In Benchmarks We Trust ... Or Not?

Ine Gevers[1], Victor De Marez[1], Jens Van Nooten[1], Jens Lemmens[1], Andriy Kosar[1], Ehsan Lotfi[1], Nikolay Banar[1], Pieter Fivez [1,2], Luna De Bruyne[1], Walter Daelemans[1],
[1]CLiPS, University of Antwerp, [2]TEXTUA,
Correspondence: ine.gevers@uantwerpen.be

Abstract

Standardized benchmarks are central to evaluating and comparing model performance in Natural Language Processing (NLP). However, Large Language Models (LLMs) have exposed shortcomings in existing benchmarks, and so far there is no clear solution. In this paper, we survey a wide scope of benchmarking issues, and provide an overview of solutions as they are suggested in the literature. We observe that these solutions often tackle a limited number of issues, neglecting other facets. Therefore, we propose concrete checklists to cover all aspects of benchmarking issues, both for benchmark creation and usage. We illustrate the use of our checklists by applying them to three popular NLP benchmarks (i.e., SuperGLUE, WinoGrande, and ARC-AGI). Additionally, we discuss the potential advantages of adding minimal-sized test-suites to benchmarking, which would ensure downstream applicability on real-world use cases.

1 Introduction

There is a rich history of benchmarking in Natural Language Processing (NLP): the field has seen an evolution from specific single-task and single-domain to more general multi-task benchmarks, following the advent of more powerful general-purpose AI models (Ruder, 2021). These benchmarks have been used as an attempt to objectively assess the performance of methods, and to track and direct progress in the field (e.g., the yearly AI Index Report, Maslej et al., 2025). In its broad sense, a benchmark is a dataset (or an ensemble of datasets) associated with one or multiple metrics, and a way to aggregate system performances (Ruder, 2021). The performance on such a benchmark is considered to be representative of the model's abilities on a task, and is used by the research community as a framework to compare methods (Raji et al., 2021). Prominent standardized benchmarks in NLP are used to promote the increasing capabilities of newly released models: technical reports introducing new Large Language Models (LLMs) often refer to their performance on a collection of standardized benchmarks (e.g., Achiam et al., 2023; Yang et al., 2024, etc.). However, recent models are outpacing the benchmark creation and benchmarks are quickly saturated, but this does not necessarily mean the model has grasped the relevant skill or knowledge (Kiela et al., 2021). Additionally, since benchmark scores have become a goal on their own, research integrity could be compromised in an attempt to optimize these scores. For instance, the LLaMA 4 team submitted 27 private variants of the model (Singh et al., 2025) to Chatbot Arena (Chiang et al., 2024), which artificially boosted the benchmark scores and obscured the distinction between the publicly released version and their best performer on this benchmark.[1] Koch and Peterson (2024) argue that the rigid consolidation of benchmarking as the sole evaluator of progress also disincentivizes exploration beyond scaling model size.

So far, there is no clear consensus on how to address the problems with benchmarking: for instance, the HuggingFace Open LLM Leaderboard introduced a way to evaluate methods across a range of tasks and metrics (Myrzakhan et al., 2024), but it eventually became outdated and is now archived.[2] Meanwhile, research is ongoing to improve existing benchmarks (e.g., by adversarial sampling, or semantic deduplication) or creating new ones (e.g., ARC-AGI, Chollet et al., 2024).

In this paper, we focus on benchmarking issues from the perspective of benchmark integrity (benchmark creation) and benchmarking practices (evaluating a method on a benchmark). While we address issues that are generally relevant regardless

[1] https://x.com/lmarena_ai/status/1909397817434816562

[2] At the time of writing: https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 23662–23676
November 4-9, 2025 ©2025 Association for Computational Linguistics

of the model type that is being evaluated, we zoom in on benchmarks currently used to evaluate LLMs because these models reveal inconsistencies and weaknesses in standardized benchmarks that were created earlier. Since benchmarks vary in format and modality, we focus here on text-based benchmarks: text as input, and text as output. These benchmarks can come in many shapes, such as classification, summarization, generation, and so on. In the scope of our paper, we consider static benchmarks that have an a priori gold label associated with each input text.

Existing position and survey papers on NLP benchmarking have provided important insights. For example, Bowman and Dahl (2021) propose core criteria for Natural Language Understanding (NLU) benchmark design. Raji et al. (2021) focus on construct validity and inappropriate community use of benchmarks given two main benchmarks, ImageNet and (Super)GLUE. McIntosh et al. (2024) address functionality and integrity of 23 benchmarks in the context of generative LLMs, while Laskar et al. (2024) examine the robustness of LLM evaluation. Banerjee et al. (2024) analyze contamination and gaming in evaluation frameworks. Finally, Reuel-Lamparth et al. (2024) propose an assessment framework that covers a wide range of AI benchmarks and provide a checklist for minimum quality assurance.

This survey paper adds to this effort by (1) surveying discrete benchmarking issues both in *creation* and *usage* without any a priori benchmark selection, (2) surveying solutions that are suggested in the literature and evaluating whether these solutions are general enough for (most of) the issues we identified, and (3) combining these insights into a concrete checklist of mitigation strategies, and exploring the added value of integrating downstream test-suites as an additional test to ensure model generalizability.

2 Survey of benchmarking flaws

In this section, we provide a survey of various problems with benchmarking that have been noted in prior literature over time. We structure them according to four types of experiment validity they undermine (Wohlin et al., 2012): (1) *internal validity*, whether results are caused by the variable(s) of interest rather than by external confounds; (2) *external validity*, whether results are generalizable to other domains, or real-world settings; (3) *statistical validity*, whether the proper methods are applied to evaluate the model outputs on the benchmark, so that the reported metrics support the claim; (4) *construct validity*, whether the task and evaluation metrics capture the phenomenon they intend to measure.

