When AI Benchmarks Plateau: A Systematic Study of Benchmark Saturation

Mubashara Akhtar [*1,2] **Anka Reuel** [*3]

Prajna Soni [◇] **Sanchit Ahuja** [◇4] **Pawan Sasanka Ammanamanchi** [◇5] **Ruchit Rawal** [◇6] **Vilém Zouhar** [◇1] **Srishti Yadav** [◇7] **Chenxi Whitehouse** [◇8] **Dayeon Ki** [◇6]

Jennifer Mickel [9] **Leshem Choshen** [10] **Marek Šuppa** [11] **Jan Batzner** [12] **Jenny Chim** [13] **Jeba Sania** [14] **Yanan Long** [15] **Hossein A. Rahmani** [16] **Christina Knight** [17] **Yiyang Nan** [18] **Jyoutir Raj** [5] **Yu Fan** [1,19] **Shubham Singh** [20] **Subramanyam Sahoo** [21] **Eliya Habba** [22] **Usman Gohar** [23] **Siddhesh Pawar** [7] **Robert Scholz** [24] **Arjun Subramonian** [5] **Jingwei Ni** [1]

Mykel J. Kochenderfer [†3] **Sanmi Koyejo** [†3] **Mrinmaya Sachan** [†1,2] **Stella Biderman** [†9] **Zeerak Talat** [†25] **Avijit Ghosh** [†26] **Irene Solaiman** [†26]

* Lead authors ◇ Top contributors † Advisors

This project was completed as part of the Evaluating Evaluations (EvalEval) Coalition: https://evalevalai.com/

Abstract

Artificial intelligence benchmarks are an important mechanism for measuring model progress and guiding deployment decisions. However, benchmarks quickly "saturate", making it difficult to differentiate models and diminishing their long-term value. In this study, we define *benchmark saturation* and analyze it across 60 language model benchmarks using 14 properties that relate to saturation. We find that nearly half of the our benchmarks exhibit saturation, with rates increasing with age. Further, we find that resilience to saturation is impacted by expert-curation, not by public test data. Our results suggest that design choices can extend benchmark longevity and inform more durable evaluation approaches.[1]

[1]ETH Zurich [2]ETH AI Center [3]Stanford University [4]Northeastern University [5]Independent Researcher [6]University of Maryland [7]University of Copenhagen [8]University of Cambridge [9]Eleuther AI [10]IBM Research, MIT-IBM Watson AI lab, MIT [11]Comenius University in Bratislava / Cisco [12]Weizenbaum Institute, Munich Center for Machine Learning, TUM [13]Queen Mary University of London [14]Harvard University [15]StickFlux Labs [16]AI Center, University College London [17]Scale AI Security and Policy Research Lab [18]Cohere [19]University of Hong Kong [20]University of Illinois Chicago [21]Berkeley AI Safety Initiative [22]The Hebrew University of Jerusalem [23]Iowa State University [24]Max Planck School of Cognition [25]University of Edinburgh [26]Hugging Face. Correspondence to: Mubashara Akhtar <mubashara.akhtar@ai.ethz.ch>, Anka Reuel <anka@cs.stanford.edu>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

[1]Data and code are available in the Github repository.

1. Introduction

Artificial Intelligence (AI) benchmarks play a central role in measuring model progress, guiding deployment decisions, and informing policy and regulation (Hardy et al., 2025; 2024; Alzahrani et al., 2024; Union, 2024). Their value depends on their ability to distinguish between models. Yet many widely used benchmarks (*e.g.,* HumanEval (Chen, 2021)) have rapidly "saturated" (Maslej et al., 2024), with top-performing systems achieving near-identical scores. When performance converges within a narrow range, benchmarks lose discriminative power and provide limited guidance for model comparison or selection (Ott et al., 2022; Chen et al., 2025). Similar dynamics have been observed in other domains–for example, ImageNet (Deng et al., 2009) exhibits near-ceiling performance for most new models.[2]

Despite its importance, *benchmark saturation* has received limited systematic study. Prior work often notes performance plateaus, increased robustness (Ashury-Tahan et al., 2026a) or introduces new benchmarks in response (Wang et al., 2024b; Jimenez et al., 2024), but rarely analyzes the mechanisms driving saturation. It remains unclear why some benchmarks saturate quickly while others retain discriminative power, and there is no agreed-upon operational definition–whether saturation reflects near-human performance, fixed ceilings, or the loss of statistical separability among state-of-the-art models. We address these gaps by defining saturation as the loss of reliable discriminative power among top-performing models and operationalizing

[2]We want to emphasize that saturation of benchmarks is not always negative–if the benchmark was valid (Salaudeen et al., 2025), saturation means that a task can be considered "solved".

it through an uncertainty-aware saturation index derived from leaderboard data. Using this framework, we analyze 60 widely used text-based LLM benchmarks across domains and evaluation settings, annotated along dimensions such as task design, linguistic scope, data construction, and accessibility to study factors associated with saturation.

This paper makes the following contributions:

- We define benchmark saturation as the loss of reliable discriminative power among state-of-the-art models and introduce a reproducible, uncertainty-aware saturation index derived from leaderboard data.
- We identify which benchmark properties are systematically associated with saturation based on an analysis of 60 benchmarks. We find that commonly assumed safeguards, such as private test sets or closed-ended formats, have limited impact on saturation, while benchmark age and scale strongly predict it.
- We derive practical recommendations for benchmark design and lifecycle management, including monitoring practices, uncertainty reporting, and criteria for benchmark retirement or revision.

The remaining paper is organized as follows: Section 2 formalizes benchmark saturation and introduces our saturation index; Section 3 outlines benchmark collection and annotation; Section 4 presents the empirical analyses; Section 5 discusses implications and actionable recommendations; Section 6 concludes with limitations and future directions.

Conflict of Interest Disclosure This work was conducted as part of a research coalition, some of whose members (including coauthors) have contributed to models or reported evaluations analyzed in this study. All artifacts were subject to the same inclusion and annotation procedure, regardless of author involvement.

2. Conceptualizing Benchmark Saturation

In this section, we formally define benchmark saturation, introduce our uncertainty-aware saturation index, and analyze its robustness to key parameter choices.

2.1. Definition and Scope

We define *benchmark saturation* as the loss of reliable discriminative power among top-performing models under comparison. A benchmark is saturated when top-performing models cannot be statistically distinguished and performance approaches the empirically observed ceiling of the benchmark. This notion corresponds to what prior work informally describes as *performance saturation*–a plateau where inter-model differences become negligible (Justen, 2025; Wang et al., 2024b; Ott et al., 2022).

Human performance ceiling. Unlike definitions based on reaching human-level performance (Gupta et al., 2025), our definition does not rely on human baselines, which are often impossible to comprehensively obtain, unavailable, or inconsistently measured (Wei et al., 2025). Moreover, human-level performance does not imply saturation, as models may still be statistically distinguishable even after reaching human-level scores, allowing the benchmark to retain discriminative power. Previous analyses describe saturation patterns descriptively (Ott et al., 2022) or emphasize lifecycle management (Hardy et al., 2024), but do not provide a quantitative criterion to determine saturation.

Saturation vs. stagnation. We therefore formalize saturation as a measurable property derived from leaderboard uncertainty. We further distinguish *stagnation* from saturation: stagnation refers to statistical indistinguishability among top models, whereas saturation additionally requires that performance is near the empirical ceiling. In practice, limited noise estimates blur the distinction between the two.

Definition: Benchmark Saturation

A benchmark is *saturated* if the evaluated models can not be reliably distinguished by their performance scores and any further improvements are not statistically distinguishable under the evaluation protocol.
Formally, saturation is characterized by:
(1) statistically alike performance among different top-performing models
(2) top performing models are approaching the benchmark's empirically inferred ceiling.

If only condition (1) holds, we refer to the benchmark as being *stagnated* rather than saturated. In this case, observed indistinguishability may arise from model-level limitations, evaluation noise, insufficient benchmark sensitivity, or artifacts in the benchmark itself (e.g., spurious correlations or repetitive patterns) and may be overcome by future architectural, training, or evaluation advances. It is often difficult to clearly distinguish stagnation from saturation, as reliable estimates of evaluation noise and benchmark ceilings are rarely available.

Our operationalization should satisfy four desiderata:

1. **Model-relative:** Defined with respect to top-performing models at a given time.
2. **Metric-agnostic:** Applicable across common metrics (accuracy, F1, BLEU).
3. **Data-driven:** Avoids reliance on externally curated performance ceilings.
4. **Reproducible:** Produces identical decisions given the same leaderboard snapshot.

To formalize this notion, we consider the performance of a set of top-performing models on each benchmark. For

a given benchmark, let $s_1 \geq \cdots \geq s_k$ denote the scores of the top k models (default $k = 5$). We introduce k as a general parameter to avoid fixing the number of models considered, and to flexibly define the set of top-performing models used to assess saturation. In our analysis, we fix $k = 5$ to ensure comparability across benchmarks. This choice reflects a practical trade-off: smaller values of k can lead to unstable estimates, while larger values risk mixing frontier models with older or less relevant ones, particularly given incomplete leaderboard coverage. Empirically, most benchmarks in our dataset report on approximately 5–7 recent, highly capable models, making $k = 5$ a reasonable and consistent choice.

