SafetyPrompts: A Systematic Review of Open Datasets for Evaluating and Improving Large Language Model Safety

Paul Röttger[1], **Fabio Pernisi**[1], **Bertie Vidgen**[2], **and Dirk Hovy**[1]

[1]Bocconi University
[2]Contextual AI

Abstract

The last two years have seen a rapid growth in concerns around the safety of large language models (LLMs). Researchers and practitioners have met these concerns by creating an abundance of datasets for evaluating and improving LLM safety. However, much of this work has happened in parallel, and with very different goals in mind, ranging from the mitigation of near-term risks around bias and toxic content generation to the assessment of longer-term catastrophic risk potential. This makes it difficult for researchers and practitioners to find the most relevant datasets for their use case, and to identify gaps in dataset coverage that future work may fill. To remedy these issues, we conduct a first systematic review of open datasets for evaluating and improving LLM safety. We review 144 datasets, which we identified through an iterative and community-driven process over the course of several months. We highlight patterns and trends, such as a trend towards fully synthetic datasets, as well as gaps in dataset coverage, such as a clear lack of non-English and naturalistic datasets. We also examine how LLM safety datasets are used in practice – in LLM release publications and popular LLM benchmarks – finding that current evaluation practices are highly idiosyncratic and make use of only a small fraction of available datasets. Our contributions are based on SafetyPrompts.com, a living catalogue of open datasets for LLM safety, which we plan to update continuously as the field of LLM safety develops.

1 Introduction

Ensuring the safety of large language models (LLMs) has become as a key priority for model developers and regulators. Consequently, in recent years, researchers and practitioners have created an abundance of datasets for evaluating and improving LLM safety. Safety, however, is a multi-faceted and contextual concept that lacks a unifying definition (Aroyo et al. 2023). This complexity is reflected in the current landscape of safety datasets, which is broad, diverse, and fast-changing. Within just two months of 2024, for example, researchers published datasets for evaluating near-term risks from LLMs, such as sociodemographic bias (Gupta et al. 2024) and toxic content generation (Bianchi et al. 2024), as well as datasets for evaluating long-term societal risk potential, around power-seeking (Mazeika et al. 2024) and sycophantic behaviours (Sharma et al. 2024). This rapid pace of dataset creation, and the variety of purposes served by different datasets, make it difficult for researchers and practitioners to find the most relevant datasets for different use cases, and to identify gaps in dataset coverage that future work may fill.

In this paper, we seek to address these issues by conducting a **first systematic review of open datasets for evaluating and improving LLM safety**. We identify 144 datasets published between June 2018 and December 2024 based on clear inclusion criteria (§2.1) using a comprehensive community-driven search method (§2.2). We examine these 144 datasets along several key dimensions, including their purpose (§3.2), intended use (§3.3), format and size (§3.4), their creation (§3.5), language (§3.6), licensing and access (§3.7), and publication (§3.8). Our review shows that a growing interest in LLM safety is driving the creation of more and more diverse open LLM safety datasets, with most contributions coming from academia. Major outstanding challenges, on the other hand, include a clear lack of safety datasets in non-English languages as well as a lack of naturalistic safety evaluations. We also review how open LLM safety datasets are used in practice – in model release publications (§4) and popular LLM benchmarks (§5) – finding that current evaluation practices are highly idiosyncratic and leverage only a small fraction of available datasets. We argue that this creates clear scope for standardisation in LLM safety evaluations, and that model safety evaluations in general could be improved by better leveraging recent progress in dataset creation (§6).

2 Dataset Review Methodology

2.1 Inclusion Criteria

At a high level, we restrict our review to *open* datasets that are relevant to *LLMs*, and specifically relevant to evaluating and improving LLM *safety*. More specifically, this means:

In terms of **data modality**, we only include text datasets. We do not include image datasets (e.g. Schwemmer et al. 2020; Zhao, Wang, and Russakovsky 2021; Ricker et al. 2022) or audio datasets (e.g. Reimao and Tzerpos 2019; Koenecke et al. 2020; Meyer et al. 2020). We also do not include datasets targeted at multimodal models, even if one modality is text, such as in the case of vision-language (e.g. Carlini et al. 2023; Hall et al. 2023; Wolfe et al. 2023) or text-to-image models (e.g. Bianchi et al. 2023; Parrish et al. 2023; Luccioni et al. 2024). Further, we do not include datasets tar-

geted at code generation models (e.g. Siddiq and Santos 2022; Bhatt et al. 2023). These modalities and models constitute natural expansions for future work.

We make only minimal restrictions in terms of **data format**. Real-world user interactions with LLMs usually take the form of text chat (Ouyang et al. 2023; Zheng et al. 2024; Zhao et al. 2024b), so we are most interested in datasets that naturally fit a chat format, like open-ended questions and instructions, but we also consider any other dataset that can meaningfully be expressed in a prompt format. This includes multiple-choice questions or autocomplete-style text snippets. We do not make any restrictions on dataset language.

For **data access**, we only include datasets that are publicly available for download via GitHub and/or Hugging Face. We do not make restrictions based on how data is licensed.

Finally, we require that all datasets are **relevant to safety**. For the purposes of our review, we adopt a wide and open definition of safety. Broadly speaking, we include datasets that relate to representational, political or other forms of sociodemographic bias; to toxicity, malicious instructions or harmful advice; to hazardous behaviours like sycophancy or power-seeking; to alignment with social, moral or ethical values; or to adversarial LLM usage (e.g. red-teaming, jailbreaking, prompt hacking). We only include datasets that explicitly focus on one or multiple of these aspects of LLM safety. We do not include datasets that target general LLM capabilities like reasoning, language understanding, or code completion (e.g. Dua et al. 2019; Hendrycks et al. 2020; Chen et al. 2021). We also do not include datasets that target factuality in LLMs, unless they directly relate to safety, like in the case of generating misinformation (Souly et al. 2024) or measuring truthfulness (Lin, Hilton, and Evans 2022).[1]

The cutoff date for our review is December 17th, 2024. We only included datasets that were first published (i.e. made publicly available online) before this date.

2.2 Finding Dataset Candidates

We used an iterative and community-driven approach combined with snowball search to identify dataset candidates for inclusion in our review. In January 2024, we released a first version of SafetyPrompts.com, with an initial list of 44 datasets that we had compiled in a heuristic fashion based on prior work and our knowledge of the LLM safety field. Over the next months, we marketed the website to the LLM safety community on Twitter and Reddit, to solicit feedback and further dataset suggestions. This resulted in 51 additional datasets, suggested by many different researchers and practitioners. We then used these 95 datasets as a starting point for snowball search, wherein we reviewed each publication corresponding to each dataset for references to other potentially relevant datasets. Whenever we identified a new dataset, we repeated this process. This resulted in 49 additional datasets. **Overall, our review includes 144 open datasets for evaluating and improving LLM safety**, which were first published between June 2018 and December 2024.

[1]Each dataset candidate was reviewed by two authors of this paper. We make detailed information on each dataset, as well as excluded datasets + exclusion reasons, available in the project repo.

We chose this review method because of two main reasons. First, LLM safety is a very fast-moving field with contributions from across academia and industry. By sharing intermittent results of our review on SafetyPrompts.com, we were able to solicit feedback from a broad range of stakeholders and expand our review, while also providing a useful resource to the community well ahead of the release of this paper. Second, traditional systematic review methods like keyword search are ill-suited to the scope of our review. Combinations of relevant keywords like "language model", "safety" and "dataset" return thousands of results on Google Scholar and similar platforms – and still fail to capture the many types of datasets that may not mention "safety" but are highly relevant to it regardless, like toxic conversation datasets or bias evaluations. Despite our best efforts, it is likely that our review is missing at least some relevant datasets. We are committed to adding these datasets, along with future relevant dataset releases, to SafetyPrompts.com.