2.1 Internal validity

Benchmarks are... only as good as their annotations Since benchmark datasets are designed to compare the performance of models on one or more specific tasks, it is crucial that the provided annotations are of high quality. If this is not the case, this comparison is not only invalid, but this also has implications for the performance of the models if they are fine-tuned on this benchmark. Moreover, Vendrow et al. (2024) show that label errors cause evaluation inconsistencies, by hiding unreliable model behavior.

In manually annotated data, one of the main causes of low annotation quality is annotator disagreement, which can occur in spite of (or because of) annotator guidelines and an extensive training procedure (Parmar et al., 2023). A second possible cause is annotator bias, which is the result of demographic and personal factors (Al Kuwatly et al., 2020).

Alternatively, automatic labeling through distant supervision may provide high-quality labels in some tasks, such as Native Language Identification. In this setup, labels are inferred from metadata associated with the input's author (i.e., their declared native language). If the gold label is not straightforward, however, labels obtained via distant supervision can be problematic. In sarcasm detection, for example, labels provided by the authors (e.g., '#sarcasm') can be mined, although this may lead to inconsistent examples of sarcasm (Loakman et al., 2023). In addition, there is no way to estimate the recall of such methods, which could lead to unrepresentative sampling from the population (Ghosh et al., 2020). As an alternative, it has been proposed to use LLMs as automatic annotators, but as argued by Felkner et al. (2024), these models are biased themselves.

Benchmarks are... gameable Dataset artifacts are superficial patterns in the data that can be exploited by the model to get the correct answer based on irrelevant correlations (Gardner et al., 2021), which is not necessarily intended by the researcher. The presence of such superficial patterns especially becomes problematic when the evaluation metrics of the benchmark encourages shortcuts. In a classic example, Mao and Lee (2019) show that in many paraphrasing datasets, repeating the input text inflated the score. Also in Natural Language Inference (NLI) tasks, models could already partially solve the task without looking at the premise at all, instead relying on lexical patterns or sentence lengths (Gururangan et al., 2018). Newer models still apply 'shortcut learning' in NLI, in which they for instance exploit lexical overlap (Yuan et al., 2024; Sun et al., 2024). In a multiple-choice question answering setup, the position of the correct answer among the possible options can also be exploited. By shuffling this order, Pezeshkpour and Hruschka (2024) observe an 85% performance drop. Similarly, Alzahrani et al. (2024) can move models up or down 8 ranks on the MMLU dataset with various small perturbations.

Recently, reasoning and explainability benchmarks were introduced to increase the transparency of LLM behavior, but they are still gameable. For example, Hsia et al. (2024) describe various methods to manipulate specific evaluation metrics such as ERASER and EVAL-X scores. Meanwhile, Mondorf and Plank (2024b) discuss how using accuracy as a metric to measure reasoning performance can obscure how LLMs rely on surface-level patterns and correlations in the training data, rather than on sophisticated reasoning abilities.

Benchmarks are... trained upon LLM benchmark evaluation is increasingly compromised by data contamination, where models are exposed to benchmark data during training (Xu et al., 2024). This leakage encompasses various forms, from entire datasets to metadata about them (Xu et al., 2024; Sainz et al., 2023b; Palavalli et al., 2024). This phenomenon is widespread, affecting popular benchmarks (such as HellaSwag and TriviaQA) within common training corpora (such as C4 and The Pile), both in open and closed source models (Sainz et al., 2023b, 2024; Singh et al., 2024a).

Detecting and mitigating contamination, which can occur during pre-training, fine-tuning, or user feedback updates (Sainz et al., 2023b; Xu et al., 2024; Balloccu et al., 2024), is challenging due to dataset scale and model opacity (Sainz et al., 2023b, 2024). Methods include string/embedding matching for open data (Xu et al., 2024; Ravaut et al., 2024), while closed models often require probing distributions and logits, or analyzing memorization (Sainz et al., 2023b; Xu et al., 2024; Sainz et al., 2023a). The growing recognition of this issue is reflected in dedicated workshops and surveys (Sainz et al., 2024; Ravaut et al., 2024; Cheng et al., 2025).

The consequences of contamination are severe: inflated scores, unfair comparisons, flawed scientific conclusions, potential performance degradation, and practical risks such as commercial, privacy, or copyright (Xu et al., 2024; Zhou et al., 2023; Sainz et al., 2023b; Cheng et al., 2025; Ravaut et al., 2024). Therefore, it is crucial to mitigate contamination through better data curation, e.g., through private/dynamic benchmarks, encryption, or licensing (Xu et al., 2024; Jacovi et al., 2023); refactoring existing benchmarks and benchmark-free evaluation, like LLM-as-judge (Xu et al., 2024; Cheng et al., 2025); and procedural safeguards such as transparency and community registries (Jacovi et al., 2023; Balloccu et al., 2024; Sainz et al., 2023b).

2.2 External validity

Benchmarks are... Anglocentric The linguistic scope of current evaluations is notably limited. Most benchmarks focus predominantly on English or a small set of high-resource languages, overlooking the vast global linguistic landscape (McIntosh et al., 2024). The few existing benchmarks for low-resource languages – such as the Uhura benchmark for evaluating question answering in six African languages (Bayes et al., 2024), or LingOly for assessing linguistic reasoning in 90 low-resource languages (Bean et al., 2024) – have demonstrated significant performance declines when LLMs are applied to under-resourced languages. Therefore, it is crucial to evaluate models on a much wider and diverse range of languages.

Early efforts to expand language coverage in benchmarks primarily relied on machine translation of existing benchmarks (Lai et al., 2023; Thellmann et al., 2024). Although this approach is fast and cost-effective, translation quality can negatively affect the validity of evaluation results (Engländer et al., 2024; Plaza et al., 2024; Singh et al., 2024b). The NLP community recognizes this issue by reducing the machine-translated content (Singh et al., 2024b) and developing human-curated evaluation resources (Enevoldsen et al., 2025). However, even human translation might leave undesirable artifacts of the source language in the translated texts ('translationese'), which is detrimental to model performance (Barth and Rehm, 2025) and might obscure evaluation results. Besides the issue of translationese, translating data may leave cultural traces from the source text, posing particular challenges in tasks involving subjectivity or cultural nuance, such as in emotion detection (De Bruyne, 2023) and topic detection (Kosar et al., 2024).

