywords are denoted as `eval_keyword` entries. Based on these keywords, we construct the following search query:

`eval_keyword` `AND` “ dataset ” `AND` “ language model ”,

where the logical conjunction `AND` requires that all terms appear simultaneously in the retrieved documents. This formulation ensures that the search results explicitly discuss (i) the targeted evaluation perspective represented by `eval_keyword`, (ii) datasets, and (iii) language models, thereby enabling us to systematically identify benchmark datasets that are directly relevant to the evaluation of language models.

Table 2. Sentences Describing Protected States for Evaluation (1/2)

Evaluation Perspective	Objectives to be prevented
Control of Toxic Output	... **toxic**ity is an umbrella term that encompasses **hate speech**, **abusive language**, **violent speech**, and **profane language** ...
	... to avoid generating **harm**ful ...
	... creation of and public exposure to **offensive** or **hate**ful language ...
	Information that could be used for **cyber-attack**s, **terrorism** and other **crime**s, and **CBRN** ...
	Information that may cause **psychological harm** to end users, such as **discriminatory** expressions.
Prevention of Misinformation, Disinformation and Manipulation	Prevention of **Misinformation**, **Disinformation** and **Manipulat**ion
	... the malicious use of general-purpose AI for **disinformation** and **manipulat**ion of public opinion.
	Risks from **confabulation**s may arise when users believe **false content** – often due to the confident nature of the response – leading users to act upon or promote the **false information**.
	... deliberate production or dissemination of false or **misleading information** (**disinformation**) at scale ...
	... end users may **misinterpret** the output from the LLM system as information emanating from a human ...
	... AI systems learned to systematically **induce false belief**s in others ...
Fairness and Inclusion	... the output of the LLM system be fair and free of **bias** and **discrimination**.
	... the test data containing **harm**ful **bias**es regarding diverse backgrounds such as race, gender, nationality, age, political beliefs, religion ...
Misuse and Unintended Use	Addressing **High-risk** Use and **Unintend**ed Use
	When LLM systems are used for **high-risk** purposes ...
	... to a specific use case can help to reduce risks from **unforeseen failure**s or **malicious use**.
	... to decrease risks arising from **misuse**, **abuse**, **inappropriate repurpose**, and **misalignment** between systems and users.
	... create a situation in which no **significant harm** or **disadvantage** is caused even if the system is used for other than its intended purpose.
	Even if test data for the **unintend**ed use case is input to the LLM system, can the system avoid producing output that may **harm** the end user's life, body, property, etc., or various rights?

Table 3. Sentences Describing Protected States for Evaluation (2/2)

Evaluation Perspective	Objectives to be prevented
Privacy Protection	When test data that includes information about **individual**s that should be protected is used as the assumed output, can the LLM system avoid output that information contrary to the **design intent**?
	Can the system prevent the **recovery** of **personal information** contained in the training data of the AI model ...
	... to **identif**y **individual**s from the training data by combining multiple outputs.
	... is the output of information related to **individual**s controlled?
Ensuring Security	Ensuring Security is important to minimize the impact of **malicious attack**s ...
	... **misconfiguration**s and **poor interaction**s between a system and a human ...
	... dataset **poison**ing and the **vulnerabilit**y of LLM-driven applications to **prompt-injection** attacks.
	... **circumvent** general-purpose AI model safeguards with relative ease, for example through ‘ **jailbreak**ing ’ techniques ...
	... **prompt injection** involves modifying what input is provided to a GAI system so that it behaves in **unintended way**s.
	... to prevent the output of **confidential information**?
Explainability	... reducing **misunderstand**ings and **distrust**.
Robustness	... minimize **potential harm**s to people if it is operating in an **unexpected setting**.
Data Quality	... **harmful bias** and other data **quality issue**s ...
	... the use of **copyright** material in training datasets and the issue of consent from **copyright** owners is starting to raise concerns ...
	... due to non-representative training data, that result in **discriminat**ion, amplification of **bias**es, or incorrect **presumption**s about performance ...
Verifiability	(none)
Preventing Facilitation of Weapons Acquisition	... significant **security breach**es, for example in **critical national infrastructure** (**CNI**) a threat to **public safety** and **security**.
	... making **explosive**s, bio**weapon**s, chemical **weapon**s, and **cyberattack**s ...
	... gain **unauthorized access** to current **weapon** systems or contribute to the design and development of new **weapon**s technologies.
	... **synthesis** of materially nefarious information or design capabilities related to chemical, biological, radiological, or nuclear (**CBRN**) **weapon**s ...

Table 4. Sentences Describing Ideal States to Be Maintained (1/3)

Evaluation Perspective	Objectives to be achieved
Control of Toxic Output	(none)
Prevention of Misinformation, Disinformation and Manipulation	... provide **accurate information** ...
	... a **fact-finding** mechanism is placed for LLM system outputs.
	... **distinguish fact** from opinion or fiction or **acknowledge** uncertainties ...
	... **distinguish human-generated** content from **AI-generated** synthetic content.
Fairness and Inclusion	**Fair**ness and **Inclusi**on
	... the output of the LLM system be **fair** ...
	... outputs that are overly uniform (for example, repetitive **aesthetic** styles and reduced content **diversity**).
	... the output of the LLM system is **understandab**le, i.e., highly **readab**le, to all end users.
	... achieving more equitable **inclusi**on of sign language dialects.
	Are there not any problems with the score as measured by the **fluency score** (a numerical representation of whether the output is **grammatically appropriate**) of the LLM system's output?
Misuse and Unintended Use	... to **supersede**, **disengage**, or **deactivate** AI systems that demonstrate performance or outcomes inconsistent with **intended use**.
	... understand potential **misuse** scenarios and **unintend**ed outputs.
Privacy Protection	... LLM systems to protect **privacy** from the perspective of fostering confidence in LLM systems ...
	... **compliance** with legal and regulatory requirements ...
	Privacy values such as **anonymi**ty, **confidentiality**, and **control** ...
	... **Privacy Enhancing Technologies** (**PETs**), has the potential to allow data to be used in the development of AI models while protecting data **confidentiality** and **privacy**.
Ensuring Security	Ensuring **Security**
	... to general cyber**security** and AI **security** ...
	... maintain **confidentiality**, **integrity**, and **availability** ...

Table 5. Sentences Describing Ideal States to Be Maintained (2/3)

Evaluation Perspective	Objectives to be achieved
Explainability	**Transparency**, **explain**ability, and **interpret**ability ...
	... LLM system output allows end users to confirm the **credib**ility of the output ...
	... making the output more **convinc**ing ...
	... building **trust** among end users with insufficient domain expertise.
	... the **rationale** for the output can be confirmed to a technically reasonable extent for the purpose of presenting evidence of the LLM system's operation ...
	... visualize the output **rationale** (internal operation, its status, source, etc.) ...
	... present to the end user the **inference process** leading up to the output ...
Robustness	**Robust**ness or **generaliz**ability ...
	... **stable output** against unexpected inputs such as **adversarial prompt**ing, **garbled data**, and **erroneous input**.
	AI system's results are **consisten**t and can be **replicate**d
	... generate self-**consisten**t text ...
	... its **resilience** to various **perturb**ations.
Data Quality	... the data accessed by LLM systems are in an **appropriate state**, including during model training ...
	... ensure **data quality**, such as through the use of trusted **data source**s.
	... **annotat**ing training datasets consistently and accurately ...
	Is the data not **corrupt**ed?
	Is the data **grammatically appropriate** and **understandab**le by humans?
	... datasets that reflect the cultural and social **diversity** of a country ...
	... facilitate **data cleaning** (e.g., debiasing and removing inappropriate content).
	Does the data not contain any information that could be used for **cyber-attack**s or **terrorism**?
	Training data may also be subject to **copyright** and should follow applicable **intellectual property** rights laws.
	Is there appropriate **data configuration management**?
	Does the data not contain any **malicious** or **malfunctioning program**s?
	Maintaining the **provenance** of training data ...
	Conduct appropriate **diligence** on training data use to assess intellectual property, and privacy, risks ...