2.3 Recording Structured Information

For each of the 144 datasets in our review, we recorded 23 pieces of structured information. At a high level, our goal was to capture the full development pipeline of each dataset: from how the dataset was created, to what entries in the dataset look like, what the dataset can or should be used for, how it can be accessed, and where it was published. We show the full codebook in the Appendix, which describes the structure and content of our main review spreadsheet. We make the spreadsheet available along with code to reproduce our analyses at github.com/paul-rottger/safetyprompts-paper.

3 Dataset Review Findings

3.1 History and Growth of Safety Datasets

Our review shows that **LLM safety builds on a rich history of research into the risks and biases of language models and dialogue agents**. The first datasets in our review were published in mid-2018, and focus on evaluating gender bias – originally for co-reference resolution systems, but equally applicable to current LLMs (Zhao et al. 2018; Rudinger et al. 2018). These datasets, in turn, build on earlier works on biases in word embeddings (e.g. Bolukbasi et al. 2016; Caliskan, Bryson, and Narayanan 2017; Garg et al. 2018), which illustrates that concerns around the negative social impacts of language models are far from new. Even the term "safety" was already used by Dinan et al. (2019) and Rashkin et al. (2019), among others, who introduced datasets for evaluating and improving the safety of dialogue agents well before the current generative LLM paradigm. By today's standards, however, interest in safety was relatively low at the time, with only 9 out of the 144 datasets in our review (6.3%) published in or before 2020, as shown in Figure 1.

LLM safety experienced a first moderate growth phase in 2021 and 2022. These two years, respectively, saw the publication of 16 and 17 open LLM safety datasets. This coincides with increased interest in generative language models, particularly among researchers, following the release of GPT-3 in mid-2020 (Brown et al. 2020).

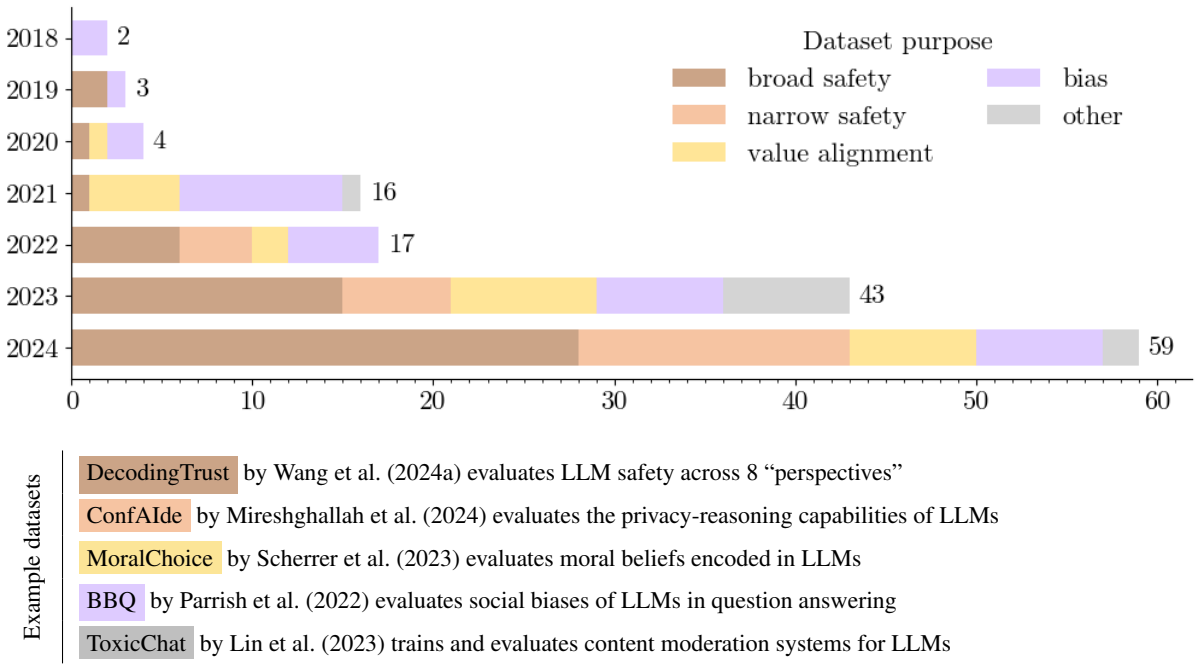

Figure 1: (top) **Number of datasets published per year, grouped by their primary purpose.** Our review includes 144 datasets published between June 2018 and December 2024. (bottom) **Example datasets** for each type of dataset purpose (§3.2).

Right now, **research into LLM safety is experiencing unprecedented growth**. This coincides with a surge in public interest in LLMs as well as concerns around LLM safety following the release of ChatGPT in November 2022. 43 out of the 144 datasets in our review (29.9%) were published in 2023, and 2024 saw even more activity, with 59 datasets published. Accordingly, it seems very likely that 2025 will surpass this record once more.

3.2 Purpose of Datasets

In our review, we differentiate between five distinct dataset purposes: **Broad safety** (n=53) denotes datasets that cover several aspects of LLM safety. This includes structured evaluation datasets like SafetyKit (Dinan et al. 2022) or SimpleSafetyTests (Vidgen et al. 2024b) as well as broad-scope red-teaming datasets like BAD (Xu et al. 2021) or AnthropicRedTeam (Ganguli et al. 2022). **Narrow safety** (n=25), conversely, denotes datasets that focus only on one specific aspect of LLM safety. SafeText (Levy et al. 2022), for example, focuses only on commonsense physical safety, while SycophancyEval (Sharma et al. 2024) focuses on sycophantic behaviour. **Value alignment** (n=23) refers to datasets that are concerned with the ethical, moral or social behaviour of LLMs. This includes datasets that seek to evaluate LLM understanding of ethical norms, like Scruples (Lourie, Le Bras, and Choi 2021) and ETHICS (Hendrycks et al. 2021), as well as opinion surveys like GlobalOpinionQA (Durmus et al. 2024). **Bias** (n=33) refers to datasets for evaluating sociodemographic biases in LLMs. BOLD (Dhamala et al. 2021), for example, evaluates bias in text completions, whereas DiscrimEval (Tamkin et al. 2023) evaluates biases in situated LLM decision-making. **Other** (n=10), in our review, includes datasets for developing LLM chat moderation systems, like FairPrism (Fleisig et al. 2023) and ToxicChat (Lin et al. 2023), as well as specialised prompts from public competitions like Gandalf (LakeraAI 2023a), Mosscap (LakeraAI 2023b) or HackAPrompt (Schulhoff et al. 2023).

Figure 1 shows that **early safety datasets were primarily concerned with evaluating biases**. 14 out of 25 datasets (56.0%) published between 2018 and 2021 were created to identify and analyse sociodemographic biases in language models. 13 of these datasets evaluate gender biases, either exclusively (e.g. Nozza, Bianchi, and Hovy 2021) or along with other categories of bias such as race and sexual orientation (e.g. Sheng et al. 2019) or religion (e.g. Li et al. 2020).