This cultural bias in benchmarks is also problematic, as many evaluations implicitly embed specific cultural norms and assume homogeneity in language use and worldview, which does not reflect reality. Singh et al. (2024b) reveal that performance on widely used MMLU is largely tied to the knowledge of Western-centric ideas, with 28% of the questions involving culturally specific information. In addition, among questions testing geographic knowledge, a striking 84.9% focus on North America or Europe. Research utilizing the BLEnD benchmark (Myung et al., 2024) highlights stark performance disparities when models process culturally diverse inputs. Additionally, studies involving AraDiCE (Mousi et al., 2025) and work by Wang et al. (2024) expose how dialectal variation and cultural context are frequently ignored or improperly handled, leading to inconsistent or inappropriate model evaluation.

Benchmarks are... corrupted The quality of the input texts is as important as the quality of the annotations, but this is not always guaranteed. Bowman and Dahl (2021) highlight that some tasks, such as NLI, occur infrequently in a natural setting (such as in social media data or product reviews). In such cases, research opts for crowd-sourcing data or generating synthetic data using LLMs. However, these approaches can cause multiple issues. For instance, crowd-sourced data is prone to contain duplicate or repetitive entries (Bowman and Dahl, 2021). Additionally, even though LLM-generated synthetic data can be an attractive alternative, this data is often biased and insufficiently representative for more complex tasks (Maheshwari et al., 2024).

Besides repetitiveness, another problem that benchmarks face is that they can contain harmful texts and data in violation of privacy and copyright laws (Rogers et al., 2021; Longjohn et al., 2024). Longjohn et al. (2024) posit that extensive quality reviews, sharing metadata and creating repositories for benchmarks can mitigate these emerging issues through updates or deprecation.

Benchmarks are... focused on the same domains Existing benchmarks are heavily skewed towards academic or general-purpose tasks. Specialized domains such as finance, legal, medical, biology, or arts receive limited attention. The existence of domain-specific models, such as BloombergGPT (Wu et al., 2023), FinLlama (Konstantinidis et al., 2024), LawLLM (Shu et al., 2024), and their superior performance compared to general-purpose models underscores the inadequacy of generic benchmarks for capturing specialized, task-specific expertise. Moreover, the lack of model generalization across domains is illustrated by performance on benchmarks like LexEval for the legal domain (Li et al., 2024), FinBen for finance (Xie et al., 2024) or a range of medical benchmarks (Pal and Sankarasubbu, 2024).

2.3 Statistical validity

Benchmarks are... evaluated too inconsistently Current evaluations of LLMs face significant inconsistencies and unreliable findings due to the complexity and variability across different benchmark evaluation setups. For instance, Mizrahi et al. (2024) show that LLMs are sensitive to prompt design, exposing a significant performance difference across benchmarks when the instruction template is paraphrased. Further, Laskar et al. (2024) describe how multiple sources of variance exist within the evaluation pipeline, including differences in prompt design and the configuration of decoding parameters, which can substantially impact reported performance. According to their criteria, only 20.7% of 212 surveyed papers sufficiently control for this variance to arrive at a fair model comparison.

Benchmarks are... reported without significance testing Recent surveys find that in most applications of AI benchmarks, including NLP ones, statistical significance testing is omitted when presenting their results (Reuel-Lamparth et al., 2024). This undermines the validity, utility and trustworthiness of these results (Biderman et al., 2024), as it remains crucial to distinguish random noise from genuine performance differences between models. For example, recent work by Zhang et al. (2024) demonstrates that the absence of statistical significance testing can obscure benchmark contamination effects in LLMs, leading to potentially misleading conclusions about model performance.

2.4 Construct validity

Benchmarks are... not representative Reliable LLM evaluation is challenged by the representativeness problem: a growing disconnect between benchmark performance and real-world capabilities (Church, 2020; Nezhurina et al., 2025). This gap stems from poor construct validity, where benchmarks are flawed proxies for general abilities, creating an *illusion of generality* (Raji et al., 2021). For instance, models frequently demonstrate high performance on standardized test-style questions yet struggle when faced with complex planning or multi-step reasoning challenges, a limitation highlighted by specialized benchmarks like PlanBench (Valmeekam et al., 2022), GPQA (Rein et al., 2023), and HLE (Phan et al., 2025).

This validity issue is compounded by an inherent evaluation bias stemming from the tension between ensuring reproducibility (favoring easily quantifiable, repeatable metrics) and achieving functionality (accurately assessing intended capabilities and real-world alignment) (McIntosh et al., 2024). The strong emphasis on reproducibility often leads to an over-reliance on convenient formats like multiple-choice question answering (MCQA). This approach reframes generative tasks as classification problems, simplifying evaluation (as seen in LegalBench, Guha et al., 2023) but significantly compromising validity, as MCQA is not a neutral setting (Balepur et al., 2024) and remains a poor proxy for real-world performance even when debiased (Cho et al., 2025; Gu et al., 2024).

Relying on these flawed, easily quantifiable proxies fosters an overfixation on leaderboard rankings, incentivizing 'benchmark gaming': optimizing specific metrics rather than cultivating genuine understanding or robust capabilities (Burden, 2024; McIntosh et al., 2024; Singh et al., 2025), a phenomenon consistent with Goodhart's Law (Burden, 2024). This results in models that appear strong on paper but are brittle in practice, failing unexpectedly when faced with minor variations unseen in the benchmarks (Nezhurina et al., 2025; Lyu et al., 2024; Mondorf and Plank, 2024a), as highlighted in Section 2.1.