2.2. Uncertainty-Aware Saturation Measurement

Performance-based evaluation. For accuracy-like metrics that are averages over a fixed test set of size n, we approximate the standard error of a model score s as

$$\mathrm{SE}(s) \approx \sqrt{\frac{s(1-s)}{n_{\text{eff}}}}. \tag{1}$$

$$\text{where } n_{\text{eff}} = n^{\alpha}, \qquad \alpha \in [0,1], \text{ default } \alpha = 0.5,$$

Note that accuracy-like metrics, metrics computed as averages over a fixed set of test samples with bounded per-sample contributions (e.g., accuracy, F1, BLEU) are broadly used. In such metrics, uncertainty can be approximated from finite-sample variability. For other metric types (e.g., Pass@k), the same framework is applicable but requires a benchmark-specific uncertainty estimate, such as bootstrap intervals or repeated-evaluation variance.

The effective test set size $n_{\text{eff}} = n^{\alpha}$ down weights the nominal test set size n to avoid an overly strong dependence of the saturation calculation on test set size. In our dataset, benchmark sizes vary substantially, ranging from a few dozen to several hundred thousand test samples, with a highly skewed distribution due to a small number of very large benchmarks. Using the raw test set size n would therefore cause the uncertainty term to be dominated by these outliers, leading to disproportionately small standard errors and artificially low saturation estimates for large benchmarks.

Thus, the standard error of the difference between the top model and k-th model is then

$$\mathrm{SE}_{\Delta} \approx \sqrt{\frac{s_1(1-s_1)}{n_{\text{eff}}} + \frac{s_k(1-s_k)}{n_{\text{eff}}}}. \tag{2}$$

Let $\Delta = s_1 - s_k$. We consider the top models to be statistically similar in performance if $\Delta \leq z \cdot \mathrm{SE}_{\Delta}$, where z is a standard normal quantile (e.g., $z = 1.96$ for a 95% confidence level). This criterion considers both dataset size and evaluation noise. Evaluation uncertainty refers to the expected variability in leaderboard scores introduced by finite test set size and metric estimation noise. We define this uncertainty through the standard error of model scores and their differences, and treat performance differences within this range as statistically indistinguishable.

Score compression. To quantify to which degree performance scores at the top of the leaderboard are collapsing, we compute the normalized score range

$$R_{\text{norm}} = \frac{s_1 - s_k}{\mathrm{SE}_{\Delta}}. \tag{3}$$

R_{norm} can be interpreted as a signal-to-noise ratio, comparing observed top-model score spread to expected evaluation uncertainty. Lower R_{norm} indicates greater saturation, with top-model differences falling within expected evaluation uncertainty and showing limited discrimination.[3]

Empirical approximation of the noise ceiling. Rather than assuming a fixed or externally defined noise ceiling, we treat the highest observed model performance (s_1) as an empirical proxy for the ceiling. Saturation is therefore assessed relative to the distribution of observed model scores, rather than with respect to an absolute performance target such as perfect accuracy (i.e., accuracy of 100%).

Strong clustering of top models at a low performance level should not be interpreted as the task being solved. Instead, such clustering indicates *model-level saturation*: the benchmark may no longer effectively distinguish between contemporary state-of-the-art models. However, as observed in prior benchmarks, this form of saturation reflects stagnation and does not preclude the benchmark from regaining discriminative power following paradigm shifts (*e.g.,* introduction of reasoning-centric or tool-augmented models) (Cobbe et al., 2021; Lewkowycz et al., 2022).

Saturation index. To capture saturation as a graded phenomenon, we combine the above signals into a continuous saturation index $S_{\text{index}} \in [0,1]$, which increases as top models become statistically indistinguishable. Benchmarks with higher values of S_{index} show stronger saturation evidence. We define the saturation index as

$$S_{\text{index}} = \exp(-R_{\text{norm}}^2), \tag{4}$$

which assigns high values when the performance differences are small relative to the evaluation uncertainty. High values of S_{index} indicate benchmarks where top-performing models are tightly clustered within evaluation noise, reflecting reduced discriminative power.

[3] In rare cases with near-zero uncertainty (e.g., deterministic near-perfect scores), we add a small ϵ-stabilization in the denominator to avoid numerical instability.

Setting comparison	Spearman correlation	Same bin (%)
$(k = 3)$ vs $(k = 5)$	0.92	48.3
$(\alpha = 0.5)$ vs $(\alpha = 0)$	0.88	23.3
$(\alpha = 0.5)$ vs $(\alpha = 1)$	0.92	18.3

Table 1. Sensitivity analysis of saturation index with respect to k and α. We report Spearman rank correlation and the percentage of benchmarks assigned to the same saturation bin.

For interpretability, we bucket benchmarks into five bins: *very low* (< 0.01), *low* ($[0.01, 0.3)$), *moderate* ($[0.3, 0.7)$), *high* ($[0.7, 0.9)$), and *very high* saturation (≥ 0.9). Notably, high saturation may also occur at lower absolute performance levels, reflecting model-level saturation rather than task-level completion. These bins are interpretable, empirically motivated ranges over a continuous score, intended to summarize broad saturation regimes rather than define strict thresholds. They reflect the spread of S_{index} observed across benchmarks while preserving the index's continuity.

2.3. Sensitivity to Parameter Selection

We conduct a sensitivity analysis with varying $k \in 3, 5$ and $\alpha \in 0, 0.5, 1$ values. Across these settings, the resulting saturation indices remain highly correlated, indicating that the relative ranking of benchmarks is preserved. Table 1 gives an overview of correlation and the fraction of benchmarks that remain in the same bins. While we observe variation in bin assignments, most changes occur between neighbouring bins rather than large shifts, which suggests that the underlying signal is stable even when scores vary. We further observe that a smaller k-value (e.g., $k = 3$) increases variance due to limited model coverage, while larger k risks mixing frontier and non-frontier models given incomplete, static leaderboard data. Similarly, $\alpha = 1$ leads to strong dependence on test set size, whereas $\alpha = 0$ ignores evaluation uncertainty. The choice $\alpha = 0.5$ is a balanced trade-off, moderating dataset size effects while preserving uncertainty awareness. Overall, absolute saturation values may shift slightly, but benchmark ordering remains stable.

3. Methodology

To study benchmark saturation, we combine structured benchmark annotations with leaderboard-based analysis.

3.1. Benchmark Collection and Annotation

(1) Initial benchmark selection. We used a three-stage, criteria-driven process to construct a representative benchmark set, focusing on benchmarks that (i) are actively used in contemporary LLM evaluation, (ii) provide sufficient longitudinal data, and (iii) vary along dimensions relevant to our hypotheses.

We compiled candidate benchmarks from two sources: *1. Evaluation reports from major model developers.* We extracted benchmarks appearing in evaluation sections of official reports (such as model cards or technical reports) released between Jan 2022 and Nov 2025 by major developers, including OpenAI, Anthropic, Google, Meta, and Alibaba, to reflect real-world evaluation practices and downstream adoption. In total, we reviewed 61 documents and identified 190 benchmarks used in at least one report. *2. Highly-cited benchmark papers.* We additionally collected widely cited benchmarks via the Semantic Scholar API using keyword-based search (details in Appendix C).

(2) Criteria-based filtering. We filtered benchmarks to ensure suitability for analysis using the following criteria: *1. Public documentation:* Benchmark documentation (*e.g.,* paper, technical report, or website) must be publicly available. *2. Sustained usage:* Benchmarks extracted from developer reports must appear in at least five distinct reports to ensure broader relevance. *3. Clear evaluation protocol:* Benchmarks with ambiguous scoring, inconsistent splits, or unclear evaluation procedures were excluded. *4. Text-only scope:* We restricted our analysis to text-based benchmarks, excluding multimodal datasets to isolate language-related saturation effects. *5. Available leaderboard data:* We included only benchmarks with sufficiently up-to-date leaderboard data and multiple evaluated models, otherwise they were excluded (*e.g.,* BIG-Bench (Srivastava et al., 2023)).

(3) Hypothesis-driven refinement. We initially develop a set of hypotheses for potential causes of benchmark saturation (see Section 3 and Section B). To ensure adequate sample sizes across hypotheses, we supplemented the filtered set with benchmarks filling gaps along key dimensions (*e.g.,* multilingual, templated, open-ended). We conducted targeted Google Scholar searches using terms such as `AI benchmark`, `leaderboard`, `evaluation`, and `dataset`, combined with hypothesis-specific keywords (*e.g.,* `multilingual`, `open-ended generation`).

After filtering and refinement, the final dataset consists of **60 benchmarks**. See Table 3 for the full list.