Broad safety emerged as a prominent theme in 2022, driven by industry contributions. Anthropic, for instance, released two broad-scope red-teaming datasets (Ganguli et al. 2022; Bai et al. 2022), while Meta published datasets on positive LLM conversations (Ung, Xu, and Boureau 2022) and general safety evaluation (Dinan et al. 2022). Most recently, broad safety has shifted towards more structured evaluation, as exemplified by benchmarks like DecodingTrust (Wang et al. 2024a) or HarmBench (Mazeika et al. 2024).

There is now a trend towards more specialised safety evaluations. Narrow safety evaluations did not emerge until 2022, but now make up a significant portion of all new datasets. In 2024 alone, 15 of the 59 datasets in our review (25.4%) were concerned with specific aspects of LLM safety, like rule-following (Mu et al. 2024) or privacy-reasoning ability (Mireshghallah et al. 2024).

Relatedly, **there is an increasing focus on adversarial robustness and LLM "jailbreaking"**. A growing body of work introduces safety datasets that are explicitly designed to elicit unsafe responses even from models trained to respond safely, often using elaborate prompting strategies, also known as "jailbreaks" (Wei, Haghtalab, and Steinhardt 2023). Since August 2023, we record 13 such datasets, which differ from most other datasets that seek to evaluate safety in response to simpler, more naturalistic prompts (e.g. Bhardwaj and Poria 2023; Shaikh et al. 2023; Vidgen et al. 2024b).

3.3 Intended Use of Datasets

Most datasets are intended for model evaluation only. 112 out of the 144 datasets in our review (77.8%) were created for benchmarking or evaluation. Only 6 datasets (4.2%), by contrast, comprise examples of positive interactions between users and LLMs, created specifically for model training (Rashkin et al. 2019; Ung, Xu, and Boureau 2022; Kim et al. 2022; Bianchi et al. 2024; Guo et al. 2024; Jiang et al. 2024). Overall, there is much more work on evaluating LLM safety than there is work on improving LLM safety.

3.4 Dataset Format and Size

The format of LLM safety datasets has changed alongside general trends in LLM development. Early datasets, especially those created for bias evaluation, often used autocomplete-style formats (n=9), where models are tasked to either fill a masked word or finish a sentence snippet (e.g. Sheng et al. 2019; Dhamala et al. 2021; Kirk et al. 2021). Such formats, which are most compatible with early language models like BERT or GPT-2, have since mostly been replaced by chat-style prompts, like open-ended questions and instructions (n=88), as well as promptable multiple-choice questions (n=24), which better suit current generative LLMs.

Dataset size varies substantially across the 144 datasets in our review. Smaller datasets ($\leq$1k entries, n=42) tend to be narrow safety evaluation datasets, written by the dataset authors. For example, the smallest dataset – ControversialInstructions by Bianchi et al. (2024) – comprises 40 author-written prompts to test the propensity of LLMs to generate hate speech. Larger datasets ($\geq$100k entries, n=21), on the other hand, are often broader in scope, designed for training and evaluation, and created through templating or LLM generation (§3.5). For example, the largest dataset – WorldValuesBench by Zhao et al. (2024a) – contains 21.5m multiple-choice prompts created programmatically from individual responses to the large-scale World Value Survey.

3.5 Dataset Creation

The use of templates is a consistently popular method for generating safety evaluation data. 26 out of the 144 datasets in our review (18.1%) consist of human-written phrase or prompt templates, which are expanded through combination into larger evaluation datasets. HolisticBias (Smith et al. 2022), for example, comprises 26 sentence templates (e.g. "I am an [attribute1] who is [attribute2]."), which are combined with around 600 individual descriptor terms to create 459.8k test cases. Generally, template approaches are most popular for bias evaluation – 16 out of 33 bias evaluation datasets in our review use a template approach – but recent work has also used templates for evaluating broad-scope LLM safety (Wang et al. 2024a) or specific concerns such as privacy reasoning (Mireshghallah et al. 2024).

A significant portion of recently-released datasets is fully synthetic. 21 of the 102 datasets released since 2023 (20.6%) consist of chat prompts or multiple-choice questions generated entirely by LLMs – most commonly by some variant of GPT-3.5. Shaikh et al. (2023), for example, prompt GPT-3.5 to generate 200 harmful questions, which they use to explore safety in chain-of-thought question answering. The resulting dataset resembles the human-written SimpleSafetyTests (Vidgen et al. 2024b). Generally, synthetic datasets vary in size and scope much like human-written datasets do.

Relatedly, instead of relying on static templates for data creation, **many recent datasets are more flexibly augmented using LLMs**. Bhatt et al. (2023), for instance, expand a small expert-written set of cyberattack instructions into a larger set of 1,000 prompts using Llama-70b-chat (Touvron et al. 2023). Wang et al. (2024a) take a similar approach to build their large-scale DecodingTrust benchmark.

By contrast, **very few datasets comprise naturalistic user interactions with LLMs**. Around half of the 144 datasets in our review are hand-written (n=70, 48.6%), but mostly by the authors of the corresponding dataset publications. A much smaller proportion of datasets is created by humans – either crowdworkers (e.g. Ganguli et al. 2022; Aroyo et al. 2023; Kirk et al. 2024) or participants of online competitions (e.g. LakeraAI 2023a,b) – interacting more naturally with LLMs.

Finally, **there is a trend towards reusing existing datasets rather than creating original data**. 34 out of the 59 datasets published in 2024 (57.6%) were entirely (n=10) or partially (n=24) sampled from older datasets. QHarm by Bianchi et al. (2024), for example, is taken entirely from AnthropicHarmlessBase (Bai et al. 2022). Another common approach is to sample existing data and augment it with templates (e.g. Tedeschi et al. 2024) or LLM generations (e.g. Yu et al. 2024). By contrast, just 10 out of 43 datasets published in 2023 (23.3%) reused existing data. This shift suggests a growing maturity of the LLM safety dataset landscape.

3.6 Dataset Languages

The vast majority of open LLM safety datasets uses English language only. 113 out of 144 datasets in our review (78.5%) are exclusively in English. 10 datasets (6.9%) focus only on Chinese (e.g. Zhou et al. 2022; Xu et al. 2023; Zhao et al. 2023), and one dataset each focuses only on Arabic, Swedish, Hindi, Korean, and French. The 16 other datasets (11.1%) cover English along with one or more other languages. Jin et al. (2024b) cover 106 languages, although test cases are auto-translated from the English. The other 143 datasets in our review together cover 31 languages.

3.7 Data Licensing and Access

When data is shared, usage licenses are mostly permissive. The most common license is the very permissive MIT License, which is used for 58 out of 144 datasets (40.3%). 23 datasets (16.0%) use the Apache 2.0 License, which provides

	Academic / Non-Profit Org.	n
1	Stanford University	15
1	Allen Institute for AI	15
3	UC Berkeley	14
4	University of Washington	13
5	Carnegie Mellon University	12

	Industry Org.	n
1	Meta* (prev. Facebook)	13
2	Anthropic	9
3	Microsoft* (incl. Research)	7
4	Google* (incl. DeepMind)	6
5	Alibaba	4

Table 1: **Organisations that published the most open LLM safety datasets** among the 144 datasets in our review. For each dataset, we count all affiliations for all co-authors. Stanford at n=15, for example, means that 15 datasets had a Stanford author.

additional patent protections. 43 datasets (29.9%) use variants of a CC BY 4.0 License, which requires dataset users to provide appropriate credit and indicate if changes were made to the dataset. Notably, 15 datasets (10.4%) prohibit commercial usage with a CC BY-NC License. Only 2 datasets (1.4%) use a more restrictive custom license. As of December 17th, 2024, 15 datasets (10.4%) do not specify any license.[2]

GitHub is still the most popular platform for sharing LLM safety data. 72 datasets (50.0%) are available only on GitHub. However, 55 datasets (38.2%) are available on both GitHub and HuggingFace, and several recent datasets are shared on Hugging Face only (e.g. Han et al. 2024; Manerba et al. 2024), suggesting a shift in how data is shared.