Ultimately, benchmark progress becomes misaligned with crucial practical goals such as usability, knowledge application, skill integration, and robustness (Pietruszka et al., 2024). Therefore, evaluation methodologies must evolve beyond convenient yet misleading proxies. While approaches like Chatbot Arena offer alternatives (Chiang et al., 2024), more robust solutions involve behavioral testing, adversarial evaluations, and the development of new benchmarks explicitly designed for validity, robustness, and real-world applicability (Raji et al., 2021; Pietruszka et al., 2024; Burden, 2024).

3 Mitigation of benchmarking issues

Research has proposed various solutions to alleviate the specific benchmarking issues surveyed in Section 2. In this section, we provide an overview of such suggestions keeping in mind all the issues we identified above. Table 1 provides an overview of the proposed solutions, and which issues they (do not) solve.

3.1 Pre-creation

Since some flaws in benchmarks stem from issues during their creation, suggestions have been made to improve relevant aspects before evaluating models on them. Specifically, research suggests to improve the quality and coverage of the data, and enrich the metadata.

For instance, **dynamically creating benchmarks** by continuously adding instances that are informed by model developments and model performances would (temporarily) alleviate the memorization issue: DynaBench (Kiela et al., 2021) and GEM (Gehrmann et al., 2021) are examples. However, the instances that are added in this process are prone to be cherry-picked based on specific failures of a model at that time, and might not be representative anymore of the task at hand (Bowman and Dahl, 2021).

Furthermore, existing benchmarks can be **augmented with refactored data**. Here, the focus is on consistency, by for instance including multiple formulations of the same instance to distinguish between genuine understanding and memorization. These instances can also be created by automatically generating perturbed versions of test instances (e.g., changing names, numbers, sentence order, logical structure slightly), where the perturbations are not related to the core task (e.g., the Alice in Wonderland problem in Nezhurina et al., 2025). However, it might be difficult to ensure that these perturbations only affect superficial features without changing the underlying task logic or the correct answer.

Additionally, benchmarks could be **filtered** to avoid easy, contaminated, and too similar examples (Gupta et al., 2025).

Besides adapting the input texts, it is argued that benchmarks should be released with more transparent and rich metadata. One aspect of this is the inclusion of **cultural bias annotations**, such as in the work of Singh et al. (2024b), where questions from MMLU were annotated based on whether cultural, geographical or dialect knowledge was needed to correctly answer the question. Another aspect is the **preservation of individual annotator responses** instead of collapsing them into a single aggregated label. This aligns with the perspectivism paradigm, which emphasizes the importance of considering diverse annotator perspectives in NLP tasks (Cabitza et al., 2023).

Finally, more **fine-grained or nuanced forms of annotation** are a possible approach as well. Sachdeva et al. (2022), for instance, use Rasch Measurement Theory (Rasch, 1960) to position social media messages on a hate speech spectrum, rather than providing an unnuanced binary label.

3.2 Post-creation

Benchmarking practices after the release of the benchmark can also be improved. A big factor is transparency and effectiveness of the evaluation metrics. On the one hand, it is suggested to **average the score** of a model across various benchmarks to ensure the generalizability (e.g., BIG-bench (Ghazal et al., 2013), and HuggingFace's Open LLM Leaderboard). On the other hand, there is more attention to evaluate models more broadly on a benchmark by including a **variety of evaluation metrics**, such as in HELM (Liang et al., 2023).

To facilitate open and reproducible evaluations, platforms such as OLMES (Gu et al., 2024) and Language Model Evaluation Harness (Gao et al., 2024) provide **open evaluation standards**.

Alternatively, there are arguments to keep the **test set of benchmarks secret**, and use private leaderboards to which the solutions are uploaded privately, and the final score is published (Rajore et al., 2024). While this would protect the test data from contamination, others argue that it would be better to **encrypt the test data** and release it together with the key to decrypt it, which would also protect it from crawlers (Jacovi et al., 2023). However, this is not a fool-proof system, and for instance exemplary instances that are provided in academic publications are still included in the pre-training data of LLMs (Gevers et al., 2025).

Chiang et al. (2024) argue that standardized NLP benchmarks fail to provide a diverse and nuanced evaluation of the expanding capabilities of LLMs, and therefore suggest to evaluate models using **human preferences**, proposing the Chatbot Arena. While, as can be seen from Table 1, this solution addresses most of the issues we identified in Section 2, this evaluation setup does not allow to measure a model's performance on a specific task, and leaves the door open for evaluation biases based on sycophancy and an overfitting to arena-specific dynamics over general model quality (Singh et al., 2025). Moreover, Singh et al. (2025) show that Chatbot Arena, which uses a normalized version of the Bradley-Terry model (Bradley and Terry, 1952), violates its assumption of unbiased sampling and full interconnection of the comparison network by providing preferential access to selected LLM providers and silently deprecating some models.

In addition, **mechanistic interpretability** (MI) can help investigate the internal mechanisms that could explain model behaviour on existing benchmarks (Bereska and Gavves, 2024; Lindsey et al., 2025). Findings from MI can validate whether a model possesses a claimed capability (construct validity) or merely mimics it. However, the methodology is hard to standardize and generalize across benchmarks.

4 Discussion

We see that many of the solutions provided in the literature are created in a vacuum, and address at best a selection of the problems we identified (see Table 1). Additionally, we note that there is more focus on some of the issues we describe than others. For example, few solutions tackle language imbalance or domain coverage.

We argue it is important to zoom out, and suggest to merge different proposed solutions so the effect is more robust against various pitfalls in benchmarking. Therefore, based on our literature review and some shortcomings it exposed, we establish two concrete checklists that could be used when (a) creating a benchmark; or (b) evaluating a method on an existing benchmark, which we present in Table 2. We demonstrate the applicability of our checklist by evaluating three widely used benchmarks (i.e., SuperGLUE, WinoGrande, and ARC-AGI) in Table 3. We note that a substantial number of our checklist items remain unmet across

Table 1: Effectiveness of proposed solutions against benchmark problems (✓: solves, ✗: doesn't solve, ∼: partially/temporarily solves), split into the following categories: Annotation Quality (AQ), Gameable (GA), Data Contamination (DaC), Language/Cultural Imbalance (LCI), Text Quality (TQ), Domain Coverage (DoC, Evaluation Inconsistency (EI), Representativeness (REP), .