(4) Annotation protocol. To test the hypotheses in Section 3, we annotated benchmarks according to the schema in Table 4. Annotations capture: (i) *temporality* (*e.g.,* release date), (ii) *saturation metrics* (*e.g.,* top-5 model scores), (iii) *data quality indicators*, (iv) *task structure* (*e.g.,* input/output format), and (v) *dataset properties* (*e.g.,* curation strategy). Annotations were collected through a structured protocol involving 23 researchers with expertise in dataset curation and evaluation. Each benchmark was independently annotated and secondarily reviewed using a predefined schema, followed by a final cross-benchmark consistency audit to resolve remaining ambiguities.

Hypothesis	Statement
H1	Public benchmarks saturate faster than private benchmarks with held-out test sets.
H2	English-only benchmarks saturate faster than multilingual or mixed-language benchmarks.
H3	Human-authored benchmarks are more resistant to performance saturation than synthetic or hybrid ones.
H4	Benchmarks that use a closed-ended response format (e.g., multiple-choice, true/false) tend to saturate faster than those requiring open-ended generation.
H5	Benchmarks that are older and more widely adopted saturate faster than newer or less-used benchmarks.
H6	Non-templated benchmarks are more resistant to performance saturation than templated benchmarks.

Table 2. Hypotheses on factors driving benchmark saturation.

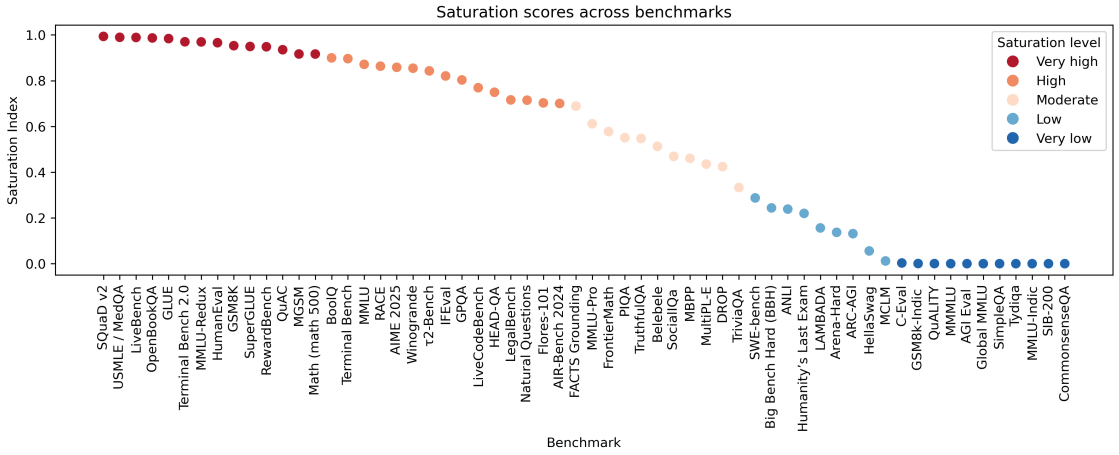

Figure 1. Overview of saturation scores across all studied benchmarks, ranked by their saturation levels.

Final benchmark set. Our benchmark selection spans a broad range of evaluation settings, including knowledge and reasoning tasks, multilingual, coding, long-context and factuality benchmarks, and recent agentic tasks. The benchmarks vary substantially in age (between 1 and 114 months), scale (from a few to hundreds of thousands of test samples), accessibility, output format, and construction style. Overall, the set includes 52 public and 8 private benchmarks, 44 English-only and 16 multilingual benchmarks, 28 closed-ended and 31 open-ended benchmarks, and 14 templated versus 46 non-templated benchmarks (Figure 1).

4. Empirical Analysis of Benchmark Saturation

We analyze saturation patterns across 60 text-based LLM benchmarks spanning domains, task formats, and evaluation settings. Using our saturation index (Section 2), we examine (i) saturation prevalence, (ii) temporal and exposure effects, and (iii) differences across benchmark properties.

4.1. Hypotheses-specific Analysis

We evaluate five hypotheses regarding potential drivers of saturation, grouped by accessibility (H_1), linguistic scope (H_2), data construction and quality (H_3), task design (H_4), popularity (H_5), and template (H_6); see App. B for details. Since benchmark age is itself positively associated with saturation and differs across several benchmark categories, age is an important cofounding factor in cross-benchmark comparisons. We therefore distinguish age-balanced comparisons (H_1, H_5, H_6), where groups have similar maturity, from age-confounded comparisons (H_2–H_4) (Figure 2).

Overall saturation patterns. Saturation is widespread. Of the 60 benchmarks analyzed, 29 exhibit high or very high saturation ($S_{\text{index}} \geq 0.7$), out of which 14 fall into the very high category ($S_{\text{index}} \geq 0.9$). These benchmarks show strong compression among top-performing models, indicating limited discriminative power at the frontier. Across benchmarks, larger test sets are associated with lower saturation indices. Benchmarks with more test items show less score compression among top models, consistent with lower evaluation uncertainty and higher resolution. This relationship persists in joint regression (Section 4.2), suggesting that measurement scale impacts discriminative power.

Temporal and exposure effects. Figure 3 shows that the average saturation index increases with benchmark age. The proportion of saturated benchmarks rises from 42.9% for benchmarks released within the past 24 months to 54.5% for those older than 60 months, with corresponding mean S_{index} values of 0.51, 0.52, and 0.60 across age bins. While the trend is modest and not statistically significant at conventional thresholds, it is directionally consistent: older benchmarks exhibit greater top-score compression. We eval-

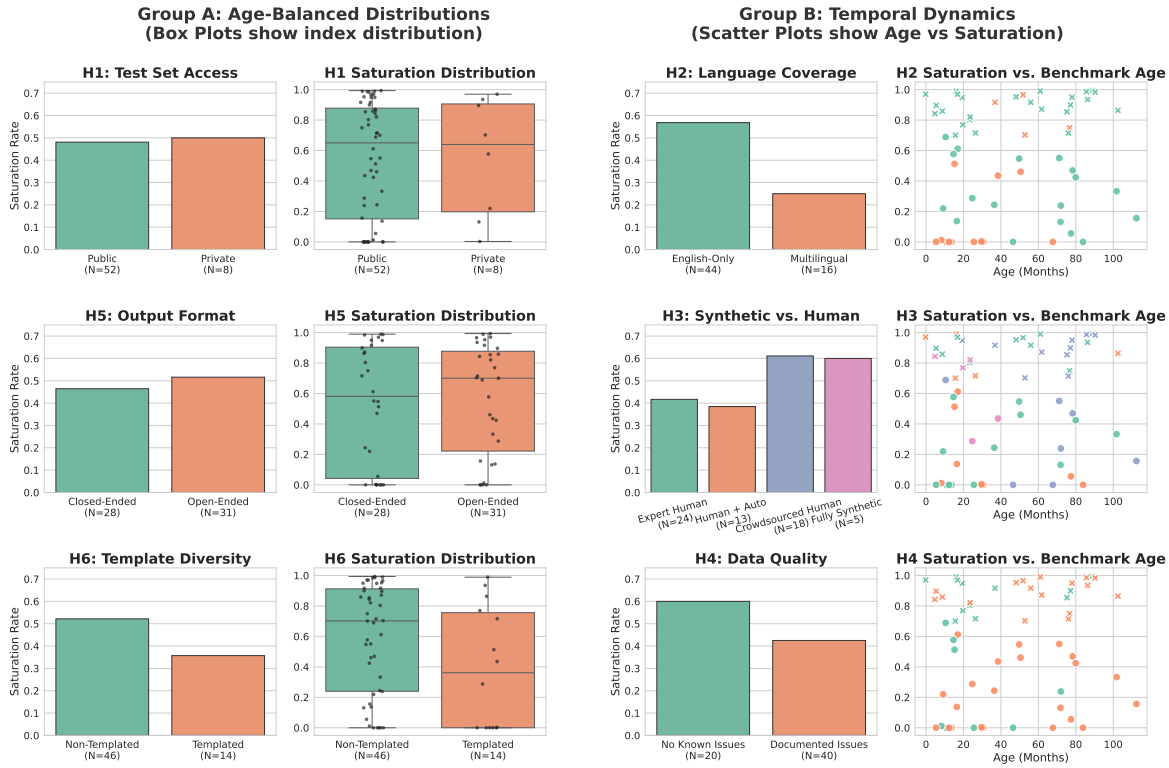

Figure 2. **Analysis of benchmark saturation ($N = 60$).** The figure is organized into two groups: **Group A (Left)** focuses on age-balanced categories (H_1, H_5, H_6), while **Group B (Right)** examines temporal dynamics (H_2, H_3, H_4), revealing that performance gaps in these categories are often driven by benchmark maturity. For each hypothesis, the first column of the group displays raw saturation rates. In the scatter plots, point colors correspond to the categories defined in the adjacent bar plots (legends omitted for brevity); $\times$ denotes saturated and $\circ$ denotes non-saturated benchmarks.

uate benchmark adoption using citation counts and inclusion in industry model release reports. Raw correlations show that benchmarks with higher citation counts tend to have higher mean saturation indices (Figure 4). However, after controlling for benchmark age, citation counts are not significantly associated with saturation ($\rho = 0.22$, $p = 0.12$). Citation growth rates ($\rho = 0.13$, $p = 0.37$) and frequency of appearance in technical reports ($\rho = 0.05$, $p = 0.73$) likewise show no significant association. These results suggest that maturity and cumulative exposure over time, rather than adoption metrics alone, better explain saturation patterns.