3.8 Dataset Publication Authors and Venues

Academic and non-profit organisations drive most of the creation of open LLM safety datasets. For 69 out of 144 datasets in our review (47.9%), all authors of the corresponding publication were affiliated only with academic or non-profit organisations. 45 datasets (31.3%) included authors from industry and academia, and only 30 datasets (20.8%) were published by fully industry teams.

The creation of LLM safety datasets is concentrated in few research hubs (Table 1). There are 156 unique affiliations across the authors of the 144 datasets in our review. 101 affiliations (64.7%) are associated with just a single dataset. The five most prolific organisations, on the other hand, are each associated with at least 13 datasets. All of the twenty most prolific organisations are located and/or headquartered in the US, with the exception of Bocconi University (Italy, n=11), the University of Oxford (UK, n=5), CUHK (HK, n=4), and Tsinghua University (China, n=4).

Most LLM safety datasets so far have been published at *ACL conferences. 68 out of the 144 datasets in our review (47.2%) were published at either ACL (n=27), EMNLP (n=30) or other *ACL venues. 33 datasets (19.7%) were published at more ML-focused conferences, i.e. NeurIPS (n=18), ICLR (n=10), or ICML (n=5). Only 11 datasets (7.6%) were published at other venues, and only 2 datasets appeared in journal publications (Jin et al. 2024a; Yu, Li, and Lan 2024). 28 datasets (19.4%), on the other hand, were accompanied only by arXiv preprints, and 4 (2.7%) only by blog posts, meaning they did not receive traditional peer review. Generally, we observe a slight trend away from *ACL conferences and towards more ML-focused venues as well as arXiv-only publication, although this could in part be explained by recent arXiv preprints still being under review.

4 Datasets in Model Release Publications

In the following, we briefly examine **how open LLM safety datasets are used in practice**. In particular, we examine in this section which safety datasets are used to evaluate current state-of-the-art LLMs ahead of their release, as documented in model release publications. In the next section (§5), we then examine which safety datasets are included in popular LLM benchmarks and leaderboards. This is to characterise current norms and common practices in evaluating LLM safety, so that we can then discuss (§6) these norms and practices in relation to the findings of our dataset review (§3).

4.1 Scope of Our Model Release Review

We examine the top 50 best-performing LLMs listed on the LMSYS Chatbot Arena Leaderboard (Chiang et al. 2024) as of July 25th, 2024.[3] The LMSYS Chatbot Arena Leaderboard is a crowdsourced platform for LLM evaluation, which ranks models based on model Elo scores calculated from over one million pairwise human preference votes. We use this leaderboard for setting the scope of our review because it is held in high regard in the LLM community, and it has up-to-date coverage of recent model releases.

The top 50 entries on the LMSYS leaderboard correspond to 29 unique model releases.[4] Of these 29 models, 16 (55.2%) are proprietary models only accessible via an API, released by companies such as OpenAI (GPT), Anthropic (Claude), Google (Gemini) and 01 AI (Yi). The other 13 models (44.8%) are open models, for which weights are publicly accessible via Hugging Face. Proprietary models generally outrank open models on the leaderboard, with Gemma 2 27b (DeepMind 2024) being the best open model at rank 14. All 29 models were released by industry labs.

4.2 Findings of Our Model Release Review

Around half of state-of-the-art LLMs are evaluated for safety ahead of their release, with substantial variation in

[2]While conducting our review, we reached out to authors of all datasets that had not specified a license and encouraged them to add one. At least five authors added a license as a result.

[3]https://lmarena.ai/?leaderboard

[4]A model release may comprise multiple model versions, such as GPT-4-0314 and GPT-4-0613.

	Dataset	Purpose	n
1	TruthfulQA (Lin, Hilton, and Evans 2022)	evaluate tendency to mimic human falsehoods	8
2	BBQ (Parrish et al. 2022)	evaluate social bias in question answering	6
3	AnthropicRedTeam (Ganguli et al. 2022)	evaluate responses to diverse red-team attacks	3
4	ToxiGen (Hartvigsen et al. 2022)	evaluate toxicity in text completions	3
5	RealToxicityPrompts (Gehman et al. 2020)	evaluate toxicity in text completions	3

Table 2: **Most popular open LLM safety datasets**, based on how often release publications for SOTA LLMs reported results on each dataset. BBQ at n=6, for example, means that 6 out of the 29 model release publications (§4.2) reported BBQ results.

the extent and nature of safety evaluations. For 16 out of 29 model releases (55.2%), model developers report at least some quantitative safety evaluation in the associated technical report, blog post or model card. 13 model release publications (44.8%) report results on at least one open LLM safety dataset. The Claude 3.5 model card, for example, reports results on WildChat (Zhao et al. 2024b) and XSTest (Röttger et al. 2024a), whereas Gemma 2 (DeepMind 2024) is evaluated on 7 different open LLM safety datasets. 13 out of the 29 models (44.8%), on the other hand, do not report any safety evaluations. This includes the strongest model, the proprietary GPT-4o, which mentions safety measures in its release blog post, but lacks any concrete quantitative safety evaluations. Other models, both proprietary (e.g. Yi-Large, Reka Core) and open (e.g. Mixtral, Command R), do not mention safety at all in their release publications.

When safety is evaluated, **proprietary data plays a large role in model release safety evaluations**. Out of the 16 model releases that report safety evaluation results, 11 (68.8%) use undisclosed proprietary data for evaluating model safety. 3 of these releases – Llama 3 (Meta 2024), Qwen 2 (Yang et al. 2024), and Phi 3 (Abdin et al. 2024) – report results only on proprietary safety datasets.

Finally, **the diversity of open LLM safety datasets used in model release evaluations is very limited**. Only 14 open LLM safety datasets are used across the 29 model releases, and 6 of these 14 datasets are used only once. Table 2 shows the 5 datasets that are used most often. Notably, TruthfulQA (Lin, Hilton, and Evans 2022) is used in 8 out of the 16 model release publications that report any safety evaluation results (50.0%), and it is often framed by model developers as a capability rather than a safety evaluation. We discuss the implications of these findings in §6.

5 Datasets Used in Popular Benchmarks

5.1 Scope of Our Benchmark Review

To complement our model release review, we briefly examine 9 popular LLM benchmarks for which safety datasets they include. 6 benchmarks are widely-used general-purpose benchmarking suites: Stanford's HELM Classic (Liang et al. 2023) and HELM Instruct (Zhang et al. 2024), Hugging Face's Open LLM Leaderboard (Beeching et al. 2023), AllenAI's RewardBench (Lambert et al. 2024), Eleuther AI's Evaluation Harness (Gao et al. 2021), and BIG-Bench (Srivastava et al. 2023). 3 benchmarks focus explicitly on LLM safety: TrustLLM (Sun et al. 2024), HELM Safety (Kaiyom et al. 2024), and the LLM Safety Leaderboard.[5]

5.2 Findings of Our Benchmark Review

There are large differences in how different benchmarks evaluate LLM safety. The 9 benchmarks make use of 26 open LLM safety datasets. 17 of these datasets are used in just one benchmark. TrustLLM (Sun et al. 2024), for example, combines 11 open LLM safety datasets, of which 5 are not used in any other benchmark. The only open LLM safety datasets that are used in more than 2 benchmarks are TruthfulQA (Lin, Hilton, and Evans 2022), which is used in 4 benchmarks, as well as RealToxicityPrompts (Gehman et al. 2020), BBQ (Parrish et al. 2022), ETHICS (Hendrycks et al. 2021), XSTest (Röttger et al. 2024a), and DoNotAnswer (Wang et al. 2024b), which are each used in 3 benchmarks.