Solution	AQ	GA	DaC	LCI	TQ	DoC	EI	REP
Pre-creation								
Dynamic benchmarks	∼	∼	✓	∼	∼	∼	✗	∼
Augment with refactored / perturbed data	✗	✓	∼	✗	∼	✗	✗	∼
Filtering benchmarks	✓	✓	✓	✗	✓	✗	✗	∼
Cultural bias annotations	✗	∼	✗	✓	✗	✗	✗	✗
Non-aggregated datasets	✓	✗	✗	∼	✗	✗	✗	✗
Fine-grained annotation scales	✓	✗	✗	✗	✗	✗	✗	✗
Post-creation								
Averaging scores	✗	✗	✗	∼	✗	∼	✗	∼
Multi-metrics	✗	∼	✗	✗	✗	✗	∼	✓
Open eval standards	✗	✗	✗	✗	✗	✗	✓	✗
Private leaderboards (secret test set)	✗	✗	✓	✗	✗	✗	✓	✗
Encrypt + license (CC BY-ND)	✗	✗	∼	✗	✗	✗	✗	✗
Human preference evaluation	✓	✓	✓	∼	✓	∼	✗	✗
Mechanistic interpretability	✗	✓	∼	✗	✗	✗	✗	∼

these benchmarks. For instance, in all three, there are no detailed annotation metadata, instance-level metadata, encryption or no-derivatives clauses for the test-set (although WinoGrande and ARC-AGI keep (part of) the test-set hidden), or allow for free-form inference. However, ARC-AGI meets more requirements than SuperGLUE and WinoGrande, since it is language-agnostic. Since our checklist is based on findings from previous literature, this highlights the weaknesses in current benchmarks that could be exploited by LLMs.

Alternatively, we must consider including less centralized and standardized strategies to evaluate LLM capabilities besides benchmarking, to ensure fair model evaluation and model generalizability. Specifically, we suggest to complement standardized benchmarks with a framework to concretely measure the model's downstream performance. Following the criterion validity, which posits that a good measure should also predict other concrete behavioral outcomes regarding the specific task/skill at hand, good performance on a benchmark should correlate with robust downstream performance (Bowman and Dahl, 2021). Therefore, as a future research direction, we suggest to create minimal-sized test-suites for real-life use cases to complement NLP benchmarking. We argue that model evaluations would be more robust by developing and using such test-suites, which should remain small enough to permit a rigorous qualitative evaluation. For example, in machine translation, small-sized test-suites including extreme edge cases are used to ensure broad, and unbiased applicability (e.g., Haddow et al., 2024). This could inspire NLP research to develop similar small datasets, in which the model is presented with the challenging cases that are relevant for real-life applications.[8] In opinion mining, for example, research could focus on Dutch COVID-19 vaccination skepticism (Lemmens et al., 2021), or on reputation analysis of governmental organizations (Boon et al., 2024). For future research, we propose to apply unsupervised sampling techniques to ensure the test-suite includes representative instances as well as informative outliers, for example by filtering for infrequent cases, gathering exemplar inputs from domain experts, and using recent case studies to ensure societal relevance and avoid data contamination. The addition of such framework to the usual model evaluation on standardized benchmarks would address all of the benchmarking issues mentioned earlier, and ensure the model performance is generalizable to real-world use cases.

[8]This differs from adversarial examples, which are designed to expose specific model weaknesses, and may not reflect genuine use-case demands (Bowman and Dahl, 2021).

Table 2: Checklist for constructing and evaluating benchmarks with the corresponding problems they solve: Annotation Quality, Gameable, Data Contamination, Domain Coverage, Evaluation Inconsistency, Language/Cultural Imbalance, Representativeness, Text Quality.

	Checklist for constructing benchmarks
☐	Provide a clear task definition with a taxonomy of intentions and assumptions of the required capabilities to solve the benchmark instances, rather than just the surface task type. Solves: **Representativeness**
☐	Clearly state the language, geographic, demographic, culture, or domain-specific limitations of the benchmark. Balance mix of domains and genres. Solves: **Language/Cultural Imbalance**, **Domain Coverage**, **Representativeness**
☐	Motivate the source of the annotations: crowdsourcing, expert annotators or synthetic. Provide annotations of all annotators (not only average), annotator guidebook and annotator metadata / demographics. Solves: **Annotation Quality**, **Language/Cultural Imbalance**
☐	Include detailed metadata, such as data sources (URLs, surrounding paragraphs), geographic and temporal information. Solves: **Data Contamination**, **Text Quality**, **Language/Cultural Imbalance**
☐	Perform extensive quality control of the texts. Pay attention to crowd-sourced texts, and (near-)duplicates. Solves: **Text Quality**, **Gameable**
☐	Include authentic data in high- and low-resource languages to guarantee cross-lingual performance. Alternatively, involve professional translators and account for cultural diversity (Barth and Rehm, 2025). Solves: **Language/Cultural Imbalance**
☐	Avoid using data that might have been memorized. For example, use tools like infini-gram[7] for web-scraped content. Solves: **Data Contamination**
☐	Add instances where surface features or irrelevant numerical details are systematically varied. Solves: **Gameable**
☐	Integrate evaluations in existing framework (e.g., OLMES, LM evaluation harness), or motivate the choice of evaluation metrics and open-source the evaluation (prompt, hyperparameters, evaluation script). Solves: **Gameable**, **Evaluation Inconsistency**
☐	Encrypt your benchmark and release the encrypted version with a no derivatives clause (Jacovi et al., 2023). Solves: **Data Contamination**
☐	Motivate the proposed inference method (e.g., probability, classification), but at least include free-form generation. Solves: **Evaluation Inconsistency**, **Representativeness**
☐	Provide relevant (open-sourced) baseline methods (which could reveal artifacts) and human performance. Solves: **Gameable**, **Representativeness**
	Checklist for evaluating methods on benchmarks
☐	Open-source the evaluation code. If available, include results using a standard prompt from the accompanying paper. Solves: **Gameable**, **Evaluation Inconsistency**
☐	Indicate if you trained your model on this benchmark and report scores without any training on the benchmark itself. Solves: **Data Contamination**
☐	Report the model version. Report score of at least one open-data LLM. Solves: **Evaluation Inconsistency**
☐	Use an appropriate interpretability technique to verify the information used for the task, such as SHAP (Mosca et al., 2022) or more recent mechanistic methods (Bereska and Gavves, 2024). Solves: **Gameable**
☐	Report a variety of evaluation metrics (cf. HELM). Solves: **Evaluation Inconsistency**, **Gameable**
☐	Report at least one statistical significance test between model, and baseline results and/or human performance (Reuel-Lamparth et al., 2024). Solves: **Evaluation Inconsistency**, **Gameable**
☐	Report whether the benchmark was used in the development phase. Solves: **Data Contamination**
☐	Release the raw model output (Laskar et al., 2024). Solves: **Gameable**, **Evaluation Inconsistency**