Accessibility and task design. Public ($N = 52$) and private ($N = 8$) benchmarks exhibit similar saturation distributions. We find no statistically meaningful difference in S_{index} between the two groups. Hiding test data does not appear to prevent saturation once benchmarks are widely adopted, rejecting hypothesis H_1. Output format is age-balanced ($p = 0.40$). We observe no meaningful difference between closed-ended ($N = 28$) and open-ended ($N = 31$) benchmarks, suggesting that generation-based evaluation does not systematically preserve longer discriminative power.

Benchmark composition and construction. English-only benchmarks ($N = 44$) show higher raw saturation rates than multilingual ones ($N = 16$). However, benchmark age is a clear cofounding factor for H_2: multilingual benchmarks in our dataset are substantially younger on average (32.9 vs. 48.9 months). This indicates that the apparent robustness of multilingual benchmarks is largely explained by their young age rather than intrinsic resistance to saturation. Accordingly, we do not find support for H_2. We further examine whether benchmark design choices influence saturation, specifically whether expert- or human-curated benchmarks are more robust than crowdsourced or synthetic ones (H_3), and whether non-templated benchmarks are more resistant than templated benchmarks (H_6). Our analysis shows that curation categories differ significantly in age ($p = 0.0017$). Crowdsourced benchmarks are older on average and exhibit higher saturation rates in raw comparisons. Expert-curated benchmarks show lower saturation at comparable ages, and several of these benchmarks (*e.g.,* ARC-AGI, BIG-Bench Hard) remain unsaturated despite prolonged exposure. Furthermore, templated benchmarks ($N = 14$) do not differ significantly from non-templated ones ($N = 46$) in saturation behaviour ($p = 0.10$). Lit-

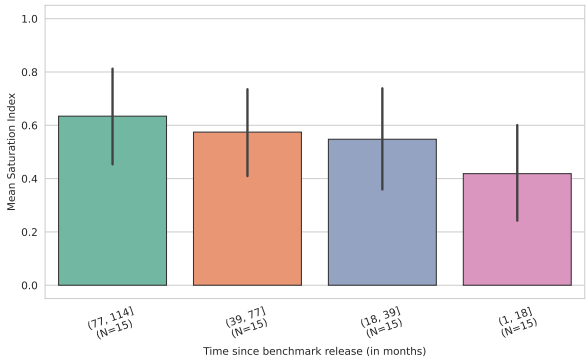

Figure 3. **Mean saturation index grouped by binned time since benchmark release (in months).** Older benchmarks exhibit higher average saturation, reflecting increasing performance compression among state-of-the-art models as benchmarks age. Error bars denote one standard deviation within each bin.

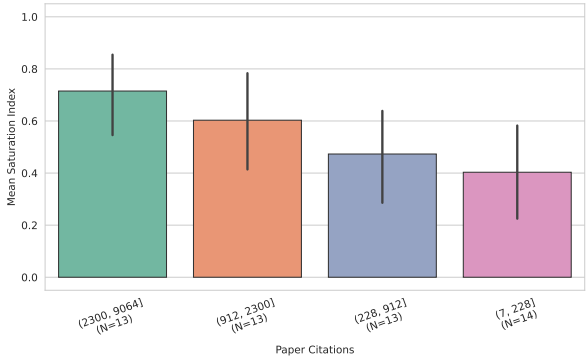

Figure 4. **Mean saturation index grouped by binned benchmark citation counts.** Benchmarks with higher citation counts exhibit higher saturation rates, suggesting that benchmark adoption and exposure are associated with reduced discriminative power over time. Error bars denote one standard deviation within each bin.

eral diversity alone does not appear to determine longevity. Fully synthetic benchmarks currently exhibit low saturation but are also relatively recent, limiting causal interpretation. These results suggest that expert-driven and adversarial design may improve robustness to saturation, though age remains a cofounding factor.

Benchmarks with documented quality issues ($N = 40$) exhibit higher saturation rates than those without ($N = 20$), but they are also significantly older on average (51.5 vs. 30.9 months; $p = 0.01$). In our annotations, documented quality issues include evidence of contamination, train-test overlap, noisy or low-quality examples, mislabeling or answer errors, documented demographic or linguistic imbalances, and other benchmark-specific problems such as unstable evaluation setups, ambiguity, or missing context. The association is consistent with multiple explanations: artifact exploitation, improved construction practices over time, or increased scrutiny of older benchmarks. Observationally, we find correlation but cannot isolate directionality.

4.2. Joint Analysis of Saturation Factors

To quantify which benchmark properties jointly explain variation in saturation, we fit a Bayesian regression model predicting S_{index} from benchmark age, test set size, adoption proxies, accessibility, output format, templating, language coverage, curation strategy, and documented quality issues. The fitted model achieves $R^2_{\text{Bayes}} = 0.884 \pm 0.012$.

Across specifications, benchmark age and test set size show the most consistent effects. Adoption metrics contribute modestly but are not robust once age is included. In contrast, accessibility (public vs. private), output format, and templating do not exhibit reliable associations with saturation. Overall, the results indicate that saturation is more strongly associated with maturity and measurement scale than with commonly assumed design safeguards.[4]

5. Synthesis and Implications

Our empirical analysis reveals a consistent pattern: benchmark saturation is primarily driven by structural exposure dynamics and measurement resolution limits, rather than by isolated design choices. While contamination, overfitting, and ceiling effects have been discussed independently in prior work (McCoy et al., 2019; Murahari et al., 2024; Schaeffer, 2023), our results clarify which factors systematically correlate with saturation across benchmarks.

5.1. Saturation as a Structural Phenomenon

Our empirical results indicate that benchmark saturation is primarily a structural consequence of exposure dynamics and measurement resolution, rather than isolated design flaws. Two variables emerge as the most consistent predictors: benchmark age and test set scale.

Age and exposure-driven compression. Older benchmarks exhibit higher saturation indices, even after controlling for adoption metrics such as citation counts or inclusion in technical reports. Once age is accounted for, these popularity proxies no longer show associations with saturation, suggesting that cumulative exposure, not popularity alone, drives convergence. Repeated optimization against a stable evaluation target progressively compresses performance differences among frontier models. Our results are consistent with this interpretation, since older benchmarks exhibit higher saturation, although our analysis does not directly identify the causal mechanism. Similar plateau dynamics have been discussed qualitatively in prior work (Ott et al., 2022). This exposure effect is consistent with known risks of familiarity and memorization. Publicly accessible benchmarks increase the possibility that evaluation data, or close

[4] See Section F in the appendix for further details.

variants, appear in training corpora (see H_1 and Section 4.1), as well as findings in the literature (Zhou et al., 2023b; Balloccu et al., 2024). However, in our analysis, private test sets do not systematically reduce saturation once benchmark age is considered, suggesting that privacy alone is not sufficient. Even without explicit leakage or contamination, our finding that older benchmarks exhibit higher saturation supports the broader mechanism that repeated exposure to fixed evaluation formats encourages benchmark-specific optimization, narrowing observable performance gaps over time.

Test set scale and measurement resolution limits. In our empirical analysis (Section 4), larger evaluation sets are consistently associated with lower saturation indices. This suggests that discriminative power depends critically on statistical resolution. When evaluation uncertainty exceeds true performance gaps, top models become statistically indistinguishable even if substantive differences remain. Smaller test sets accelerate this effect, as variance dominates observed score differences. Moreover, reliance on coarse aggregate metrics (*e.g.,* single accuracy scores) can mask residual behavioral variation across subskills or input types (Murahari et al., 2024; Saxon et al., 2024). Taken together with our finding that older benchmarks tend to be more saturated, these results suggest that saturation often reflects loss of *relative separability* among top-performing models rather than complete task mastery (which would be desirable; importantly, benchmark saturation is a neutral, not a negative phenomenon. It only becomes an issue if saturation does not reflect task mastery). Benchmark maturity increases optimization pressure, while finite evaluation resolution constrains the ability to detect incremental gains. Saturation therefore emerges from the interaction between cumulative exposure and statistical measurement limits, even in the absence of explicit contamination or fundamental capability ceilings.

5.2. Safeguards That Do Not Prevent Saturation

Although benchmark age emerges as the strongest factor and most consistently correlates with saturation, we test the remaining hypotheses to evaluate whether commonly assumed safeguards retain explanatory power once we take age into account. Our results show that, these safeguards do not show robust associations with saturation in our data.