There is currently no single LLM safety benchmark with a truly comprehensive scope. The TrustLLM benchmark (Sun et al. 2024), has the broadest scope relevant to safety among the 9 benchmarks we examined, covering malicious instruction-following, bias, and value alignment. However, it does not, for example, test for longer-term risk potential with evaluations for sycophancy (Perez et al. 2023) or chemical weapon knowledge (Li et al. 2024), as the Evaluation Harness does (Gao et al. 2021). Similarly, the LLM Safety Leaderboard (Beeching et al. 2023) tests for toxic content generation and malicious instruction-following, but not for false refusal (e.g. Röttger et al. 2024a) or sociodemographic biases (e.g. Parrish et al. 2022)

6 Discussion & Future Directions

6.1 The State of the Safety Dataset Landscape

Overall, our review shows that **growing interest in LLM safety is driving the creation of more and more diverse open LLM safety datasets**. More datasets were published in 2023 than ever before, and 2024 again surpassed this record (§3.1). Existing datasets span varied purposes (§3.2) and formats (§3.4), which have adapted over time to meet the needs and requirements of LLM users and developers. Researchers and practitioners are making creative use of new methods for dataset creation (§3.5), and when data is shared, usage licenses are mostly permissive (§3.7). These are encouraging signs for the health of the open LLM safety community and

[5] https://huggingface.co/spaces/AI-Secure/llm-trustworthy-leaderboard, based on Wang et al. (2024a).

its ability to address emerging challenges and fill gaps in dataset coverage as they become apparent.

Among these gaps, the most apparent today is that **there is a clear lack of safety datasets in non-English languages**. We found that English dominates the current safety dataset landscape (§3.6), mirroring long-standing trends in NLP research (Bender 2011; Joshi et al. 2020; Holtermann et al. 2024). To some extent, this imbalance also reflects an imbalance in who is publishing safety datasets (§3.8). The lack of non-English resources for evaluating and improving LLM safety is a problem because it means that billions of non-English speakers across the world are potentially more at risk of harm when using current language technologies. Further, interpretations of safety are known to vary across cultures and geographies (Aroyo et al. 2023; Kirk et al. 2024). These two factors create an urgent need for the LLM safety community to prioritise the creation of non-English datasets going forward. Recent datasets like AyaRedTeaming (Aakanksha et al. 2024), which were created for language-specific cultural contexts with the involvement of local stakeholders, may serve as a useful blueprint for future work in this direction.

The second major concern apparent from our review is that **current safety evaluation datasets largely fail to reflect real-world LLM usage**, which undermines their ecological validity. Data generation and augmentation, from templates to more recent LLM-based methods, have enabled rapid dataset creation (§3.5), but it is not clear that they can emulate the "messiness" and diversity that is apparent in real-world user interactions with LLMs (Ouyang et al. 2023; Zhao et al. 2024b; Zheng et al. 2024). An increasing focus on safety against "jailbreaking" (§3.2) may match the behaviour of sophisticated adversarial users, but does not capture the potential risks from vulnerable populations interacting with LLMs without malicious intent (Vidgen et al. 2024a). Addressing these concerns will require the creation of more naturalistic evaluations, and clearer communication about which user personas (e.g. vulnerable, malicious, or adversarial) are targeted by individual safety evaluations. Encouraging developments in this direction include WildGuard (Han et al. 2024) and WildJailBreak (Jiang et al. 2024), which provide evaluations based on real user prompts. Future work may similarly look to collections of real user interactions with LLMs, such as WildChat (Zhao et al. 2024b) and LMSYS (Zheng et al. 2024), to improve the ecological validity of LLM safety evaluations.

6.2 The Use of Safety Datasets in Practice

Our analysis of how open LLM safety datasets are used in practice shows that **there is clear scope for standardisation in LLM safety evaluations**. Safety is a key priority for model developers and users, as evidenced by the inclusion of safety evaluations in the majority of model release publications (§4) and popular LLM benchmarks (§5). However, there is much idiosyncrasy in *which* datasets are used for evaluating safety across model release publications and benchmarks. For commercial model releases, these datasets are often proprietary and undisclosed. More standardised and open evaluations, on the other hand, would enable more meaningful model comparisons and incentivise the development of safer LLMs.

We see three main directions towards this goal. First, current evaluation practices could **better leverage recent progress in safety dataset creation**. The most popular open datasets for evaluating safety in model release publications, for example, are all from 2020 or 2022 (Table 2), despite more than two thirds of the datasets in our review being published in or after 2023. Prior publication date is not a sign of lacking quality, but older autocomplete-style datasets like RealToxicityPrompts (Gehman et al. 2020) or ToxiGen (Hartvigsen et al. 2022) simply do not reflect realistic usage of current chat-optimised LLMs (Ouyang et al. 2023; Zheng et al. 2024; Zhao et al. 2024b), which undermines their ecological validity. Our review highlights many datasets that could be used in their stead. Second, there is a clear need for **unified standards of dataset quality**. In this paper, we refrained from making quality judgments about the datasets we reviewed, mainly because different datasets serve different purposes, so that their utility is highly context dependent. However, recent efforts like BetterBench (Reuel et al. 2024) as applied to safety datasets could help create consensus on which datasets to use for safety benchmarking. Third, safety research would benefit from **unified safety evaluation protocols**. The open-ended nature of safety-relevant behaviours complicates standardised evaluation compared to many factual tasks. Frameworks like Eleuther AI's Evaluation Harness, adapted for safety tasks, would make it easier for model developers and users to run the same safety evaluations with reproducible results, and thus serve to reduce idiosyncrasy in prevailing safety evaluation practices.

7 Conclusion

In recent years, researchers and practitioners have sought to meet concerns around the safety of large language models by creating an abundance of datasets for evaluating and improving LLM safety. However, the rapid pace of dataset creation and the variety of purposes served by different datasets have made it difficult for researchers and practitioners to find the most relevant datasets for different use cases, and to identify gaps in dataset coverage that future work may fill. In this paper, we addressed these issues by conducting a first systematic review of open LLM safety datasets.

In our review, which includes 144 datasets published between June 2018 and December 2024, we showed that, encouragingly, existing datasets span varied purposes and formats, which have adapted over time to meet the changing needs and requirements of LLM users and developers. However, we also highlighted major outstanding challenges, including a clear lack of non-English datasets as well as naturalistic safety evaluations. Further, when examining how open LLM safety datasets are used in practice – in model release publications and popular LLM benchmarks – we found that current evaluation practices are highly idiosyncratic and make use of only a small fraction of available datasets, which presents clear scope for rejuvenation and standardisation.

Overall, we hope that with our review, as well as the living dataset catalogue we make available on SafetyPrompts.com, we can enable such positive change, by helping researchers and practitioners make the best use of existing datasets, and providing a strong foundation for future dataset development.