Table 3: Evaluation of three popular NLP benchmarks using our checklist for benchmark creation (✓: is applied, ✗: is not applied, ∼: irrelevant (e.g., not natural language)).

Benchmark	**SuperGLUE**	**WinoGrande**	**ARC-AGI**
Provide task definition	✓	✓	✓
State limitations, mix of domains and genres	✓	✓	✓
Motivate source of annotations, provide detailed annotation metadata	✗	✗	✗
Include metadata of texts	✗	✗	∼
Quality control of texts	✓	✓	✓
Authentic data in high- and low-resource languages, or professional translations	✗	✗	∼
Avoid memorized data	✗	✗	✓
Systematically adapt surface features	✓	✗	✓
Integrate evaluation in existing framework, or motivate and open-source evaluation metrics	✓	✓	✗
Encrypt and shared with no-derivatives clause	✗	✗	✗
Motivate inference method, at least include free-form	✗	✗	✗
Include open-source baselines and human performance	✓	✓	✓

5 Conclusion

Benchmarks are ubiquitous in the NLP community. It is the go-to method to evaluate model capabilities, and compare systems to each other. However, especially with the rise of powerful LLMs, weaknesses in benchmarking practices are revealed, questioning the validity of existing benchmarks in their creation, dissemination, and usage. However, as of now there is no one-fits-all solution to fix benchmarking.

In this study, we survey benchmarking issues that are identified in prior literature, grouped according to experimental validity types. Then, we survey proposed solutions in the literature for these issues. However, we find that it is important not to overestimate the usability of single solutions, since they are often created with only one or a few issues in mind, neglecting other pitfalls. Therefore, we combine specific recommendations from the literature in concrete checklists, which can be used to improve benchmarking practices. Last, we suggest to include downstream minimal-size test-suites to ensure the model's benchmark performance is generalizable to real-world use cases as a future research direction.

Limitations

This study is subject to a few limitations. First of all, this paper attempts to provide a comprehensive overview of discrete issues within NLP benchmarks and their proposed solutions, but it is inherently challenging to compile an exhaustive list. There are likely other issues present in NLP benchmarking, and potentially additional solutions suggested in the literature, that have not been captured within the scope of this work. However, the issues and solutions we include are representative of the overall problem we set out to address.

Second, to the best of our knowledge, there is little research focusing on the potential interactions between the different suggested solutions for benchmarking issues. For example, does addressing one issue inadvertently exacerbate others? These interdependencies should be further researched.

Third, our proposed checklist functions as a guideline for benchmarking practices. We do not claim this is a final product, and it should be updated with new insights from the community. Additionally, it might not be universally applicable across all benchmarking scenarios, so we encourage benchmark practitioners to adapt and tailor it to their specific contexts.

Another limitation of this study is that only the text modality was considered, even though benchmarks for other modalities, such as vision, are affected by similar issues, as reported in Li et al. (2025). Nonetheless, the issues raised and checklists provided in this study are still relevant to non-textual benchmarks.

Acknowledgments

This research was made possible with a grant from the Fonds Wetenschappelijk Onderzoek (FWO) project 11P3824N, and funding from the Flemish

Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, Bangkok, Thailand. Association for Computational Linguistics.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. Artifacts or abduction: How do LLMs answer multiple-choice questions without the question? *Preprint*, arXiv:2402.12483.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Sourav Banerjee, Ayushi Agarwal, and Eishkaran Singh. 2024. The vulnerability of language model benchmarks: Do they accurately reflect true LLM performance? *arXiv preprint arXiv:2412.03597*.

Fabio Barth and Georg Rehm. 2025. Multilingual European language models: Benchmarking approaches and challenges. *Preprint*, arXiv:2502.12895.

Edward Bayes, Israel Abebe Azime, Jesujoba Oluwadara Alabi, Jonas Kgomo, Tyna Eloundou, Elizabeth Proehl, Kai Chen, Imaan Khadir, Naome A. Etori, Shamsuddeen Hassan Muhammad, Choice Mpanza, Igneciah Pocia Thete, Dietrich Klakow, and David Ifeoluwa Adelani. 2024. Uhura: A benchmark for evaluating scientific question answering and truthfulness in low-resource African languages. *ArXiv*, abs/2412.00948.

Andrew M. Bean, Simi Hellsten, Harry Mayne, Jabez Magomere, Ethan A. Chi, Ryan Chi, Scott A. Hale, and Hannah Rose Kirk. 2024. Lingoly: A benchmark of olympiad-level linguistic reasoning puzzles in low-resource and extinct languages. *ArXiv*, abs/2406.06196.

Leonard Bereska and Stratis Gavves. 2024. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*. Survey Certification, Expert Certification.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, and 1 others. 2024. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*.