Private test sets. *Benchmark creators should not rely on private or held-out test sets alone as a long-term defense against saturation.* In our H1 analysis, we observe similar saturation distributions and no statistically meaningful difference in S_{index} between public and private benchmarks. While contamination and memorization are well-documented risks (Zhou et al., 2023b; Balloccu et al., 2024; Deng et al., 2024; Sainz et al., 2024), secrecy alone does not prevent compression once distributional characteristics become widely known. Direct fine-tuning on evaluation data can trivially inflate scores (Schaeffer, 2023), but our results suggest that even without explicit leakage, prolonged exposure drives convergence.

Open-ended output formats. *Benchmark creators should not assume that switching from multiple-choice to open-ended generation alone will meaningfully extend benchmark usefulness over time.* In evaluating hypothesis H4, we observe no meaningful difference in saturation distributions between closed-ended (N=28) and open-ended (N=31) benchmarks. The output format comparison is age-balanced (p=0.40). Although multiple-choice benchmarks may enable overfitting strategies (Chandak et al., 2025), and models can exploit superficial cues (McCoy et al., 2019; Pacchiardi et al., 2024), format alone does not determine longevity. Compression seems to occur in both settings.

Template diversity and multilinguality. *Benchmark creators should prioritize refresh mechanisms, substantive difficulty and measurement resolution over surface-level diversity features such as templating or multilingual scope alone.* In evaluating hypothesis H6, we find that templated benchmarks (N=14) do not differ significantly from non-templated benchmarks (N=46) in saturation behaviour (p=0.10), suggesting that template diversity alone does not delay saturation. Multilingual benchmarks appear more robust in raw comparisons, but this effect is largely explained by recency. In evaluating hypothesis H2, we find that multilingual benchmarks (N=16) show lower raw saturation rates than English-only benchmarks (N=44), but this apparent advantage is confounded by benchmark maturity: multilingual benchmarks in our sample are substantially younger on average (32.9 vs. 48.9 months). While English-dominant pretraining corpora may accelerate ceiling effects on English-only tasks (Touvron et al., 2023; Wang et al., 2024a), age remains the dominant predictor.

Recent evidence from SWE-bench Verified also illustrates how plateaus can arise from evaluation artifacts rather than capability ceilings. OpenAI (2024) report that many frequently-failed tasks contain narrow or wide tests that reject functionally correct solutions and performance increasingly reflects training exposure to benchmark-associated repositories rather than general coding ability.

5.3. Structural Resistance to Saturation

A minority of benchmarks remain unsaturated despite substantial exposure. Qualitatively, these benchmarks tend to share structural properties that alter one or both of the above mechanisms. Benchmarks with adversarial or dynamic data collection (*e.g.,* Dynabench (Kiela et al., 2021)) reduce optimization stability by continuously updating the evaluation

distribution. Broad, capability-spanning initiatives such as BIG-Bench (Srivastava et al., 2023) expand coverage and limit narrow over-fitting. Holistic evaluation frameworks that track multiple behavioural dimensions (Liang et al., 2023) increase measurement granularity.

5.4. Implications for Benchmark Lifecycle Management

Our findings suggest that sustainable evaluation requires monitoring benchmark's discriminative power rather than relying on absolute score improvements. Benchmarks should be treated as evolving measurement instruments whose usefulness can reduce as models adapt to them.

Benchmark design considerations. Our findings suggest four actionable takeaways during benchmark design. *(1) Increase evaluation resolution.* Across our analyses, test set scale is one of the strongest predictors of lower saturation. Benchmark designers should therefore make sure that score differences between models exceed expected evaluation uncertainty. This can require larger test sets, harder examples, stratified reporting by subgroups of items (e.g., according to subskills), or multiple complementary metrics providing more insights into performance differences rather than a single aggregate score. *(2) Integrate dynamic benchmark updates.* Static benchmarks become easier optimization targets over time. Periodic refreshes, adversarial data collection, rotating hidden subsets, or continuously updated evaluation pools can reduce benchmark convergence resulting from exposure and prolong benchmark usefulness. *(3) Report uncertainty-aware statistics.* Integrate into leaderboard reporting confidence intervals, the spread of scores among top systems, and compression indicators in addition to aggregated peak scores. Small improvements that fall within evaluation noise should not be interpreted as meaningful progress. *(4) Define criteria for lifecycle management.* Benchmarks creators should include explicit revision, expansion, or retirement procedures during benchmark design once frontier systems become statistically indistinguishable, as highlighted in lifecycle-oriented evaluation frameworks (Hardy et al., 2024).

When is saturation desirable? Benchmark saturation is not inherently negative. If a benchmark is well-designed, valid, and measures a clearly defined capability, then convergence of top-performing models new the benchmark's ceiling may indicate genuine task mastery. In such cases, saturation reflects substantive progress: models can reliably perform the task the benchmark was intended to measure. However, saturation becomes problematic when it reflects loss of measurement resolution rather than capability completion. If score compression arises because evaluation noise exceeds true performance gaps, or because the benchmark lacks sufficient depth or coverage to differentiate increasingly capable systems, then apparent convergence may mask unresolved weaknesses. In this scenario, models may appear indistinguishable despite meaningful differences in robustness, calibration, or generalization. The key distinction is whether saturation reflects true capability attainment or reduced discriminative power: the former signals progress, while the latter calls for revision or expansion.

6. Limitations and Future work

Our benchmark selection, though criteria-driven, reflects current evaluation practices and may overrepresent widely-adopted benchmarks. Top-N leaderboard snapshots may miss saturation dynamics for sparse or inconsistently evaluated benchmarks. The saturation index further depends on currently available frontier model evaluations, which may be incomplete, selectively reported, or inconsistently updated. We assume benchmark properties are time-invariant, yet attributes like annotation diversity evolve post-release. Similarly, benchmarks themselves may change over time through revised splits, refreshed test sets, or updated protocols, which are not captured in our static annotations.

Our analysis relies on publicly available leaderboard data, which posed several challenges. Multiple leaderboards may exist for a benchmark, often differing in evaluation setups (*e.g.*, LLM-judge prompts) and scoring criteria. Many leaderboards are not regularly updated and may omit newly released models. We therefore prioritized leaderboards based on visibility, recency, and result verification, though inconsistencies remain. Finally, our uncertainty estimates are designed for accuracy-like metrics over fixed test sets; metrics such as Elo ratings, pass@k, or judge-based evaluations require tailored variance estimates.

Future work should incorporate continuous-time leaderboard data and distinguish genuine saturation from temporary plateaus. Longitudinal analysis and causal studies comparing different exposure patterns could further clarify the mechanisms driving saturation. Studying performance shifts following major model innovations could clarify whether saturation is transient or persistent.

7. Conclusion

In this work, we present a systematic analysis of benchmark saturation. By introducing an uncertainty-aware saturation index and characterizing benchmarks across multiple design dimensions, we identify which properties are associated with saturation dynamics. Our findings challenge common assumptions (e.g. the protective role of private test sets) and highlight the importance of benchmark design, scale, and lifecycle management. This work provides a foundation for more robust and sustainable evaluation practices, designing benchmarks such that they remain informative over time.

Impact Statement

Benchmark scores increasingly shape public discourse, model deployment, investment, marketing, policy decisions, and resource allocation in AI development. When saturated benchmarks are reported without appropriate context, they risk misinforming stakeholders about meaningful capability differences. Our analysis demonstrates that near-ceiling scores often fail to discriminate between models in ways that matter for downstream applications. This work encourages more careful communication of evaluation results, particularly when such results inform decisions in high-stakes domains such as healthcare, education, and public services.

Acknowledgements

We thank Siva Kantha Rao Vanama, Abhijit Ubale, Vijaya Kumar Reddy Palreddy, Shivaprasad Chitta, Sasikanth Kotti, Wm. Matthew Kennedy, and Alexander Hoyle who supported this project through annotation efforts, discussions, and comments on the paper draft.

Mubashara Akhtar was supported by the ETH AI Center through an ETH AI Center postdoctoral fellowship. Hossein A. Rahmani's effort was supported by the Engineering and Physical Sciences Research Council (EP/S021566/1). Vilém Zouhar gratefully acknowledges the support of the Google PhD Fellowship. Marek Suppa was funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I02-03-V01-00029. Jan Batzner was supported by the Federal Ministry of Research, Technology, and Space of Germany [Grant Number 16DII131]. Yanan Long thanks the TPU Research Cloud for computational support. Anka Reuel was supported by the Stanford Interdisciplinary Graduate Fellowship. Sanmi Koyejo is partially supported by NSF 2046795 and 2205329, IES R305C240046, ARPA-H, the MacArthur Foundation, Schmidt Sciences, Stanford HAI, RAISE Health, OpenAI, Microsoft, and Google.

References

Adelani, D. I., Liu, H., Shen, X., Vassilyev, N., Alabi, J. O., Mao, Y., Gao, H., and Lee, E.-S. A. SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 226–245, St. Julian's, Malta, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.14. URL https://aclanthology.org/2024.eacl-long.14.