Acknowledgments

Thank you for feedback and dataset suggestions to Giuseppe Attanasio, Steven Basart, Federico Bianchi, Marta R. Costa-Jussà, Daniel Hershcovic, Kexin Huang, Hyunwoo Kim, George Kour, Bo Li, Hannah Lucas, Marta Marchiori Manerba, Norman Mu, Niloofar Mireshghallah, Matus Pikuliak, Verena Rieser, Felix Röttger, Sam Toyer, Ryan Tsang, Pranav Venkit, Laura Weidinger, and Linhao Yu. Special thanks to Hannah Rose Kirk for the initial logo suggestion.

PR, FP, and DH are members of the Data and Marketing Insights research unit of the Bocconi Institute for Data Science and Analysis, and are supported by a MUR FARE 2020 initiative under grant agreement Prot. R20YSMBZ8S (INDOMITA) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 949944, INTEGRATOR).

For the appendices to this paper, including our codebook and a full list of all 144 datasets covered in our review, please see our SafetyPrompts arxiv preprint (Röttger et al. 2024b). AAAI formatting guidelines prohibit us from publishing the appendices here.

References

Aakanksha; Ahmadian, A.; Ermis, B.; Goldfarb-Tarrant, S.; Kreutzer, J.; Fadaee, M.; and Hooker, S. 2024. The Multilingual Alignment Prism: Aligning Global and Local Preferences to Reduce Harm. In *EMNLP 2024*, 12027–12049. ACL.

Abdin, M.; Jacobs, S. A.; Awan, A. A.; et al. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219.

Aroyo, L.; Taylor, A.; Diaz, M.; Homan, C.; Parrish, A.; Serapio-García, G.; Prabhakaran, V.; and Wang, D. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. *NeurIPS 2023*, 36.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv:2204.05862*.

Beeching, E.; Fourrier, C.; Habib, N.; Han, S.; Lambert, N.; Rajani, N.; Sanseviero, O.; Tunstall, L.; and Wolf, T. 2023. Open LLM Leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

Bender, E. M. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in LT*, 6.

Bhardwaj, R.; and Poria, S. 2023. Red-Teaming Large Language Models using Chain of Utterances for Safety-Alignment. arXiv:2308.09662.

Bhatt, M.; Chennabasappa, S.; Nikolaidis, C.; Wan, S.; Evtimov, I.; Gabi, D.; Song, D.; Ahmad, F.; Aschermann, C.; Fontana, L.; et al. 2023. Purple llama cyberseceval: A secure coding benchmark for language models. *arXiv preprint*.

Bianchi, F.; Kalluri, P.; Durmus, E.; Ladhak, F.; Cheng, M.; Nozza, D.; Hashimoto, T.; Jurafsky, D.; Zou, J.; and Caliskan, A. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *FAccT 2023*, 1493–1504. ACM.

Bianchi, F.; Suzgun, M.; Attanasio, G.; Rottger, P.; Jurafsky, D.; Hashimoto, T.; and Zou, J. 2024. Safety-Tuned LLaMAs: Lessons From Improving the Safety of Large Language Models that Follow Instructions. In *ICLR 2024*.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? *NeurIPS 2016*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *NeurIPS 2020*, 33: 1877–1901.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.

Carlini, N.; Nasr, M.; Choquette-Choo, C. A.; Jagielski, M.; Gao, I.; Koh, P. W.; Ippolito, D.; Tramèr, F.; and Schmidt, L. 2023. Are aligned neural networks adversarially aligned? In *NeurIPS 2023*.

Chen, M.; Tworek, J.; Jun, H.; et al. 2021. Evaluating large language models trained on code. *arXiv:2107.03374*.

Chiang, W.-L.; Zheng, L.; Sheng, Y.; Angelopoulos, A. N.; Li, T.; Li, D.; Zhang, H.; Zhu, B.; Jordan, M.; Gonzalez, J. E.; and Stoica, I. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. arXiv:2403.04132.

DeepMind, G. 2024. Gemma 2: Improving Open Language Models at a Practical Size.

Dhamala, J.; Sun, T.; Kumar, V.; Krishna, S.; Pruksachatkun, Y.; Chang, K.-W.; and Gupta, R. 2021. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *FAccT 2021*, 862–872. ACM.

Dinan, E.; Abercrombie, G.; Bergman, A.; Spruit, S.; Hovy, D.; Boureau, Y.-L.; and Rieser, V. 2022. SafetyKit: First Aid for Measuring Safety in Open-domain Conversational Systems. In *ACL 2022*, 4113–4133. ACL.

Dinan, E.; Humeau, S.; Chintagunta, B.; and Weston, J. 2019. Build it Break it Fix it for Dialogue Safety: Robustness from Adversarial Human Attack. In *EMNLP-IJCNLP 2019*, 4537–4546. ACL.

Dua, D.; Wang, Y.; Dasigi, P.; Stanovsky, G.; Singh, S.; and Gardner, M. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *NAACL 2019*, 2368–2378. ACL.

Durmus, E.; Nguyen, K.; Liao, T.; Schiefer, N.; et al. 2024. Towards Measuring the Representation of Subjective Global Opinions in Language Models. In *COLM 2024*.

Fleisig, E.; Amstutz, A.; Atalla, C.; Blodgett, S. L.; Daumé III, H.; Olteanu, A.; Sheng, E.; Vann, D.; and Wallach, H. 2023. FairPrism: Evaluating Fairness-Related Harms in Text Generation. In *ACL 2023*, 6231–6251. ACL.

Ganguli, D.; Lovitt, L.; Kernion, J.; Askell, A.; Bai, Y.; Kadavath, S.; Mann, B.; Perez, E.; Schiefer, N.; Ndousse, K.; et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv:2209.07858*.

Gao, L.; Tow, J.; Biderman, S.; et al. 2021. A framework for few-shot language model evaluation. In *Zenodo*. Zenodo.

Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *PNAS*, 115(16): E3635–E3644.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *EMNLP 2020 (Findings)*. ACL.

Guo, Y.; Cui, G.; Yuan, L.; Ding, N.; Sun, Z.; Sun, B.; Chen, H.; Xie, R.; Zhou, J.; Lin, Y.; Liu, Z.; and Sun, M. 2024. Controllable Preference Optimization: Toward Controllable Multi-Objective Alignment. In *EMNLP 2024*, 1437–1454. ACL.

Gupta, V.; Venkit, P. N.; Laurençon, H.; Wilson, S.; and Passonneau, R. J. 2024. CALM : A Multi-task Benchmark for Comprehensive Assessment of Language Model Bias. In *COLM 2024*.

Hall, S. M.; Abrantes, F. G.; Zhu, H.; Sodunke, G.; Shtedritski, A.; and Kirk, H. R. 2023. VisoGender: A dataset for benchmarking gender bias in image-text pronoun resolution. In *NeurIPS 2023 (D&B)*.

Han, S.; Rao, K.; Ettinger, A.; Jiang, L.; Lin, B. Y.; Lambert, N.; Choi, Y.; and Dziri, N. 2024. WildGuard: Open One-stop Moderation Tools for Safety Risks, Jailbreaks, and Refusals of LLMs. In *NeurIPS 2024 (D&B)*.

Hartvigsen, T.; Gabriel, S.; Palangi, H.; Sap, M.; Ray, D.; and Kamar, E. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *ACL 2022*, 3309–3326. ACL.

Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2021. Aligning AI With Shared Human Values. In *ICLR*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring Massive Multitask Language Understanding. In *ICLR 2020*.

Holtermann, C.; Röttger, P.; Dill, T.; and Lauscher, A. 2024. Evaluating the Elementary Multilingual Capabilities of Large Language Models with MultiQ. *arXiv:2403.03814*.

Jiang, L.; Rao, K.; Han, S.; Ettinger, A.; Brahman, F.; Kumar, S.; Mireshghallah, N.; Lu, X.; Sap, M.; Choi, Y.; and Dziri, N. 2024. WildTeaming at Scale: From In-the-Wild Jailbreaks to (Adversarially) Safer Language Models. In *NeurIPS 2024*.

Jin, J.; Kim, J.; Lee, N.; Yoo, H.; Oh, A.; and Lee, H. 2024a. KoBBQ: Korean Bias Benchmark for Question Answering. *TACL*, 12: 507–524.

Jin, Z.; Kleiman-Weiner, M.; Piatti, G.; Levine, S.; Liu, J.; Adauto, F. G.; Ortu, F.; Strausz, A.; Sachan, M.; Mihalcea, R.; Choi, Y.; and Schölkopf, B. 2024b. Multilingual Trolley Problems for Language Models. In *Pluralistic Alignment Workshop at NeurIPS 2024*.

Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; and Choudhury, M. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *ACL 2020*, 6282–6293. ACL.

Kaiyom, F.; Ahmed, A.; Mai, Y.; Klyman, K.; Bommasani, R.; and Liang, P. 2024. HELM Safety: Towards Standardized Safety Evaluations of Language Models. *Stanford Center for Research on Foundation Models*.

Kim, H.; Yu, Y.; Jiang, L.; Lu, X.; Khashabi, D.; Kim, G.; Choi, Y.; and Sap, M. 2022. ProsocialDialog: A Prosocial Backbone for Conversational Agents. In *EMNLP 2022*, 4005–4029. ACL.

Kirk, H. R.; Jun, Y.; Volpin, F.; Iqbal, H.; Benussi, E.; Dreyer, F.; Shtedritski, A.; and Asano, Y. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *NeurIPS 2021*, 34: 2611–2624.

Kirk, H. R.; Whitefield, A.; Röttger, P.; et al. 2024. The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. In *NeurIPS 2024 (D&B)*.

Koenecke, A.; Nam, A.; Lake, E.; Nudell, J.; Quartey, M.; Mengesha, Z.; Toups, C.; Rickford, J. R.; Jurafsky, D.; and Goel, S. 2020. Racial disparities in automated speech recognition. *PNAS*, 117(14): 7684–7689.

LakeraAI. 2023a. Gandalf Prompt Injection. In *Hugging Face Dataset*.

LakeraAI. 2023b. Mosscap Prompt Injection. In *Hugging Face Dataset*.

Lambert, N.; Pyatkin, V.; Morrison, J.; et al. 2024. RewardBench: Evaluating Reward Models for Language Modeling. arXiv:2403.13787.

Levy, S.; Allaway, E.; Subbiah, M.; Chilton, L.; Patton, D.; McKeown, K.; and Wang, W. Y. 2022. SafeText: A Benchmark for Exploring Physical Safety in Language Models. In *EMNLP*, 2407–2421. ACL.

Li, N.; Pan, A.; Gopal, A.; et al. 2024. The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning. In *ICML 2024*.

Li, T.; Khashabi, D.; Khot, T.; Sabharwal, A.; and Srikumar, V. 2020. UNQOVERing Stereotyping Biases via Underspecified Questions. In Cohn, T.; He, Y.; and Liu, Y., eds., *EMNLP 2020 (Findings)*, 3475–3489. ACL.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; et al. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *ACL 2022*, 3214–3252. ACL.

Lin, Z.; Wang, Z.; Tong, Y.; Wang, Y.; Guo, Y.; Wang, Y.; and Shang, J. 2023. ToxicChat: Unveiling Hidden Challenges of Toxicity Detection in Real-World User-AI Conversation. In *EMNLP 2023 (Findings)*, 4694–4702. ACL.

Lourie, N.; Le Bras, R.; and Choi, Y. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 13470–13479.

Luccioni, S.; Akiki, C.; Mitchell, M.; and Jernite, Y. 2024. Stable bias: Evaluating societal representations in diffusion models. *NeurIPS*, 36.

Manerba, M.; Stanczak, K.; Guidotti, R.; and Augenstein, I. 2024. Social Bias Probing: Fairness Benchmarking for Language Models. In *EMNLP 2024*, 14653–14671. ACL.

Mazeika, M.; Phan, L.; Yin, X.; Zou, A.; Wang, Z.; Mu, N.; Sakhaee, E.; Li, N.; Basart, S.; Li, B.; Forsyth, D.; and Hendrycks, D. 2024. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. In *ICML 2024*.

Meta. 2024. The Llama 3 Herd of Models.

Meyer, J.; Rauchenstein, L.; Eisenberg, J. D.; and Howell, N. 2020. Artie Bias Corpus: An Open Dataset for Detecting Demographic Bias in Speech Applications. In *LREC 2020*, 6462–6468. European Language Resources Association. ISBN 979-10-95546-34-4.

Mireshghallah, N.; Kim, H.; Zhou, X.; Tsvetkov, Y.; Sap, M.; Shokri, R.; and Choi, Y. 2024. Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory. In *ICLR 2024*.

Mu, N.; Chen, S.; Wang, Z.; Chen, S.; Karamardian, D.; Aljeraisy, L.; Alomair, B.; Hendrycks, D.; and Wagner, D. 2024. Can LLMs Follow Simple Rules? arXiv:2311.04235.

Nozza, D.; Bianchi, F.; and Hovy, D. 2021. HONEST: Measuring Hurtful Sentence Completion in Language Models. In *NAACL 2021*, 2398–2406. ACL.

Ouyang, S.; Wang, S.; Liu, Y.; et al. 2023. The Shifted and The Overlooked: A Task-oriented Investigation of User-GPT Interactions. In *EMNLP 2023*, 2375–2393. ACL.

Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. 2022. BBQ: A hand-built bias benchmark for question answering. In *ACL 2022 (Findings)*, 2086–2105. ACL.

Parrish, A.; Kirk, H. R.; Quaye, J.; et al. 2023. Adversarial Nibbler: A Data-Centric Challenge for Improving the Safety of Text-to-Image Models. *arXiv preprint*.

Perez, E.; Ringer, S.; Lukosiute, K.; ...; and Kaplan, J. 2023. Discovering Language Model Behaviors with Model-Written Evaluations. In *ACL 2023 (Findings)*, 13387–13434. ACL.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *ACL 2019*, 5370–5381. ACL.

Reimao, R.; and Tzerpos, V. 2019. For: A dataset for synthetic speech detection. In *SpeD 2019*, 1–10. IEEE.

Reuel, A.; Hardy, A.; Smith, C.; Lamparth, M.; Hardy, M.; and Kochenderfer, M. 2024. BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. In *NeurIPS 2024 (D&B)*.

Ricker, J.; Damm, S.; Holz, T.; and Fischer, A. 2022. Towards the detection of diffusion model deepfakes. *arXiv:2210.14571*.

Röttger, P.; Kirk, H.; Vidgen, B.; Attanasio, G.; Bianchi, F.; and Hovy, D. 2024a. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. In *NAACL 2024*, 5377–5400. ACL.