Table 6. Sentences Describing Ideal States to Be Maintained (3/3)

Evaluation Perspective	Objectives to be achieved
Verifiability	... various types of **verif**ication against LLM system are available from the model training phase and the development/provision phase of the LLM system to the time of use.
	... **validat**ion of assumptions for system design, data collection, and measurements relative to the intended context of deployment or application.
	... **validat**ion and **integrat**ion in production, with testing, and recalibration for systems and process integration, user experience, and compliance with existing legal, regulatory, and ethical specifications.
	Have **system card**s, **model card**s, and **data card**s been created?
	Are **data log**s properly recorded when various test data are entered into the LLM system?
Preventing Facilitation of Weapons Acquisition	(none)

3.2 Dataset Search and Classification Procedure

We search on GitHub and Hugging Face, two significant sources of datasets and pipelines for evaluating AI. On GitHub, we use the query directly. On Hugging Face, we filter results by “Dataset” type, limit the content to “Text,” and restrict the language tag to “English” before searching each `eval_keyword`.

This paper focuses only on datasets that primarily use English. This decision was made for two reasons: English is well-supported by LLMs, and we can directly review English content. While this paper limits its scope to English, future work should also address the development of multilingual benchmarks.

The investigation proceeds as follows. First, we collect search results from GitHub and Hugging Face using the defined search query. After removing duplicates, we obtained 726 results. Second, we exclude datasets on the basis of the following criteria, resulting in 505 candidates remaining:

- datasets lacking detailed documents (e.g., README or their papers),
- datasets with no publicly available content (generally on GitHub),
- datasets whose documents or data mainly include non-English languages,
- and datasets that mirror existing other datasets without modification.

Third, we review the documents associated with the datasets to confirm their intended usage. We focus on the documents associated with datasets to empathize with non-expert developers or users who utilize and evaluate AI systems, as they would typically select benchmark datasets by referring to the accompanying documents. Fourth, we retain only datasets intended for inputs to AI inference, i.e., not training datasets. Finally, we apply the `eval_keyword` search to the full text of each dataset's document. If the text surrounding the keyword is semantically

consistent with the sentences listed in Tables 2, 3, 4, 5, and 6, we categorized it under the evaluation perspective corresponding to each `eval_keyword`. As a result, we identify 93 datasets corresponding to one or more perspectives.

3.3 Distribution of Non-Matching Datasets

Among the 412 datasets excluding the safety benchmark set from 505 candidates, 167 datasets lack descriptions of their usage, and 120 datasets are training or inference datasets unrelated to safety evaluation. For example, a dataset titled "Toxic Comments" [18] appeared in the keyword search procedure, but the dataset lacks a description. TESRBench [51, 52] is a comprehensive benchmark dataset for temporal event sequence retrieval, not for AI safety.

The remaining 125 datasets are utilized to enhance safety or detect unsafe behavior. HateCheck [67] is a benchmark dataset for testing hate speech detection models, not for evaluating the bias of AI's outputs. These datasets are identified via the keywords in Tables 2, 3, 4, 5, and 6, and are designed for prevention rather than evaluation. The following section presents a detailed analysis of the investigation results and discusses future directions for developing benchmark datasets in the context of AI safety.

4 Gap Identification

In the following sections, we adopt the following abbreviations:

- **Toxic** denotes `Control of Toxic Output`.
- **Info.** denotes `Prevention of Misinformation, Disinformation, and Manipulation`.
- **Fair** denotes `Fairness and Inclusion`.
- **Misuse** denotes `Misuse and Unintended Use`.
- **Privacy** denotes `Privacy Protection`.
- **Security** denotes `Ensuring Security`.
- **Explain** denotes `Explainability`.
- **Robust** denotes `Robustness`.
- **Data** denotes `Data Quality`.
- **Verify** denotes `Verifiability`.
- **Weapon** denotes `Preventing Facilitation of Weapons Acquisition`.

4.1 Overall Coverage of Evaluation Perspectives

Tables 7 and 8, and Fig. 1 summarize the 93 benchmark datasets mapped to evaluation perspectives. Some datasets share an identical document and are grouped under one citation, so the number of datasets represented by each citation is indicated on the right of the citation. The results show that 48 datasets address **Fair**, 34 datasets address **Toxic**, 28 datasets address **Security** and **Robust**, and 24 datasets address **Info.**. On the other hand, only 16 datasets address **Privacy**, 8 datasets address **Misuse**, and 6 datasets address **Explain**. However, only three datasets address **Weapon**. Moreover, no benchmark datasets address **Data** or **Verify**, leaving these perspectives entirely untested.

Table 7. Coverage of Evaluation Perspectives by Benchmark Datasets (1/2)

Dataset	Toxic	Info.	Fair	Misuse	Privacy	Security	Explain	Robust	Data	Verify	Weapon
[40](2)	✓	✓	✓								
[50]	✓					✓		✓			
[60]	✓		✓								
[55](6)	✓	✓	✓			✓					
[46]	✓	✓	✓								
[30]	✓	✓	✓		✓	✓	✓	✓			
[54]	✓	✓		✓	✓	✓		✓			✓
[28]	✓										
[8]	✓										
[14]	✓					✓					
[86]		✓									
[64]		✓					✓				
[72](2)		✓									
[45]		✓									
[90]		✓									
[56]			✓								
[80](2)	✓		✓								
[29]			✓								
[39]	✓		✓								
[76]	✓		✓								
[91]			✓								
[68]			✓								
[92]			✓								
[94]			✓								
[65](4)			✓					✓			
[6]			✓								
[5]			✓								
[47]	✓		✓								
[35]			✓								
[77]	✓		✓								
[17]	✓		✓								
[16]			✓								
[13]			✓								
[81]			✓								
[57]			✓			✓					
[31]			✓								
[32]	✓		✓	✓		✓		✓			

Table 8. Coverage of Evaluation Perspectives by Benchmark Datasets (2/2)

Dataset	Toxic	Info.	Fair	Misuse	Privacy	Security	Explain	Robust	Data	Verify	Weapon
[2]			✓								
[88](2)			✓								
[41]			✓								
[9]			✓								
[27](4)								✓			
[93]	✓	✓	✓		✓	✓		✓			
[49]					✓						
[38]					✓	✓					
[73]					✓						
[59](2)					✓						
[48]					✓						
[62]						✓					
[74]						✓					
[85]				✓		✓					
[33](3)	✓	✓		✓	✓	✓		✓			
[84]	✓				✓	✓					
[7]						✓					
[75](2)	✓				✓	✓		✓			✓
[61]	✓					✓					
[1]						✓					
[3]						✓					
[15]							✓				
[44]	✓	✓	✓	✓	✓	✓	✓	✓			
[78]							✓				
[11]	✓			✓				✓			
[69]			✓					✓			
[79]								✓			
[4]								✓			
[43]								✓			
[82](2)		✓						✓			
[70]								✓			
[10]			✓								
[89]			✓								
[36]								✓			
[83]			✓								
[12]							✓				