AI, S. GSM8K-Indic: A multilingual version of GSM8K for indian languages, 2024a. URL https://huggingface.co/datasets/sarvamai/gsm8k-indic.

AI, S. MMLU-Indic: A multilingual version of MMLU for indian languages, 2024b. URL https://huggingface.co/datasets/sarvamai/mmlu-indic.

Alzahrani, N., Alyahya, H., Alnumay, Y., AlRashed, S., Alsubaie, S., Almushayqih, Y., Mirza, F., Alotaibi, N., Al-Twairesh, N., Alowisheq, A., Bari, M. S., and Khan, H. 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)*, pp. 13787–13805, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.744. URL https://aclanthology.org/2024.acl-long.744.

Ashury-Tahan, S., Gera, A., Bandel, E., Shmueli-Scheuer, M., and Choshen, L. Robustness as an Emergent Property of Task Performance, February 2026a. URL http://arxiv.org/abs/2602.03344. arXiv:2602.03344 [cs.LG].

Ashury-Tahan, S., Mai, Y., C, R., Gera, A., Perlitz, Y., Yehudai, A., Bandel, E., Choshen, L., Shnarch, E., Liang, P., and Shmueli-Scheuer, M. The Mighty ToRR: A Benchmark for Table Reasoning and Robustness, February 2026b. URL http://arxiv.org/abs/2502.19412. arXiv:2502.19412 [cs.CL].

Balloccu, S., Schmidtová, P., Lango, M., and Dusek, O. 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)*, pp. 67–93, St. Julian's, Malta, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.5. URL https://aclanthology.org/2024.eacl-long.5.

Bandarkar, L., Liang, D., Muller, B., Artetxe, M., Shukla, S. N., Husa, D., Goyal, N., Krishnan, A., Zettlemoyer, L., and Khabsa, M. The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44.

Barres, V., Dong, H., Ray, S., Si, X., and Narasimhan, K. τ^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment, 2025. URL https://arxiv.org/abs/2506.07982. Version Number: 1.

Bisk, Y., Zellers, R., Le Bras, R., Gao, J., and Choi, Y. PIQA: Reasoning about Physical Commonsense in Natural Language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7432–7439, April 2020. doi: 10.1609/aaai.v34i05.6239. URL https://ojs.aaai.org/index.php/AAAI/article/view/6239.

Cassano, F., Gouwar, J., Nguyen, D., Nguyen, S., Phipps-Costin, L., Pinckney, D., Yee, M.-H., Zi, Y., Anderson, C. J., Feldman, M. Q., Guha, A., Greenberg, M., and Jangda, A. MultiPL-E: A Scalable and Extensible Approach to Benchmarking Neural Code Generation, 2022. URL https://arxiv.org/abs/2208.08227. Version Number: 4.

Chandak, N., Goel, S., Prabhu, A., Hardt, M., and Geiping, J. Answer Matching Outperforms Multiple Choice for Language Model Evaluation, 2025. URL https://arxiv.org/abs/2507.02856. Version Number: 1.

Chen, M. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. Evaluating Large Language Models Trained on Code, 2021. URL https://arxiv.org/abs/2107.03374. Version Number: 2.

Chen, S., Chen, Y., Li, Z., Jiang, Y., Wan, Z., He, Y., Ran, D., Gu, T., Li, H., Xie, T., and Ray, B. Benchmarking Large Language Models Under Data Contamination: A Survey from Static to Dynamic Evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10091–10109, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.511. URL https://aclanthology.org/2025.emnlp-main.511.

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., Liang, P., and Zettlemoyer, L. QuAC: Question Answering in Context. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2174–2184, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1241. URL https://aclanthology.org/D18-1241/.

Chollet, F. On the Measure of Intelligence, November 2019. URL http://arxiv.org/abs/1911.01547. arXiv:1911.01547 [cs.AI].

Chollet, F., Knoop, M., Kamradt, G., Landers, B., and Pinkard, H. ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems, 2025. URL https://arxiv.org/abs/2505.11831. Version Number: 2.

Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., and Toutanova, K. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., and Palomaki, J. T y D i QA: A Benchmark for Information-Seeking Question Answering in *Ty* pologically *Di* verse Languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, December 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00317. URL https://direct.mit.edu/tacl/article/96451.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training Verifiers to Solve Math Word Problems, 2021. URL https://arxiv.org/abs/2110.14168. Version Number: 2.

Das, D., De Langis, K., Martin-Boyle, A., Kim, J., Lee, M., Kim, Z. M., Hayati, S. A., Owan, R., Hu, B., Parkar, R., Koo, R., Park, J., Tyagi, A., Ferland, L., Roy, S., Liu, V., and Kang, D. Under the Surface: Tracking the Artifactuality of LLM-Generated Data, 2024. URL https://arxiv.org/abs/2401.14698. Version Number: 2.

Deng, C., Zhao, Y., Tang, X., Gerstein, M., and Cohan, A. Investigating Data Contamination in Modern Benchmarks for Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8706–8719,

Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.482. URL https://aclanthology.org/2024.naacl-long.482.

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Miami, FL, June 2009. IEEE. ISBN 978-1-4244-3992-8. doi: 10.1109/CVPR.2009.5206848. URL https://ieeexplore.ieee.org/document/5206848/.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., and Gardner, M. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2368–2378, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1246. URL https://aclanthology.org/N19-1246/.

Gema, A. P., Leang, J. O. J., Hong, G., Devoto, A., Mancino, A. C. M., Saxena, R., He, X., Zhao, Y., Du, X., Ghasemi Madani, M. R., Barale, C., McHardy, R., Harris, J., Kaddour, J., Van Krieken, E., and Minervini, P. Are We Done with MMLU? 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)*, pp. 5069–5096, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.262. URL https://aclanthology.org/2025.naacl-long.262.

Gill, A., Ravichander, A., and Marasovic, A. What Has Been Lost with Synthetic Evaluation? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 9902–9945, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.526. URL https://aclanthology.org/2025.findings-emnlp.526.

Glazer, E., Erdil, E., Besiroglu, T., Chicharro, D., Chen, E., Gunning, A., Olsson, C. F., Denain, J.-S., Ho, A., Santos, E. d. O., Järviniemi, O., Barnett, M., Sandler, R., Vrzala, M., Sevilla, J., Ren, Q., Pratt, E., Levine, L., Barkley, G., Stewart, N., Grechuk, B., Grechuk, T., Enugandla, S. V., and Wildon, M. FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI, 2024. URL https://arxiv.org/abs/2411.04872. Version Number: 7.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., and Fan, A. The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, May 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00474. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00474/110993/The-Flores-101-Evaluation-Benchmark-for-Low.

Guha, N., Nyarko, J., Ho, D. E., Ré, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D. N., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G. M., Porat, H., Hegland, J., Wu, J., Nudell, J., Niklaus, J., Nay, J., Choi, J. H., Tobia, K., Hagan, M., Ma, M., Livermore, M., Rasumov-Rahe, N., Holzenberger, N., Kolt, N., Henderson, P., Rehaag, S., Goel, S., Gao, S., Williams, S., Gandhi, S., Zur, T., Iyer, V., and Li, Z. LEGALBENCH: a collaboratively built benchmark for measuring legal reasoning in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Gupta, V., Ross, C., Pantoja, D., Passonneau, R. J., Ung, M., and Williams, A. 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)*, pp. 4595–4615, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.235. URL https://aclanthology.org/2025.naacl-long.235.

Haas, L., Yona, G., D'Antonio, G., Goldshtein, S., and Das, D. SimpleQA Verified: A Reliable Factuality Benchmark to Measure Parametric Knowledge, 2025. URL https://arxiv.org/abs/2509.07968. Version Number: 2.

Habba, E., Arviv, O., Itzhak, I., Perlitz, Y., Bandel, E., Choshen, L., Shmueli-Scheuer, M., and Stanovsky, G. DOVE: A Large-Scale Multi-Dimensional Predictions Dataset Towards Meaningful LLM Evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11744–11763, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.611. URL https://aclanthology.org/2025.findings-acl.611.

Hardy, A., Hardy, M., Kochenderfer, M., Lamparth, M., Reuel, A., and Smith, C. BetterBench: Assessing AI

Benchmarks, Uncovering Issues, and Establishing Best Practices. In *Advances in Neural Information Processing Systems 37*, pp. 21763–21813, Vancouver, BC, Canada, 2024. Neural Information Processing Systems Foundation, Inc. (NeurIPS). ISBN 979-8-3313-1438-5. doi: 10.52202/079017-0685. URL http://www.proceedings.com/079017-0685.html.