Jan Boon, Jan Wynen, Walter Daelemans, Jens Lemmens, and Koen Verhoest. 2024. Agencies on the parliamentary radar: Exploring the relations between media attention and parliamentary attention for public agencies using machine learning methods. *Public Administration*, 102(3):1026–1044.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in Natural Language Understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

John Burden. 2024. Evaluating AI Evaluation: Perils and Prospects. *arXiv preprint*. ArXiv:2407.09221 [cs].

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Yuxing Cheng, Yi Chang, and Yuan Wu. 2025. A Survey on Data Contamination for Large Language Models. *arXiv preprint*. ArXiv:2502.14425 [cs].

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Gyeongje Cho, Yeonkyoung So, and Jaejin Lee. 2025. ANPMI: Assessing the true comprehension capabilities of LLMs for multiple choice questions. *Preprint*, arXiv:2502.18798.

Francois Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. 2024. Arc prize 2024: Technical report. *arXiv preprint arXiv:2412.04604*.

Kenneth Ward Church. 2020. Benchmarks and goals. *Natural Language Engineering*, 26(5):579–592.

Luna De Bruyne. 2023. The paradox of multilingual emotion detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 458–466, Toronto, Canada. Association for Computational Linguistics.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, and 1 others. 2025. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.

Leon Engländer, Hannah Sterz, Clifton Poth, Jonas Pfeiffer, Ilia Kuznetsov, and Iryna Gurevych. 2024. M2qa: Multi-domain multilingual question answering. *arXiv preprint arXiv:2407.01091*.

Virginia Felkner, Jennifer Thompson, and Jonathan May. 2024. GPT is not an annotator: The necessity of human annotation in fairness benchmark construction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14104–14115, Bangkok, Thailand. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1813, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, and 37 others. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Ine Gevers, Victor De Marez, Luna De Bruyne, and Walter Daelemans. 2025. WinoWhat: A parallel corpus of paraphrased WinoGrande sentences with common sense categorization. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 68–80, Vienna, Austria. Association for Computational Linguistics.

Ahmad Ghazal, Tilmann Rabl, Minqing Hu, Francois Raab, Meikel Poess, Alain Crolotte, and Hans-Arno Jacobsen. 2013. Bigbench: Towards an industry standard benchmark for big data analytics. In *Proceedings of the 2013 ACM SIGMOD international conference on Management of data*, pages 1197–1208.

Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. 2020. A report on the 2020 sarcasm detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 1–11, Online. Association for Computational Linguistics.

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. 2024. OLMES: A standard for language model evaluations. *Preprint*, arXiv:2406.08446.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

Vipul Gupta, Candace Ross, David Pantoja, Rebecca J. Passonneau, Megan Ung, and Adina Williams. 2025. Improving model evaluation using SMART filtering of benchmark datasets. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4595–4615, Albuquerque, New Mexico. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in Natural Language Inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors. 2024. *Proceedings of the Ninth Conference on Machine Translation*. Association for Computational Linguistics, Miami, Florida, USA.

Jennifer Hsia, Danish Pruthi, Aarti Singh, and Zachary Lipton. 2024. Goodhart's law applies to NLP's explanation benchmarks. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1322–1335, St. Julian's, Malta. Association for Computational Linguistics.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop Uploading Test Data in Plain Text: Practical Strategies for Mitigating Data Contamination by Evaluation Benchmarks. *arXiv preprint*. ArXiv:2305.10160 [cs].

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Bernard J Koch and David Peterson. 2024. From protoscience to epistemic monoculture: How benchmarking set the stage for the deep learning revolution. *arXiv preprint arXiv:2404.06647*.

Thanos Konstantinidis, Giorgos Iacovides, Mingxue Xu, Tony G Constantinides, and Danilo Mandic. 2024. Finllama: Financial sentiment classification for algorithmic trading applications. *arXiv preprint arXiv:2403.12285*.

Andriy Kosar, Guy De Pauw, and Walter Daelemans. 2024. Comparative evaluation of topic detection: Humans vs. LLMs. *Computational Linguistics in the Netherlands Journal*, 13:91–120.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuật Nguyễn, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13785–13816, Miami, Florida, USA. Association for Computational Linguistics.

Jens Lemmens, Tess Dejaeghere, Tim Kreutz, Jens Van Nooten, Ilia Markov, and Walter Daelemans. 2021. Vaccinpraat: Monitoring vaccine skepticism in Dutch Twitter and Facebook comments. *Computational Linguistics in the Netherlands Journal*, 11:173–188.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *arXiv preprint arXiv:2409.20288*.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *Preprint*, arXiv:2501.02189.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Preprint*, arXiv:2211.09110.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, and 8 others. 2025. On the biology of a large language model. *Transformer Circuits Thread*.

Tyler Loakman, Aaron Maladry, and Chenghua Lin. 2023. The iron(ic) melting pot: Reviewing human evaluation in humour, irony and sarcasm generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6676–6689, Singapore. Association for Computational Linguistics.

Rachel Longjohn, Markelle Kelly, Sameer Singh, and Padhraic Smyth. 2024. Benchmark data repositories for better benchmarking. *ArXiv*, abs/2410.24100.

Chenyang Lyu, Minghao Wu, and Alham Aji. 2024. Beyond Probabilities: Unveiling the Misalignment in Evaluating Large Language Models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 109–131, Bangkok, Thailand. Association for Computational Linguistics.

Gaurav Maheshwari, Dmitry Ivanov, and Kevin El Haddad. 2024. Efficacy of synthetic data as a benchmark. *ArXiv*, abs/2409.11968.

Hong-Ren Mao and Hung-Yi Lee. 2019. Polly want a cracker: Analyzing performance of parroting on paraphrase generation datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5960–5968, Hong Kong, China. Association for Computational Linguistics.

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, and 4 others. 2025. Artificial intelligence index report 2025. *Preprint*, arXiv:2504.07139.