Röttger, P.; Pernisi, F.; Vidgen, B.; and Hovy, D. 2024b. Safetyprompts: a systematic review of open datasets for evaluating and improving large language model safety. *arXiv preprint arXiv:2404.05399*.

Rudinger, R.; Naradowsky, J.; Leonard, B.; and Van Durme, B. 2018. Gender Bias in Coreference Resolution. In *NAACL 2018*, 8–14. ACL.

Scherrer, N.; Shi, C.; Feder, A.; and Blei, D. 2023. Evaluating the Moral Beliefs Encoded in LLMs. In *NeurIPS 2023*.

Schulhoff, S.; Pinto, J.; Khan, A.; et al. 2023. Ignore This Title and HackAPrompt: Exposing Systemic Vulnerabilities of LLMs Through a Global Prompt Hacking Competition. In *EMNLP 2023*, 4945–4977. ACL.

Schwemmer, C.; Knight, C.; Bello-Pardo, E. D.; Oklobdzija, S.; Schoonvelde, M.; and Lockhart, J. W. 2020. Diagnosing gender bias in image recognition systems. *Socius*, 6: 2378023120967171.

Shaikh, O.; Zhang, H.; Held, W.; Bernstein, M.; and Yang, D. 2023. On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning. In *ACL 2023*, 4454–4470. ACL.

Sharma, M.; Tong, M.; Korbak, T.; et al. 2024. Towards Understanding Sycophancy in Language Models. In *ICLR 2024*.

Sheng, E.; Chang, K.-W.; Natarajan, P.; and Peng, N. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *EMNLP-IJCNLP 2019*, 3407–3412. ACL.

Siddiq, M. L.; and Santos, J. C. S. 2022. SecurityEval dataset: mining vulnerability examples to evaluate machine learning-based code generation techniques. In *MSR4P&S 2022*, 29–33. ACM.

Smith, E. M.; Hall, M.; Kambadur, M.; Presani, E.; and Williams, A. 2022. "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. In *EMNLP 2022*, 9180–9211. ACL.

Souly, A.; Lu, Q.; Bowen, D.; Trinh, T.; et al. 2024. A StrongREJECT for Empty Jailbreaks. In *ICLR 2024 (R2FM Workshop)*.

Srivastava, A.; Rastogi, A.; Rao, A.; et al. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *TMLR*.

Sun, L.; Huang, Y.; Wang, H.; Wu, S.; Zhang, Q.; Gao, C.; Huang, Y.; Lyu, W.; Zhang, Y.; Li, X.; et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint*.

Tamkin, A.; Askell, A.; Lovitt, L.; et al. 2023. Evaluating and mitigating discrimination in language model decisions. *arXiv:2312.03689*.

Tedeschi, S.; Friedrich, F.; Schramowski, P.; Kersting, K.; Navigli, R.; Nguyen, H.; and Li, B. 2024. ALERT: A Comprehensive Benchmark for Assessing Large Language Models' Safety through Red Teaming. *arXiv:2404.08676*.

Touvron, H.; Lavril, T.; Izacard, G.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv:2302.13971*.

Ung, M.; Xu, J.; and Boureau, Y.-L. 2022. SaFeRDialogues: Taking Feedback Gracefully after Conversational Safety Failures. In *ACL 2022*, 6462–6481. ACL.

Vidgen, B.; Agrawal, A.; Ahmed, A. M.; et al. 2024a. Introducing v0.5 of the AI Safety Benchmark from MLCommons. arXiv:2404.12241.

Vidgen, B.; Scherrer, N.; Kirk, H. R.; Qian, R.; Kannappan, A.; Hale, S. A.; and Röttger, P. 2024b. SimpleSafetyTests: a Test Suite for Identifying Critical Safety Risks in Large Language Models. arXiv:2311.08370.

Wang, B.; Chen, W.; Pei, H.; Xie, C.; Kang, M.; Zhang, C.; Xu, C.; Xiong, Z.; Dutta, R.; Schaeffer, R.; et al. 2024a. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. *NeurIPS*, 36.

Wang, Y.; Li, H.; Han, X.; Nakov, P.; and Baldwin, T. 2024b. Do-Not-Answer: Evaluating Safeguards in LLMs. In Graham, Y.; and Purver, M., eds., *EACL 2024 (Findings)*, 896–911. ACL.

Wei, A.; Haghtalab, N.; and Steinhardt, J. 2023. Jailbroken: How Does LLM Safety Training Fail? *arXiv:2307.02483*.

Wolfe, R.; Yang, Y.; Howe, B.; and Caliskan, A. 2023. Contrastive Language-Vision AI Models Pretrained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias. In *FAccT 2023*, 1174–1185. ACM.

Xu, G.; Liu, J.; Yan, M.; et al. 2023. Cvalues: Measuring the values of chinese large language models from safety to responsibility. *arXiv:2307.09705*.

Xu, J.; Ju, D.; Li, M.; Boureau, Y.-L.; Weston, J.; and Dinan, E. 2021. Bot-Adversarial Dialogue for Safe Conversational Agents. In *NAACL 2021*, 2950–2968. ACL.

Yang, A.; Yang, B.; Hui, B.; et al. 2024. Qwen2 Technical Report. arXiv:2407.10671.

Yu, E.; Li, J.; Liao, M.; Wang, S.; Zuchen, G.; Mi, F.; and Hong, L. 2024. CoSafe: Evaluating Large Language Model Safety in Multi-Turn Dialogue Coreference. In *EMNLP 2024*, 17494–17508. ACL.

Yu, J.; Li, L.; and Lan, Z. 2024. Beyond Binary Classification: A Fine-grained Safety Dataset for Large Language Models. *IEEE Access*.

Zhang, Y.; Mai, Y.; Roberts, J. S.; Bommasani, R.; Dubois, Y.; and Liang, P. 2024. HELM Instruct: A Multidimensional Instruction Following Evaluation Framework with Absolute Ratings.

Zhao, D.; Wang, A.; and Russakovsky, O. 2021. Understanding and evaluating racial biases in image captioning. In *ICCV 2021*, 14830–14840.

Zhao, J.; Fang, M.; Shi, Z.; Li, Y.; Chen, L.; and Pechenizkiy, M. 2023. CHBias: Bias Evaluation and Mitigation of Chinese Conversational Language Models. In *ACL 2023*, 13538–13556. ACL.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *NAACL 2018*, 15–20. ACL.

Zhao, W.; Mondal, D.; Tandon, N.; Dillion, D.; Gray, K.; and Gu, Y. 2024a. WorldValuesBench: A Large-Scale Benchmark Dataset for Multi-Cultural Value Awareness of Language Models. In *LREC-COLING 2024*, 17696–17706. ELRA and ICCL.

Zhao, W.; Ren, X.; Hessel, J.; Cardie, C.; Choi, Y.; and Deng, Y. 2024b. (InThe)WildChat: 570K ChatGPT Interaction Logs In The Wild. In *ICLR 2024*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Li, T.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Li, Z.; Lin, Z.; Xing, E.; Gonzalez, J. E.; Stoica, I.; and Zhang, H. 2024. LMSYS-1M: A Large-Scale Real-World LLM Conversation Dataset. In *ICLR 2024*.

Zhou, J.; Deng, J.; Mi, F.; Li, Y.; Wang, Y.; Huang, M.; Jiang, X.; Liu, Q.; and Meng, H. 2022. Towards Identifying Social Bias in Dialog Systems: Framework, Dataset, and Benchmark. In *EMNLP 2022 (Findings)*, 3576–3591. ACL.