Hardy, A., Reuel, A., Jafari Meimandi, K., Soder, L., Griffith, A., Asmar, D. M., Koyejo, S., Bernstein, M. S., and Kochenderfer, M. J. More than Marketing? On the Information Value of AI Benchmarks for Practitioners. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pp. 1032–1047, Cagliari Italy, March 2025. ACM. ISBN 979-8-4007-1306-4. doi: 10.1145/3708359.3712152. URL https://dl.acm.org/doi/10.1145/3708359.3712152.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring Massive Multitask Language Understanding, 2020. URL https://arxiv.org/abs/2009.03300. Version Number: 3.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring Mathematical Problem Solving With the MATH Dataset, 2021. URL https://arxiv.org/abs/2103.03874. Version Number: 2.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. XTREME: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Huang, Y., Bai, Y., Zhu, Z., Zhang, J., Zhang, J., Su, T., Liu, J., Lv, C., Zhang, Y., lei, j., Fu, Y., Sun, M., and He, J. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 62991–63010. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/c6ec1844bec96d6d32ae95ae694e23d8-Paper-Datasets_and_Benchmarks.pdf.

Jacovi, A., Wang, A., Alberti, C., Tao, C., Lipovetz, J., Olszewska, K., Haas, L., Liu, M., Keating, N., Bloniarz, A., Saroufim, C., Fry, C., Marcus, D., Kukliansky, D., Tomar, G. S., Swirhun, J., Xing, J., Wang, L., Aaron, M., Ambar, M., Fellinger, R., Wang, R., Sims, R., Zhang, Z., Goldshtein, S., Matias, Y., and Das, D. FACTS leaderboard. https://kaggle.com/facts-leaderboard, 2024. URL https://www.kaggle.com/benchmarks/google/facts. Google DeepMind, Google Research, Google Cloud, Kaggle.

Jacovi, A., Wang, A., Alberti, C., Tao, C., Lipovetz, J., Olszewska, K., Haas, L., Liu, M., Keating, N., Bloniarz, A., Saroufim, C., Fry, C., Marcus, D., Kukliansky, D., Tomar, G. S., Swirhun, J., Xing, J., Wang, L., Gurumurthy, M., Aaron, M., Ambar, M., Fellinger, R., Wang, R., Zhang, Z., Goldshtein, S., and Das, D. The FACTS Grounding Leaderboard: Benchmarking LLMs' Ability to Ground Responses to Long-Form Input, 2025. URL https://arxiv.org/abs/2501.03200. Version Number: 1.

Jain, N., Han, Gu, A., Li, W.-D., Yan, F., Zhang, T., Wang, S., Solar-Lezama, A., Sen, K., and Stoica, I. LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. In Yue, Y., Garg, A., Peng, N., Sha, F., and Yu, R. (eds.), *International Conference on Learning Representations*, volume 2025, pp. 58791–58831, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/94074dd5a072d28ff75a76dabed43767-Paper-Conference.pdf.

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. SWE-bench: Can Language Models Resolve Real-world Github Issues? In Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., and Sun, Y. (eds.), *International Conference on Learning Representations*, volume 2024, pp. 54107–54157, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/edac78c3e300629acfe6cbe9ca88fb84-Paper-Conference.pdf.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., and Szolovits, P. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences*, 11(14):6421, July 2021. ISSN 2076-3417. doi: 10.3390/app11146421. URL https://www.mdpi.com/2076-3417/11/14/6421.

Joshi, M., Choi, E., Weld, D., and Zettlemoyer, L. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL http://aclweb.org/anthology/P17-1147.

Justen, L. LLMs Outperform Experts on Challenging Biology Benchmarks, 2025. URL https://arxiv.org/abs/2505.06108. Version Number: 3.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma,

Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., and Williams, A. 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*, pp. 4110–4124, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324.

Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Gowda, T., Graham, Y., Grundkiewicz, R., Haddow, B., Knowles, R., Koehn, P., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Novák, M., Popel, M., and Popović, M. Findings of the 2022 Conference on Machine Translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 1–45, Abu Dhabi, United Arab Emirates (Hybrid), 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.wmt-1.1. URL https://aclanthology.org/2022.wmt-1.1.

Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Koehn, P., Marie, B., Monz, C., Morishita, M., Murray, K., Nagata, M., Nakazawa, T., Popel, M., Popović, M., and Shmatova, M. Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet. In *Proceedings of the Eighth Conference on Machine Translation*, pp. 1–42, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.1. URL https://aclanthology.org/2023.wmt-1.1.

Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Karpinska, M., Koehn, P., Marie, B., Monz, C., Murray, K., Nagata, M., Popel, M., Popović, M., Shmatova, M., Steingrímsson, S., and Zouhar, V. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In *Proceedings of the Ninth Conference on Machine Translation*, pp. 1–46, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.1. URL https://aclanthology.org/2024.wmt-1.1.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:453–466, November 2019. ISSN 2307-387X. doi: 10.1162/tacl_a_00276. URL https://direct.mit.edu/tacl/article/43518.

Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1082. URL http://aclweb.org/anthology/D17-1082.

Lambert, N., Pyatkin, V., Morrison, J., Miranda, L., Lin, B. Y., Chandu, K., Dziri, N., Kumar, S., Zick, T., Choi, Y., Smith, N. A., and Hajishirzi, H. RewardBench: Evaluating Reward Models for Language Modeling, 2025. URL https://aclanthology.org/2025.findings-naacl.96.

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., Wu, Y., Neyshabur, B., Gur-Ari, G., and Misra, V. Solving Quantitative Reasoning Problems with Language Models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 3843–3857. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/18abbeef8cfe9203fdf9053c9c4fe191-Paper-Conference.pdf.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Re, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., Wang, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N. S., Khattab, O., Henderson, P., Huang, Q., Chi, R. A., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., and Koreeda, Y. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*, 2023. URL https://mlanthology.org/tmlr/2023/liang2023tmlr-holistic/.

Lin, S., Hilton, J., and Evans, O. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

Li∗, T., Chiang∗, W.-L., Frick, E., Dunlap, L., Zhu, B., Gonzalez, J. E., and Stoica, I. From live data to high-quality benchmarks: The arena-hard pipeline, April 2024. URL https://lmsys.org/blog/2024-04-19-arena-hard/.

Liu, Y. L., Blodgett, S. L., Cheung, J. C. K., Liao, Q. V., Olteanu, A., and Xiao, Z. ECBD: Evidence-Centered Benchmark Design for NLP, 2024. URL https://arxiv.org/abs/2406.08723. Version Number: 1.

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., and Clark, J. Artificial Intelligence Index Report 2024, 2024. URL https://arxiv.org/abs/2405.19522. Version Number: 1.

McCoy, R. T., Pavlick, E., and Linzen, T. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334.

Merrill, M. A., Shaw, A. G., Carlini, N., Li, B., Raj, H., Bercovich, I., Shi, L., Shin, J. Y., Walshe, T., Buchanan, E. K., Shen, J., Ye, G., Lin, H., Poulos, J., Wang, M., Nezhurina, M., Jitsev, J., Lu, D., Mastromichalakis, O. M., Xu, Z., Chen, Z., Liu, Y., Zhang, R., Chen, L. L., Kashyap, A., Uslu, J.-L., Li, J., Wu, J., Yan, M., Bian, S., Sharma, V., Sun, K., Dillmann, S., Anand, A., Lanpouthakoun, A., Koopah, B., Hu, C., Guha, E., Dreiman, G. H. S., Zhu, J., Krauth, K., Zhong, L., Muennighoff, N., Amanfu, R., Tan, S., Pimpalgaonkar, S., Aggarwal, T., Lin, X., Lan, X., Zhao, X., Liang, Y., Wang, Y., Wang, Z., Zhou, C., Heineman, D., Liu, H., Trivedi, H., Yang, J., Lin, J., Shetty, M., Yang, M., Omi, N., Raoof, N., Li, S., Zhuo, T. Y., Lin, W., Dai, Y., Wang, Y., Chai, W., Zhou, S., Wahdany, D., She, Z., Hu, J., Dong, Z., Zhu, Y., Cui, S., Saiyed, A., Kolbeinsson, A., Hu, J., Rytting, C. M., Marten, R., Wang, Y., Dimakis, A., Konwinski, A., and Schmidt, L. Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces, 2026. URL https://arxiv.org/abs/2601.11868. Version Number: 1.

Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL http://aclweb.org/anthology/D18-1260.

Murahari, V., Deshpande, A., Clark, P., Rajpurohit, T., Sabharwal, A., Narasimhan, K., and Kalyan, A. QualEval: Qualitative Evaluation for Model Improvement, May 2024. URL http://arxiv.org/abs/2311.02807. arXiv:2311.02807 [cs.LG].

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4885–4901, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://www.aclweb.org/anthology/2020.acl-main.441.

OpenAI. Dataset Language Statistics. Technical report, OpenAI, 2020. URL https://github.com/openai/gpt-3/tree/master/dataset_statistics.

OpenAI. MMMLU: Multilingual massive multitask language understanding, 2024. URL https://huggingface.co/datasets/openai/MMMLU.

OpenAI. Why we no longer evaluate on swe-bench verified. https://openai.com/index/why-we-no-longer-evaluate-swe-bench-verified/, 2024. Accessed: 2026-05-19.