Timothy R. McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Paul Watters, and Malka N. Halgamuge. 2024. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *Preprint*, arXiv:2402.09880.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Philipp Mondorf and Barbara Plank. 2024a. Beyond Accuracy: Evaluating the Reasoning Behavior of Large Language Models – A Survey. *arXiv preprint*. ArXiv:2404.01869 [cs].

Philipp Mondorf and Barbara Plank. 2024b. Beyond accuracy: Evaluating the reasoning behavior of large language models - a survey. In *First Conference on Language Modeling*.

Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. SHAP-based explanation methods: a review for NLP interpretability. In *Proceedings of the 29th international conference on computational linguistics*, pages 4593–4603.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. AraDiCE: Benchmarks for Dialectal and Cultural Capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218.

Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. 2024. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. *arXiv preprint arXiv:2406.07545*.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, and 1 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. 2025. Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models. *arXiv preprint*. ArXiv:2406.02061 [cs].

Ankit Pal and Malaikannan Sankarasubbu. 2024. Gemini goes to med school: exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 21–46.

Medha Palavalli, Amanda Bertsch, and Matthew R. Gormley. 2024. A Taxonomy for Data Contamination in Large Language Models. *arXiv preprint*. ArXiv:2407.08716 [cs].

Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2023. Don't blame the annotator: Bias already starts in the annotation instructions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1779–1789, Dubrovnik, Croatia. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, and 716 others. 2025. Humanity's last exam. *Preprint*, arXiv:2501.14249.

Michał Pietruszka, Łukasz Borchmann, Aleksander Jędrosz, and Paweł Morawiecki. 2024. Can Models Help Us Create Better Models? Evaluating LLMs as Data Scientists. *arXiv preprint*. ArXiv:2410.23331 [cs].

Irene Plaza, Nina Melero, Cristina del Pozo, Javier Conde, Pedro Reviriego, Marina Mayor-Rocher, and María Grandury. 2024. Spanish and llm benchmarks: is mmlu lost in translation? *arXiv preprint arXiv:2406.17789*.

Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Tanmay Rajore, Nishanth Chandran, Sunayana Sitaram, Divya Gupta, Rahul Sharma, Kashish Mittal, and Manohar Swaminathan. 2024. Truce: Private benchmarking to prevent contamination and improve comparative evaluation of llms. *arXiv preprint arXiv:2403.00393*.

Georg Rasch. 1960. Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.

Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, Ruochen Zhao, Chengwei Qin, Caiming Xiong, and Shafiq Joty. 2024. How Much are LLMs Contaminated? A Comprehensive Survey and the LLMSanitize Library. *arXiv preprint*. ArXiv:2404.00699 [cs] version: 1.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *ArXiv*, abs/2311.12022.

Anka Reuel-Lamparth, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J Kochenderfer. 2024. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. *Advances in Neural Information Processing Systems*, 37:21763–21813.

Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. Just What do You Think You're Doing, Dave?' a Checklist for Responsible Data Use in NLP. *ArXiv*, abs/2109.06598.

Sebastian Ruder. 2021. Challenges and Opportunities in NLP Benchmarking. `http://ruder.io/nlp-benchmarking`.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, and Eneko Agirre. 2023a. Did ChatGPT cheat on your test? `https://hitz-zentroa.github.io/lm-contamination/blog/`. Accessed: 2025-04-16.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023b. NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. *arXiv preprint*. ArXiv:2310.18018 [cs].

Oscar Sainz, Iker García-Ferrero, Alon Jacovi, Jon Ander Campos, Yanai Elazar, Eneko Agirre, Yoav Goldberg, Wei-Lin Chen, Jenny Chim, Leshem Choshen, Luca D'Amico-Wong, Melissa Dell, Run-Ze Fan, Shahriar Golchin, Yucheng Li, Pengfei Liu, Bhavish Pahwa, Ameya Prabhu, Suryansh Sharma, and 9 others. 2024. Data Contamination Report from the 2024 CONDA Shared Task.

Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. 2024. LawLLM: Law large language model for the US legal system. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4882–4889.

Aaditya K. Singh, Muhammed Yusuf Kocyigit, Andrew Poulton, David Esiobu, Maria Lomeli, Gergely Szilvasy, and Dieuwke Hupkes. 2024a. Evaluation data contamination in LLMs: how do we measure it and (when) does it matter? *arXiv preprint*. ArXiv:2411.03923 [cs].

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah Smith, and 1 others. 2025. The Leaderboard Illusion. *arXiv preprint arXiv:2504.20879*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, and 1 others. 2024b. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.

Zechen Sun, Yisheng Xiao, Juntao Li, Yixin Ji, Wenliang Chen, and Min Zhang. 2024. Exploring and mitigating shortcut learning for generative large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6883–6893, Torino, Italia. ELRA and ICCL.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and 1 others. 2024. Towards Multilingual LLM Evaluation for European Languages. *arXiv preprint arXiv:2410.08928*.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. In *Neural Information Processing Systems*.

Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. 2024. Large language model benchmarks do not test reliability. In *Neurips Safe Generative AI Workshop 2024*.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384.

Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, Anders Wesslén, and 1 others. 2012. *Experimentation in software engineering*, volume 236. Springer.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, and 1 others. 2024. Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743.

Cheng Xu, Shuhao Guan, Derek Greene, and M.-Tahar Kechadi. 2024. Benchmark Data Contamination of Large Language Models: A Survey. *arXiv preprint*. ArXiv:2406.04244 [cs].

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. 2024. Do LLMs overcome shortcut learning? an evaluation of shortcut challenges in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12188–12200.

Huixuan Zhang, Yun Lin, and Xiaojun Wan. 2024. PaCoST: Paired confidence significance testing for benchmark contamination detection in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1794–1809, Miami, Florida, USA. Association for Computational Linguistics.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't Make Your LLM an Evaluation Benchmark Cheater. *arXiv preprint*. ArXiv:2311.01964 [cs].