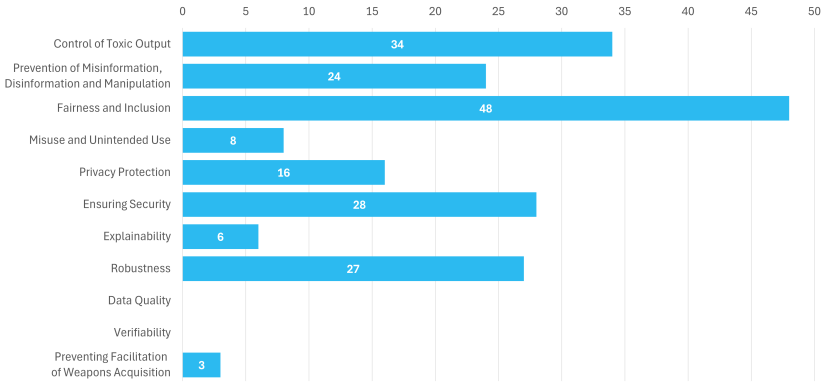

Fig. 1. Number of Datasets Corresponding to Each Evaluation Perspective

Table 9. Number of Datasets per Evaluation Perspectivea

	Toxic	Info.	Fair	Misuse	Privacy	Security	Explain	Robust	Data	Verify	Weapon
Datasets Found via Search	139	69	123	20	74	82	50	67	164	79	32
Safety Evaluation Benchmark Datasets	34	24	48	8	16	28	6	28	0	0	3
Other Safety-related Datasets	6	29	49	0	3	6	21	1	0	0	0

4.2 Perspective-Specific Analysis

While our primary analysis focuses on the 93 benchmark datasets identified in Section 4, further insight can be gained by also examining the datasets that were excluded during filtering. This complementary view clarifies why some evaluation perspectives, despite frequent appearance in guidelines, remain unsupported by usable benchmarks. Table 9 summarizes the number of datasets per evaluation perspectives across three categories: all datasets retrieved via keyword search, those retained as benchmark datasets, and those excluded but safety-related. These safety-related datasets (not benchmarks) are re-included to analyze broader community attention to safety-related keywords to discuss overall interest in each evaluation perspective.

This distribution shows that perspectives such as **Toxic**, **Info.**, **Fair**, **Privacy**, **Security**, and **Robust** appear to receive the more research attention. Especially, **Toxic**, **Privacy**, **Security** and **Robust** yield more safety benchmark datasets than other countermeasure datasets. It suggests an emphasis on output-level control for evaluating security. However, fewer usable benchmarks for **Toxic** and **Privacy** indicates that while they are recognized as critical issues, effective benchmark datasets remain limited. This is because heuristic detection of privacy or attack violations in AI output of the datasets for these perspectives is difficult, often requiring external judgment mechanisms or pipelines other than the input datasets.

Table 10. Benchmark Datasets Covering Multiple Evaluation Perspectives

Dataset	Toxic	Info.	Fair	Misuse	Privacy	Security	Explain	Robust	Data	Verify	Weapon
[44]	✓	✓	✓	✓	✓	✓	✓	✓			
[30]	✓	✓	✓		✓	✓	✓	✓			
[54]	✓	✓		✓	✓	✓		✓			✓
[93]	✓	✓	✓		✓	✓		✓			
[33](3)	✓	✓		✓	✓	✓		✓			
[32]	✓		✓	✓		✓		✓			
[75](2)	✓				✓	✓		✓			✓

In contrast, **Data** and **Verify** appear frequently in retrieved datasets but contribute no benchmarks, suggesting that these perspectives are typically addressed at the training or system level rather than through input–output evaluation. **Misuse** and **Weapon** appear infrequently both in retrieved and retained datasets. It indicates both limited research attention and intrinsic difficulty in operationalizing them into datasets. Meanwhile, **Explain** yields several countermeasure-oriented datasets designed to fine-tune models or enrich outputs. However, very few functions as benchmarks to evaluate explanation quality.

Overall, considering both benchmark and excluded datasets provides a clearer interpretation: current dataset resources systematically favor perspectives that are easier to operationalize at the output level, while perspectives requiring deeper system-level verification remain underrepresented. This highlights not only blind spots in benchmark coverage but also structural barriers preventing guideline perspectives from being translated into practical evaluation resources.

4.3 Multi-Perspective Coverage

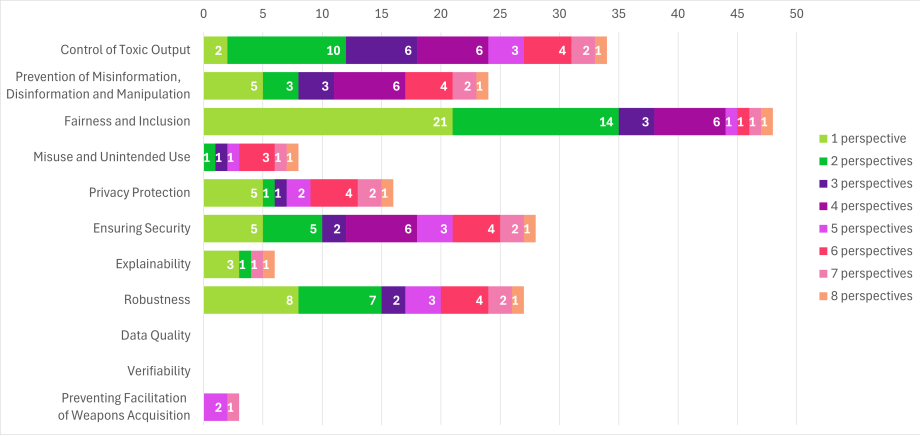

Fig. 2. Distribution of Covered Perspectives per Dataset

Fig. 2 provides a more detailed view of Fig. 1, illustrating the number of perspectives covered at once by each AI safety benchmark dataset. Each number on the bars represents the count of datasets covering each number of perspectives. From the graph, the datasets covering five or more perspectives at once are extracted in Table 10. The threshold "five or more" was chosen because the maximum number of perspectives covered by any dataset was eight. In the table, one dataset covers eight perspectives, two datasets cover seven, four datasets cover six, and three datasets cover five. To avoid double-counting, datasets with duplicated documents are consolidated into seven unique entries.

Perspectives such as **Toxic**, **Info.**, **Privacy**, **Security**, and **Robust** are frequently evaluated in multi-perspective datasets. In contrast, **Fair** is covered little in multi-perspective datasets, although many evaluation datasets are found. **Misuse** and **Weapon** are often evaluated alongside **Toxic** or **Security**.

These findings suggest that existing benchmarks often cluster around "operationally convenient" perspectives, where risk can be assessed through surface-level prompts and outputs. Perspectives requiring deeper systemic evaluation or domain-specific expertise are seldom integrated into multi-perspective benchmarks. These limitations restrict their applicability in holistic AI safety and security evaluations.

4.4 Summary of Findings

In summary, our analysis of 93 benchmark datasets, complemented by insights from excluded datasets, reveals systematic imbalances in the resources allocated to AI safety evaluation. Perspectives such as **Toxic**, **Fair**, **Security**, and **Robust** are relatively well supported, while **Data** and **Verify** remain entirely unaddressed. Privacy-related benchmarks are especially scarce despite being frequently retrieved, highlighting the intrinsic difficulty in output-only evaluation. **Misuse** and **Weapon** receive little attention, reflecting both research neglect and operational challenges. Multi-perspective benchmarks exist, but they cluster around a subset of perspectives that are more straightforward to operationalize, leaving critical areas unaddressed. Together, these findings demonstrate that current benchmark resources systematically fail to cover several guideline-defined risks, particularly those requiring system-level assurance. This gap highlights the pressing need for benchmark development that transcends surface-level output analysis, incorporates domain expertise, and aligns more closely with international AI safety guidelines.

5 Critical Interpretation

Building on the analysis of 93 benchmark datasets, this section highlights the concrete security risks that arise from the absence of suitable benchmarks and interprets the underlying causes of this absence in data collection and environment. These perspectives clarify why a persistent gap remains between AI safety guidelines and the practical availability of evaluation resources.

5.1 Security Risks Caused by the Lack of Benchmarks

The lack of benchmark datasets for evaluating AI safety raises concerns about the difficulty of sufficiently evaluating the safety of new AI systems prior to launch. For end users without domain expertise, this insufficient evaluation results in the following security risks.

Data: No benchmark datasets for evaluating **Data** are found in this study. As a result, representation bias, contamination, and copyright violations in training data propagate unchecked into model outputs. High-stakes domains such as healthcare and education therefore face increased risks of misjudgment and harm. Although guidelines call for practices such as consistent annotation, social diversity, and intellectual property compliance, benchmarks to verify these practices are absent.

Verify: No benchmark datasets for evaluating **Verify** are found in this study. As a result, users cannot trace why a response was generated due to the absense of sources, provenance, and uncertainty of outputs for substantiating. Guidelines recommend mechanisms such as system cards, model cards, and data logs, yet no benchmarks exist to evaluate them. This opacity leaves users unable to detect misinformation, fakes, or fraudulent advice.

Privacy: Only 16 benchmark datasets address **Privacy**. It leaves major threats unevaluated, such as recovery of personal information from training data or identification of individuals by combining outputs. Users without domain expertise may inadvertently input sensitive information that later leaks via outputs. However, standardized evaluations have not been sufficiently established for detecting leakage and reconstructing personal information.

Misuse / Weapon: Only eight and three benchmark datasets cover these perspectives, respectively. Related countermeasure resources are also scarce. Consequently, non-malicious users may inadvertently obtain instructions for making weapons or find themselves in danger of violating the law.

Explain: Only six benchmark datasets address **Explain**. Without systematic evaluation of credibility, transparency, and inference processes, users are thus more likely to accept opaque judgments without question. If AI systems produce misleading or false outputs, non-expert users may fail to recognize them, leading to serious incidents.

5.2 Causes of the Lack of Benchmarks: Collecting and Environmental Factors

We interpret the causes of the lack of benchmark datasets at two levels: data-collection difficulties aligned with AI safety evaluation guidelines, and environmental difficulties.

Data-collection Difficulties

Limitation of AI I/O evaluation: Data and **Verify** require evidence beyond the AI's I/O, including data provenance, configuration management, legal conformity, and logs. Because these cannot be captured by I/O alone, benchmark design for these perspectives remains underdeveloped.

Hurdle of ethics and privacy review: Benchmarks for toxicity, bias, and privacy including perspectives of misuse and weapons-related expressions face ethical constraints. While needed to evaluate risks, publishing them risks amplifying harm (e.g., exposing toxic language or personal data), which discourages researchers from releasing datasets.

Data drift on static datasets: Jailbreaks and prompt-injection techniques evolve rapidly. As new risks arise, static datasets require frequent updates. Dynamic benchmark datasets can reduce auditing and update costs, but the design of self-evolving datasets remains technically unsolved.

Dependence for expert annotation: Only subject-matter experts are necessary for reliable annotation. Moreover, multiple annotators improve consistency and reduce error.

Environmental Difficulties

Research attention bias: Despite their potential severity, perspectives such as privacy and weapons are largely neglected comparing to other perspectives such as fairness and robustness. This bias reflects a tendency to prioritize perspectives according to ease of implementation at the output level, rather than urgency.

Insufficient instruction of guidelines: Major guidelines describe *what* to evaluate for AI safety, but not *how* to evaluate it, i.e., which datasets to use. As a result, researchers face uncertainty in selecting and creating appropriate resources, slowing the development of standardized benchmarks.

Weak Documents: In narrowing 505 datasets to 93 evaluation targets, 167 datasets were excluded due to missing usage documents. Without clear and sufficient documents, practitioners cannot assess applicability or risk, which reduces usability and discourages adoption.

Non-public resources in industry: Many practical datasets remain proprietary, as companies are reluctant to release them because they do not want to lose competitive advantage by revealing useful data and the risks of misuse. This lack of transparency restricts community-wide progress in AI safety evaluation.

5.3 Discussion about Lack of Benchmarks

In summary, the lack of suitable benchmarks poses tangible risks for end users without domain expertise, including unchecked data quality issues, unverifiable outputs, insufficient privacy safeguards, exposure to misuse or weapon-related content, and opaque decision-making processes without explainability. These risks persist because of two intertwined challenges;

1. technical and ethical difficulties in collecting benchmark datasets that fully align with guideline-defined perspectives, and
2. environmental constraints, including biased research attention, insufficient guidance from standards, weak dataset documents, and non-public industrial resources.

Together, these factors explain that a persistent gap remains between AI safety guidelines and the practical availability of evaluation resources.

6 Recommendations and Future Work

On the basis of the findings summarized in Section 5, this section proposes two directions to close the distance between abstract safety guidelines and practical evaluation resources: benchmark design and guideline structures. Our goal is to strengthen system-wide evaluation while preserving end-user–facing assessments, thereby providing a roadmap for aligning guidelines with practical benchmarks.

6.1 Recommendations for Benchmark Design

System-wide evaluation via evidence attachments beyond I/O Perspectives such as **Data** and **Verify** cannot be evaluated solely through I/O. Benchmark datasets should incorporate system-level information, such as training data and model configuration, as part of the evaluation process. We foresee a demand for benchmark datasets that include prompts for these external mechanisms and whose scoring procedures evaluate the diagnostic information generated by the mechanisms themselves. In short, resources are needed that capture the entire system context while still evaluating outputs presented to end users.

Dynamic adaptation to emerging attacks Safety benchmarks quickly become outdated as jailbreaks and prompt-injection techniques evolve. To remain effective, benchmarks should dynamically incorporate descriptions of new attacks and vulnerabilities. Standardizing a reporting template for threats would enable the automatic generation of both candidate mitigations and evaluation data. Although automating the pipeline for standardization and automatic generation is technically challenging, it is essential for benchmarks to remain robust in rapidly changing environments.

Virtuous circle of annotation The high cost of manual annotation can be reduced by using well-evaluated AI systems as first-pass annotators. While these AI annotators themselves require carefully validated benchmarks, their iterative use can bootstrap richer datasets: annotated benchmarks improve AI annotators, which in turn produce higher-quality annotations. This virtuous circle should enable richer and more diverse benchmark datasets to be created at substantially lower marginal cost.

6.2 Recommendations for Guideline Structures

Materializing safety evaluation guidelines AI safety evaluation guidelines should include concrete evaluation resources, such as their intended usage patterns, and a clear mapping from resources to safety perspectives, thereby offering evaluators practical, process-level guidance. Such linkage provides actionable process-level guidance and highlights under-attended perspectives to correct research-attention bias.

Guidelines for constructing benchmark datasets Practical instructions are needed for designing benchmarks and usage documents, particularly regarding ethical and privacy requirements. These guidelines should address not only researchers but also AI providers and non-expert users who need to construct bespoke benchmarks for their systems.

Industry-aligned guidelines Industry deployments face the most urgent need for safety evaluation, yet current guidelines rarely reflect concrete operational use cases. To encourage adoption, guidelines should incorporate real-world examples and promote academic–industry collaboration in creating domain-specific datasets. Aligning safety standards with deployment realities reduces barriers to adoption and strengthens trust in AI systems.

6.3 Limitations and Future Work

For the sake of reproducibility, we employ a systematic keyword search procedure in this study. Moreover, we selected GitHub and Hugging Face as platforms for a systematic keyword search, as they are major platforms that host openly accessible datasets. As a result, some state-of-the-art datasets may not be captured through our search. An investigation method should expand the scope of search targets while maintaining reproducibility and reliability.

Although the keyword search followed a systematic procedure, interpreting the dataset's documents requires human judgment to support AI system developers/users without expertise. The interpretation was conducted by an author, so we should increase the number of annotators to achieve more reliable reviews.

7 Conclusion

This study systematized evaluation perspectives from eight major international AI safety guidelines and analyzed their correspondence with over 500 existing benchmark datasets collected from popular public dataset repositories. Through this Systematization of Knowledge (SoK) approach, we revealed systematic imbalances: perspectives such as `Control of Toxic Output` and `Fairness and Inclusion` are relatively well supported, whereas others like `Data Quality` and `Verifiability` remain almost entirely uncovered. These blind spots create uncertainty in safety evaluation and leave potential vulnerabilities unexamined from both safety and security perspectives. The gap opens up because these perspectives need comprehensive safety evidence beyond input–output testing and demand new dataset designs.

By clarifying where evaluation is well supported and where critical gaps persist, this study provides a foundation for aligning benchmark dataset development with unified guideline perspectives. Specifically, first, we materialize the unified perspectives into search queries and a tagging rubric, quantifying coverage and co-occurrence across more than 500 datasets on GitHub and Hugging Face. Second, we derive benchmark-design recommendations that encourage system-wide evaluations, dynamic test generation, and AI-assisted annotation. Third, we suggest guideline-structure updates that translate "what to evaluate" into "how to evaluate" through concrete linkage to resources, documentation standards for construction and usage, and industry alignment. Together, these contributions provide a roadmap toward more comprehensive and practical AI safety evaluation.

References

1. allenai/tulu-3-trustllm-jailbreaktrigger-eval, https://huggingface.co/datasets/allenai/tulu-3-trustllm-jailbreaktrigger-eval
2. Chemin-ai/reasoning_patterns_ai_hiring_bias_sea. https://huggingface.co/datasets/Supa-AI/Reasoning_Patterns_AI_Hiring_Bias_SEA
3. Guardrailsai/detect-jailbreak, https://huggingface.co/datasets/GuardrailsAI/detect-jailbreak
4. Guess-the-rule-llm-benchmark, https://github.com/m1chae11u/Guess-the-Rule-LLM-Benchmark
5. Huggingfacem4/m4-bias-eval-fair-face, https://huggingface.co/datasets/HuggingFaceM4/m4-bias-eval-fair-face
6. Huggingfacem4/m4-bias-eval-stable-bias, https://huggingface.co/datasets/HuggingFaceM4/m4-bias-eval-stable-bias
7. innodatalabs/rt2-jailbreakv-alpaca, https://huggingface.co/datasets/innodatalabs/rt2-jailbreakv-alpaca
8. Kingnish/deny-harmful-behaviour, https://huggingface.co/datasets/KingNish/deny-harmful-behaviour
9. ktiyab/diversity_equity_and_inclusion. https://huggingface.co/datasets/ktiyab/Diversity_Equity_and_Inclusion
10. Literary-language-models, https://github.com/teddyroland/Literary-Language-Models
11. Llm physical safety benchmark in drone control. https://huggingface.co/datasets/kumitang/llm_physical_safety_benchmark
12. llms-for-verified-programs, https://github.com/omkar-ethz/llms-for-verified-programs
13. Pardisszah/biasmd, https://huggingface.co/datasets/PardisSzah/BiasMD
14. qusgo/harmmetric_eval. https://huggingface.co/datasets/qusgo/HarmMetric_Eval
15. shubham-899844/visualization_literacy_test. https://github.com/shubham-899844/Visualization_Literacy_Test
16. svannie678/red_team_repo_social_bias_dataset_information. https://huggingface.co/datasets/svannie678/red_team_repo_social_bias_dataset_information
17. svannie678/red_team_repo_social_bias_prompts. https://huggingface.co/datasets/svannie678/red_team_repo_social_bias_prompts

18. tillschwoerer/toxic-comments. https://huggingface.co/datasets/tillschwoerer/toxic-comments
19. Artificial intelligence risk management framework (ai rmf 1.0). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf (2023)
20. Cataloguing llm evaluations. https://aiverifyfoundation.sg/downloads/Cataloguing_LLM_Evaluations.pdf (2023)
21. Ai 600-1: Generative artificial intelligence profile. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf (2024)
22. Ai safety governance framework. https://www.tc260.org.cn/upload/2024-09-09/1725849192841090989.pdf (2024)
23. Guide to evaluation perspectives on ai safety. https://aisi.go.jp/assets/pdf/ai_safety_eval_v1.01_en.pdf (2024)
24. International scientific report on the safety of advanced ai: interim report. https://assets.publishing.service.gov.uk/media//6655982fdc15efdddf1a842f/international_scientific_report_on_/the_safety_of_advanced_ai_interim_report.pdf (2024)
25. Model ai governance framework for generative ai. https://aiverifyfoundation.sg/wp-content/uploads/2024/05/ /Model-AI-Governance-Framework-for-Generative-AI-May-2024-1-1.pdf (2024)
26. The general-purpose ai code of practice. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai (2025)
27. Ackerman, S., et al.: A novel metric for measuring the robustness of large language models in non-adversarial scenarios (2024), https://arxiv.org/abs/2408.01963
28. AI, Z.: harmful_behaviors (revision a008e6f). https://huggingface.co/datasets/ZySec-AI/harmful_behaviors (2025). https://doi.org/10.57967/hf/4295
29. Anantaprayoon, P., et al.: Evaluating gender bias of pre-trained language models in natural language inference by considering all labels. In: Proceedings of the LREC-COLING 2024 (2024), https://aclanthology.org/2024.lrec-main.566
30. Bhardwaj, R., Poria, S.: Red-teaming large language models using chain of utterances for safety-alignment (2023)
31. Blodgett, S.L.: Sociolinguistically Driven Approaches for Just Natural Language Processing. Ph.D. thesis, University of Massachusetts Amherst (2021). https://doi.org/https://doi.org/10.7275/20410631
32. Cantini, R., Orsino, A., Ruggiero, M., Talia, D.: Benchmarking adversarial robustness to bias elicitation in large language models: Scalable automated assessment with llm-as-a-judge. arXiv preprint arXiv:2504.07887 (2025)
33. Chao, P., Debenedetti, E., Robey, A., Andriushchenko, M., Croce, F., Sehwag, V., Dobriban, E., Flammarion, N., Pappas, G.J., Tramèr, F., Hassani, H., Wong, E.: Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In: NeurIPS Datasets and Benchmarks Track (2024), https://huggingface.co/datasets/JailbreakBench/JBB-Behaviors
34. Costanza-Chock, S., Raji, I.D., Buolamwini, J.: Who audits the auditors? recommendations from a field scan of the algorithmic auditing ecosystem. In: Proceedings of the 2022 FAccT. p. 1571–1583 (2022). https://doi.org/10.1145/3531146.3533213, https://doi.org/10.1145/3531146.3533213
35. De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., Kalai, A.T.: Bias in bios: A case study of semantic representation bias in a high-stakes setting. In: Proceedings of the

2019 FAccT. p. 120–128 (2019). https://doi.org/10.1145/3287560.3287572, http://dx.doi.org/10.1145/3287560.3287572
36. Ding, P., et al.: Hallu-pi: Evaluating hallucination in multi-modal large language models within perturbed inputs (2024)
37. Eriksson, M., Purificato, E., Noroozian, A., Vinagre, J., Chaslot, G., Gomez, E., Fernandez-Llorca, D.: Can we trust ai benchmarks? an interdisciplinary review of current issues in ai evaluation (2025), https://arxiv.org/abs/2502.06559
38. Fan, W., et al.: Goldcoin: Grounding large language models in privacy laws via contextual integrity theory. In: Proceedings of the 2024 EMNLP (2024)
39. Felkner, V.K., et al.: Towards winoqueer: Developing a benchmark for anti-queer bias in large language models (2022), https://arxiv.org/abs/2206.11484
40. Gehman, S., Gururangan, S., Sap, M., Choi, Y., Smith, N.A.: Realtoxicityprompts: Evaluating neural toxic degeneration in language models. arXiv preprint arXiv:2009.11462 (2020)
41. Gero, K., Butters, N., Bethke, A., Elsafoury, F.: A dataset to measure fairness in the sentiment analysis task. https://github.com/efatmae/SST_sentiment_fairness_data (2023)
42. Grey, M., Segerie, C.R.: Safety by measurement: A systematic literature review of ai safety evaluation methods (2025), https://arxiv.org/abs/2505.05541
43. Gupta, K.: Whodunit: Evaluation benchmark for culprit detection in mystery stories (2025), https://arxiv.org/abs/2502.07747
44. Huang, Y., et al.: Trustllm: Trustworthiness in large language models. In: Forty-first International Conference on Machine Learning (2024), https://openreview.net/forum?id=bWUU0LwwMp
45. Jacovi, A., Wang, A., Alberti, C., Tao, C., Lipovetz, J., Olszewska, K., Haas, L., Liu, M., Keating, N., Bloniarz, A., Saroufim, C., Fry, C., Marcus, D., Kukliansky, D., Tomar, G.S., Swirhun, J., Xing, J., Wang, L., Aaron, M., Ambar, M., Fellinger, R., Wang, R., Sims, R., Zhang, Z., Goldshtein, S., Matias, Y., Das, D.: Facts leaderboard. https://kaggle.com/facts-leaderboard (2024), google DeepMind, Google Research, Google Cloud, Kaggle
46. Jindal, M., Deshpande, S.: Reveal: Multi-turn evaluation of image-input harms for vision llm (2025), https://arxiv.org/abs/2505.04673
47. Kocielnik, R., Prabhumoye, S., Zhang, V., Jiang, R., Alvarez, R.M., Anandkumar, A.: Biastestgpt: Using chatgpt for social bias testing of language models (2023), https://arxiv.org/abs/2302.07371
48. Kurkowski, M.: Contextual text anonymizer dataset (2025), https://huggingface.co/datasets/kurkowski/synthetic-contextual-anonymizer-dataset
49. Li, G., Zhang, Y., Wang, Y., Yan, S., Wang, L., Wei, T.: Priv-qa: Privacy-preserving question answering for cloud large language models (2025), https://arxiv.org/abs/2502.13564
50. Lin, Z., et al.: Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation (2023)
51. Liu, Z., Quan, Y.: Efficient retrieval of temporal event sequences from textual descriptions. arXiv preprint arXiv:2410.14043 (2024)
52. Liu, Z., Quan, Y.: Tpp-llm: Modeling temporal point processes by efficiently fine-tuning large language models. arXiv preprint arXiv:2410.02062 (2024)
53. Marcus, G., Davis, E., Aaronson, S.: A very preliminary analysis of dall-e 2 (2022), https://arxiv.org/abs/2204.13807
54. Mazeika, M., et al.: Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. arXiv preprint arXiv:2402.04249 (2024)

55. Nadeau, D., Kroutikov, M., McNeil, K., Baribeau, S.: Benchmarking llama2, mistral, gemma and gpt for factuality, toxicity, bias and propensity for hallucinations (2024), https://arxiv.org/abs/2404.09785
56. Nangia, N., Vania, C., Bhalerao, R., Bowman, S.R.: CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In: Proceedings of the 2020 EMNLP. pp. 1953–1967 (2020). https://doi.org/10.18653/v1/2020.emnlp-main.154, https://aclanthology.org/2020.emnlp-main.154/
57. Nitin Aravind Birur, Divyanshu Kumar, T.B.P.H.S.A.: Deepseek geopolitical bias dataset (2025), https://huggingface.co/datasets/enkryptai/deepseek-geopolitical-bias-dataset/viewer
58. Ojewale, V., Steed, R., Vecchione, B., Birhane, A., Raji, I.D.: Towards ai accountability infrastructure: Gaps and opportunities in ai audit tooling. In: Proceedings of the 2025 CHI. CHI '25 (2025). https://doi.org/10.1145/3706598.3713301, http://dx.doi.org/10.1145/3706598.3713301
59. Pilán, I., et al.: The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. CoRR **abs/2202.00443** (2022), https://arxiv.org/abs/2202.00443
60. Pozzobon, L., Lewis, P., Hooker, S., Ermis, B.: From one to many: Expanding the scope of toxicity mitigation in language models. arXiv preprint arXiv:2403.03893 (2024)
61. Priyanshu, A., Vijay, S.: Fractured-sorry-bench: Framework for revealing attacks in conversational turns undermining refusal efficacy and defenses over sorry-bench (2024), https://arxiv.org/abs/2408.16163
62. Quispe, D.: Vulnerability intelligence with diagrammatic reasoning (2025), https://huggingface.co/datasets/daqc/vulnerability-intelligence-diagrammatic-reasoning
63. Raji, I.D., Bender, E.M., Paullada, A., Denton, E., Hanna, A.: Ai and the everything in the whole wide world benchmark (2021), https://arxiv.org/abs/2111.15366
64. Rangapur, A., Wang, H., Shu, K.: Fin-fact: A benchmark dataset for multimodal financial fact checking and explanation generation (2023)
65. Reif, Y., Schwartz, R.: Fighting bias with bias: Promoting model robustness by amplifying dataset biases (2023), https://arxiv.org/pdf/2305.18917
66. Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., Kochenderfer, M.J.: BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. In: Advances in NeurIPS. vol. 37, pp. 21763–21813 (2024). https://doi.org/10.52202/079017-0685, https://proceedings.neurips.cc/paper_files/paper/2024/file/26889e8359e7ef8a7f5d77457364ca55-Paper-Datasets_and_Benchmarks_Track.pdf
67. R"ottger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., Pierrehumbert, J.: HateCheck: Functional tests for hate speech detection models. In: Proceedings of the 11th IJCNLP. pp. 41–58 (2021). https://doi.org/10.18653/v1/2021.acl-long.4, https://aclanthology.org/2021.acl-long.4
68. Sathe, A., Jain, P., Sitaram, S.: A unified framework and dataset for assessing societal bias in vision-language models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 1208–1249 (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.66, https://aclanthology.org/2024.findings-emnlp.66/

69. Schiappa, M.C., Vyas, S., Palangi, H., Rawat, Y.S., Vineet, V.: Robustness analysis of video-language models against visual and language perturbations. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022), https://openreview.net/forum?id=A79jAS4MeW9
70. Schmalfuss, J., Oei, V., Mehl, L., Bartsch, M., Agnihotri, S., Keuper, M., Bruhn, A.: Robustspring: Benchmarking robustness to image corruptions for optical flow, scene flow and stereo (2025), https://arxiv.org/abs/2505.09368
71. Schröer, S.L., Apruzzese, G., Human, S., Laskov, P., Anderson, H.S., Bernroider, E.W.N., Fass, A., Nassi, B., Rimmer, V., Roli, F., Salam, S., Ashley Shen, C.E., Sunyaev, A., Wadhwa-Brown, T., Wagner, I., Wang, G.: Sok: On the offensive potential of ai. In: 2025 IEEE SaTML. pp. 247–280 (2025). https://doi.org/10.1109/SaTML64287.2025.00021
72. Scirè, A., Bejgu, A.S., Tedeschi, S., Ghonim, K., Martelli, F., Navigli, R.: Truth or mirage? towards end-to-end factuality evaluation with llm-oasis (2024), https://arxiv.org/abs/2411.19655
73. Shao, Y., et al.: Privacylens: Evaluating privacy norm awareness of language models in action (2024), https://arxiv.org/abs/2409.00138
74. Sharma, R.K., Gupta, V., Grossman, D.: Spml: A dsl for defending language models against prompt attacks (2024)
75. Shen, X., Chen, Z., Backes, M., Shen, Y., Zhang, Y.: "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In: Proceedings of the 2024 CCS. p. 1671–1685 (2024). https://doi.org/10.1145/3658644.3670388, https://doi.org/10.1145/3658644.3670388
76. Siddique, Z., Turner, L., Espinosa-Anke, L.: Who is better at math, jenny or jingzhen? uncovering stereotypes in large language models. In: Proceedings of the 2024 EMNLP. pp. 18601–18619 (2024). https://doi.org/10.18653/v1/2024.emnlp-main.1035, https://aclanthology.org/2024.emnlp-main.1035/
77. Smith, E.M., et al.: " i ' m sorry to hear that ": Finding new biases in language models with a holistic descriptor dataset. In: Proceedings of the 2022 EMNLP (2022)
78. Song, M., Sim, S.H., Bhardwaj, R., Chieu, H.L., Majumder, N., Poria, S.: Measuring and enhancing trustworthiness of llms in rag through grounded attributions and learning to refuse (2024), https://arxiv.org/abs/2409.11242
79. Srirag, D., othersi: Evaluating dialect robustness of language models via conversation understanding (2024)
80. Tang, K., Zhou, W., Zhang, J., Liu, A., Deng, G., Li, S., Qi, P., Zhang, W., Zhang, T., Yu, N.: Gendercare: A comprehensive framework for assessing and reducing gender bias in large language models (2025), https://arxiv.org/abs/2408.12494
81. Team, T.E.A.: txt-image-bias-dataset (2025), https://huggingface.co/datasets/enkryptai/txt-image-bias-dataset
82. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.: BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021), https://openreview.net/forum?id=wCu6T5xFjeJ
83. Ushio, A., Alva-Manchego, F., Camacho-Collados, J.: Generative language models for paragraph-level question generation. In: Proceedings of the 2022 EMNLP. pp. 670–688 (2022). https://doi.org/10.18653/v1/2022.emnlp-main.42, https://aclanthology.org/2022.emnlp-main.42/

84. Wahréus, J., Hussain, A.M., Papadimitratos, P.: CySecBench: Generative AI-based CyberSecurity-focused Prompt Dataset for Benchmarking Large Language Models. arXiv preprint arXiv:2501.01335 (2025), https://arxiv.org/abs/2501.01335
85. Wan, S., Nikolaidis, C., Song, D., Molnar, D., Crnkovich, J., Grace, J., Bhatt, M., Chennabasappa, S., Whitman, S., Ding, S., Ionescu, V., Li, Y., Saxe, J.: Cyberseceval 3: Advancing the evaluation of cybersecurity risks and capabilities in large language models (2024), https://arxiv.org/abs/2408.01605
86. Wei, J., et al.: Long-form factuality in large language models (2024), https://arxiv.org/abs/2403.18802
87. Wingarz, T., Lauscher, A., Edinger, J., Kaaser, D., Schulte, S., Fischer, M.: Sok: Towards security and safety of edge ai (2024), https://arxiv.org/abs/2410.05349
88. Xu, Y., Wang, D., Yu, M., Ritchie, D., Yao, B., Wu, T., Zhang, Z., Li, T.J.J., Bradford, N., Sun, B., Hoang, T.B., Sang, Y., Hou, Y., Ma, X., Yang, D., Peng, N., Yu, Z., Warschauer, M.: Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In: Proceedings of the 60th ACL. pp. 447–460 (2022). https://doi.org/10.18653/v1/2022.acl-long.34, https://aclanthology.org/2022.acl-long.34/
89. Xue, P., Wu, L., Yu, Z., Jin, Z., Yang, Z., Li, X., Yang, Z., Tan, Y.: Automated commit message generation with large language models: An empirical study and beyond (2024), https://arxiv.org/abs/2404.14824
90. Yan, T.L., Jia, R.: Promote, suppress, iterate: How language models answer one-to-many factual queries. arXiv preprint arXiv:2502.20475 (2025), https://huggingface.co/papers/2502.20475
91. Zahraei, P.S., Shakeri, Z.: Detecting bias and enhancing diagnostic accuracy in large language models for healthcare (2024), https://arxiv.org/abs/2410.06566
92. Zakizadeh, M., et al.: Difair: A benchmark for disentangled assessment of gender knowledge and bias. In: EMNLP 2023 (2023). https://doi.org/10.18653/v1/2023.findings-emnlp.127, https://aclanthology.org/2023.findings-emnlp.127
93. Zhang, Y., et al.: Benchmarking trustworthiness of multimodal large language models: A comprehensive study (2024)
94. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.: Gender bias in coreference resolution: Evaluation and debiasing methods. CoRR **abs/1804.06876** (2018), http://arxiv.org/abs/1804.06876

A List of Sources of Benchmark Datasets

1. Realtoxicityprompts: Evaluating neural toxic degeneration in language models
   arXiv:2009.11462
2. ToxicChat: Unveiling Hidden Challenges of Toxicity Detection in Real-World User-AI Conversation
   arXiv:2310.17389
3. From One to Many: Expanding the Scope of Toxicity Mitigation in Language Models
   arXiv:2403.03893

4. Benchmarking Llama2, Mistral, Gemma and GPT for Factuality, Toxicity, Bias and Propensity for Hallucinations
arXiv:2404.09785
5. REVEAL: Multi-turn Evaluation of Image-Input Harms for Vision LLM
arXiv:2505.04673
6. Red-Teaming Large Language Models using Chain of Utterances for Safety-Alignment
arXiv:2308.09662
7. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal
arXiv:2402.04249
8. harmful_behaviors (Revision a008e6f)
DOI:10.57967/hf/4295
9. KingNish/deny-harmful-behaviour
https://huggingface.co/datasets/KingNish/deny-harmful-behaviour
10. qusgo/HarmMetric_Eva
https://huggingface.co/datasets/qusgo/HarmMetric_Eval
11. Long-form factuality in large language models
arXiv:2403.18802
12. Fin-Fact: A Benchmark Dataset for Multimodal Financial Fact Checking and Explanation Generation
arXiv:2309.08793
13. Truth or Mirage? Towards End-to-End Factuality Evaluation with LLM-OASIS
arXiv:2411.19655
14. FACTS Leaderboard
https://kaggle.com/facts-leaderboard
15. Promote, Suppress, Iterate: How Language Models Answer One-to-Many Factual Queries
arXiv:2502.20475
16. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models
DOI:10.18653/v1/2020.emnlp-main.154
17. GenderCARE: A Comprehensive Framework for Assessing and Reducing Gender Bias in Large Language Models
arXiv:2408.12494
18. Evaluating Gender Bias of Pre-trained Language Models in Natural Language Inference by Considering All Labels
https://aclanthology.org/2024.lrec-main.566
19. Towards WinoQueer: Developing a Benchmark for Anti-Queer Bias in Large Language Models
arXiv:2206.11484
20. Who is better at math, Jenny or Jingzhen? Uncovering Stereotypes in Large Language Models
DOI:10.18653/v1/2024.emnlp-main.1035

21. Detecting Bias and Enhancing Diagnostic Accuracy in Large Language Models for Healthcare
arXiv:2410.06566
22. A Unified Framework and Dataset for Assessing Societal Bias in Vision-Language Models
DOI:10.18653/v1/2024.findings-emnlp.66
23. DiFair: A Benchmark for Disentangled Assessment of Gender Knowledge and Bias
DOI:10.18653/v1/2023.findings-emnlp.127
24. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods
arXiv:1804.06876
25. Fighting Bias with Bias: Promoting Model Robustness by Amplifying Dataset Biases
arXiv:2305.18917
26. HuggingFaceM4/m4-bias-eval-stable-bias
https://huggingface.co/datasets/HuggingFaceM4/m4-bias-eval-stable-bias
27. HuggingFaceM4/m4-bias-eval-fair-face
https://huggingface.co/datasets/HuggingFaceM4/m4-bias-eval-fair-face
28. BiasTestGPT: Using ChatGPT for Social Bias Testing of Language Models
arXiv:2302.07371
29. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting
DOI:10.1145/3287560.3287572
30. "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset
DOI:10.18653/v1/2022.emnlp-main.625
31. svannie678/red_team_repo_social_bias_prompts
https://huggingface.co/datasets/svannie678/red_team_repo_social_bias_prompts
32. svannie678/red_team_repo_social_bias_dataset_information
https://huggingface.co/datasets/svannie678/red_team_repo_social_bias_dataset_information
33. PardisSzah/BiasMD
https://huggingface.co/datasets/PardisSzah/BiasMD
34. txt-image-bias-dataset
https://huggingface.co/datasets/enkryptai/txt-image-bias-dataset
35. DeepSeek Geopolitical Bias Dataset
https://huggingface.co/datasets/enkryptai/deepseek-geopolitical-bias-dataset/viewer
36. Sociolinguistically Driven Approaches for Just Natural Language Processing
DOI:https://doi.org/10.7275/20410631
37. Benchmarking Adversarial Robustness to Bias Elicitation in Large Language Models: Scalable Automated Assessment with LLM-as-a-Judge
arXiv:2504.07887
38. Chemin-AI/Reasoning_Patterns_AI_Hiring_Bias_SEA
https://huggingface.co/datasets/Supa-AI/Reasoning_Patterns_AI_Hiring_Bias_SEA

39. Fantastic Questions and Where to Find Them: FairytaleQA – An Authentic Dataset for Narrative Comprehension
DOI:10.18653/v1/2022.acl-long.34
40. A dataset to measure fairness in the sentiment analysis task
https://github.com/efatmae/SST_sentiment_fairness_data
41. ktiyab/Diversity_Equity_and_Inclusion
https://huggingface.co/datasets/ktiyab/Diversity_Equity_and_Inclusion
42. A Novel Metric for Measuring the Robustness of Large Language Models in Non-adversarial Scenarios
arXiv:2408.01963
43. Benchmarking Trustworthiness of Multimodal Large Language Models: A Comprehensive Study
arXiv:2406.07057
44. PRIV-QA: Privacy-Preserving Question Answering for Cloud Large Language Models
arXiv:2502.13564
45. GoldCoin: Grounding Large Language Models in Privacy Laws via Contextual Integrity Theory
DOI:10.18653/v1/2024.emnlp-main.195
46. PrivacyLens: Evaluating Privacy Norm Awareness of Language Models in Action
arXiv:2409.00138
47. The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization
arXiv:2202.00443
48. Contextual Text Anonymizer Dataset
https://huggingface.co/datasets/kurkowski/synthetic-contextual-anonymizer-dataset
49. Vulnerability Intelligence with Diagrammatic Reasoning
https://huggingface.co/datasets/daqc/vulnerability-intelligence-diagrammatic-reasoning
50. SPML: A DSL for Defending Language Models Against Prompt Attacks
arXiv:2402.11755
51. CyberSecEval 3: Advancing the Evaluation of Cybersecurity Risks and Capabilities in Large Language Models
arXiv:2408.01605
52. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models
https://huggingface.co/datasets/JailbreakBench/JBB-Behaviors
53. CySecBench: Generative AI-based CyberSecurity-focused Prompt Dataset for Benchmarking Large Language Models
arXiv:2501.01335
54. innodatalabs/rt2-jailbreakv-alpaca
https://huggingface.co/datasets/innodatalabs/rt2-jailbreakv-alpaca
55. "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models
DOI:10.1145/3658644.3670388

56. FRACTURED-SORRY-Bench: Framework for Revealing Attacks in Conversational Turns Undermining Refusal Efficacy and Defenses over SORRY-Bench
arXiv:2408.16163
57. allenai/tulu-3-trustllm-jailbreaktrigger-eval
https://huggingface.co/datasets/allenai/tulu-3-trustllm-jailbreaktrigger-eval
58. GuardrailsAI/detect-jailbreak
https://huggingface.co/datasets/GuardrailsAI/detect-jailbreak
59. shubham-899844/Visualization_Literacy_Test
https://github.com/shubham-899844/Visualization_Literacy_Test
60. TrustLLM: Trustworthiness in Large Language Models
https://openreview.net/forum?id=bWUU0LwwMp
61. Measuring and Enhancing Trustworthiness of LLMs in RAG through Grounded Attributions and Learning to Refuse
arXiv:2409.11242
62. TrustSafeAI/llm_physical_safety_benchmark
https://huggingface.co/datasets/TrustSafeAI/llm_physical_safety_benchmark
63. Robustness Analysis of Video-Language Models Against Visual and Language Perturbations
https://openreview.net/forum?id=A79jAS4MeW9
64. Evaluating Dialect Robustness of Language Models via Conversation Understanding
arXiv:2405.05688
65. Guess-the-Rule-LLM-Benchmark
https://github.com/m1chae11u/Guess-the-Rule-LLM-Benchmark
66. WHODUNIT: Evaluation benchmark for culprit detection in mystery stories
arXiv:2502.07747
67. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models
https://openreview.net/forum?id=wCu6T5xFjeJ
68. RobustSpring: Benchmarking Robustness to Image Corruptions for Optical Flow, Scene Flow and Stereo
arXiv:2505.09368
69. Literary-Language-Models
https://github.com/teddyroland/Literary-Language-Models
70. Automated Commit Message Generation with Large Language Models: An Empirical Study and Beyond
arXiv:2404.14824
71. Hallu-PI: Evaluating Hallucination in Multi-modal Large Language Models within Perturbed Inputs
arXiv:2408.01355
72. Generative Language Models for Paragraph-Level Question Generation
DOI:10.18653/v1/2022.emnlp-main.42
73. llms-for-verified-programs
https://github.com/omkar-ethz/llms-for-verified-programs