Ott, S., Barbosa-Silva, A., Blagec, K., Brauner, J., and Samwald, M. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13(1):6793, November 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34591-0. URL https://www.nature.com/articles/s41467-022-34591-0.

Pacchiardi, L., Tesic, M., Cheke, L. G., and Hernández-Orallo, J. Leaving the barn door open for Clever Hans: Simple features predict LLM benchmark answers, 2024. URL https://arxiv.org/abs/2410.11672. Version Number: 1.

Pang, R. Y., Parrish, A., Joshi, N., Nangia, N., Phang, J., Chen, A., Padmakumar, V., Ma, J., Thompson, J., He, H., and Bowman, S. QuALITY: Question Answering with Long Input Texts, Yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5336–5358, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.391. URL https://aclanthology.org/2022.naacl-main.391.

Paperno, D., Kruszewski, G., Lazaridou, A., Pham, N. Q., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., and Fernandez, R. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1525–1534, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1144. URL http://aclweb.org/anthology/P16-1144.

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., Choi, M., Agrawal, A., Chopra, A., Khoja, A., Kim, R., Ren, R., Hausenloy, J., Zhang, O., Mazeika, M., Dodonov, D., Nguyen, T., Lee, J., Anderson, D., Doroshenko, M., Stokes, A. C., Mahmood, M., Pokutnyi, O., Iskra, O., Wang, J. P., Levin, J.-C., Kazakov, M., Feng, F., Feng, S. Y., Zhao, H., Yu, M., Gangal, V., Zou, C., Wang, Z., Popov, S., Gerbicz, R., Galgon, G., Schmitt, J., Yeadon, W., Lee, Y., Sauers, S., Sanchez, A., Giska, F., Roth, M., Riis, S., Utpala, S., Burns, N., Goshu, G. M., Naiya, M. M., Agu, C., Giboney, Z., Cheatom, A., Fournier-Facio, F., Crowson, S.-J., Finke, L., Cheng, Z., Zampese, J., Hoerr, R. G., Nandor, M., Park, H., Gehrunger, T., Cai, J., McCarty, B., Garretson, A. C., Taylor, E., Sileo, D., Ren, Q., Qazi, U., Li, L., Nam, J., Wydallis, J. B., Arkhipov, P., Shi, J. W. L., Bacho, A., Willcocks, C. G., Cao, H., Motwani, S., Santos, E. d. O., Veith, J., Vendrow, E., Cojoc, D., Zenitani, K., Robinson, J., Tang, L., Li, Y., Vendrow, J., Fraga, N. W., Kuchkin, V., Maksimov, A. P., Marion, P., Efremov, D., Lynch, J., Liang, K., Mikov, A., Gritsevskiy, A., Guillod, J., Demir, G., Martinez, D., Pageler, B., Zhou, K., Soori, S., Press, O., Tang, H., Rissone, P., Green, S. R., Brüssel, L., Twayana, M., Dieuleveut, A., Imperial, J. M., Prabhu, A., Yang, J., Crispino, N., Rao, A., Zvonkine, D., Loiseau, G., Kalinin, M., Lukas, M., Manolescu, C., Stambaugh, N., Mishra, S., Hogg, T., Bosio, C., Coppola, B. P., Salazar, J., Jin, J., Sayous, R., Ivanov, S., Schwaller, P., Senthilkuma, S., Bran, A. M., Algaba, A., Houte, K. V. d., Van Der Sypt, L., Verbeken, B., Noever, D., Kopylov, A., Myklebust, B., Li, B., Schut, L., Zheltonozhskii, E., Yuan, Q., Lim, D., Stanley, R., Yang, T., Maar, J., Wykowski, J., Oller, M., Sahu, A., Ardito, C. G., Hu, Y., Kamdoum, A. G. K., Jin, A., Vilchis, T. G., Zu, Y., Lackner, M., Koppel, J., Sun, G., Antonenko, D. S., Chern, S., Zhao, B., Arsene, P., Cavanagh, J. M., Li, D., Shen, J., Crisostomi, D., Zhang, W., Dehghan, A., Ivanov, S., Perrella, D., Kaparov, N., Zang, A., Sucholutsky, I., Kharlamova, A., Orel, D., Poritski, V., Ben-David, S., Berger, Z., Whitfill, P., Foster, M., Munro, D., Ho, L., Sivarajan, S., Hava, D. B., Kuchkin, A., Holmes, D., Rodriguez-Romero, A., Sommerhage, F., Zhang, A., Moat, R., Schneider, K., Kazibwe, Z., Clarke, D., Kim, D. H., Dias, F. M., Fish, S., Elser, V., Kreiman, T., Vilchis, V. E. G., Klose, I., Anantheswaran, U., Zweiger, A., Rawal, K., Li, J., Nguyen, J., Daans, N., Heidinger, H., Radionov, M., Rozhoň, V., Ginis, V., Stump, C., Cohen, N., Poświata, R., Tkadlec, J., Goldfarb, A., Wang, C., Padlewski, P., Barzowski, S., Montgomery, K., Stendall, R., Tucker-Foltz, J., Stade, J., Rogers, T. R., Goertzen, T., Grabb, D., Shukla, A., Givré, A., Ambay, J. A., Sen, A., Aziz, M. F., Inlow, M. H., He, H., Zhang, L., Kaddar, Y., Ängquist, I., Chen, Y., Wang, H. K., Ramakrishnan, K., Thornley, E., Terpin, A., Schoelkopf, H., Zheng, E., Carmi, A., Brown, E. D. L., Zhu, K., Bartolo, M., Wheeler, R., Stehberger, M., Bradshaw, P., Heimonen, J., Sridhar, K., Akov, I., Sandlin, J., Makarychev, Y., Tam, J., Hoang, H., Cunningham, D. M., Goryachev, V., Patramanis, D., Krause, M., Redenti, A., Aldous, D., Lai, J., Coleman, S., Xu, J., Lee, S., Magoulas, I., Zhao, S., Tang, N., Cohen, M. K., Paradise, O., Kirchner, J. H., Ovchynnikov, M., Matos, J. O., Shenoy, A., Wang, M., Nie, Y., Sztyber-Betley, A., Faraboschi, P., Riblet, R., Crozier, J., Halasyamani, S., Verma, S., Joshi, P., Meril, E., Ma, Z., Andréoletti, J., Singhal, R., Platnick, J., Nevirkovets, V., Basler, L., Ivanov, A., Khoury, S., Gustafsson, N., Piccardo, M., Mostaghimi, H., Chen, Q., Singh, V., Khánh, T. Q., Rosu, P., Szlyk, H., Brown, Z., Narayan, H., Menezes, A., Roberts, J., Alley, W., Sun, K., Patel, A., Lamparth, M., Reuel, A., Xin, L., Xu, H., Loader, J., Martin, F., Wang, Z., Achilleos, A., Preu, T., Korbak, T., Bosio, I., Kazemi, F., Chen, Z., Bálint, B., Lo, E. J. Y., Wang, J., Nunes, M. I. S., Milbauer, J., Bari, M. S., Wang, Z., Ansarinejad, B., Sun, Y., Durand, S., Elgnainy, H., Douville, G., Tordera, D., Balabanian, G., Wolff, H., Kvistad, L., Milliron, H., Sakor, A., Eron, M., O., A. F. D., Shah, S., Zhou, X., Kamalov, F., Abdoli, S., Santens, T., Barkan, S., Tee, A., Zhang, R., Tomasiello, A., De Luca, G. B., Looi, S.-Z., Le, V.-K., Kolt, N., Pan, J., Rodman, E., Drori, J., Fossum, C. J., Muennighoff, N., Jagota, M., Pradeep, R., Fan, H., Eicher, J., Chen, M., Thaman, K., Merrill, W., Firsching, M., Harris, C., Ciobâcă, S., Gross, J., Pandey, R., Gusev, I., Jones, A., Agnihotri, S., Zhelnov, P., Mofayezi, M., Piperski, A., Zhang, D. K., Dobarskyi, K., Leventov, R., Soroko, I., Duersch, J., Taamazyan, V., Ho, A., Ma, W., Held, W., Xian, R., Zebaze, A. R., Mohamed, M., Leser, J. N., Yuan, M. X., Yacar, L., Lengler, J., Olszewska, K., Di Fratta, C., Oliveira, E., Jackson, J. W., Zou, A., Chidambaram, M., Manik, T., Haffenden, H., Stander, D., Dasouqi, A., Shen, A., Golshani, B., Stap, D., Kretov, E., Uzhou, M., Zhidkovskaya, A. B., Winter, N., Rodriguez, M. O., Lauff, R., Wehr, D., Tang, C., Hossain, Z., Phillips, S., Samuele, F., Ekström, F., Hammon, A., Patel, O., Farhidi, F., Medley, G., Mohammadzadeh, F., Peñaflor, M., Kassahun, H., Friedrich, A., Perez, R. H., Pyda, D., Sakal, T., Dhamane, O., Mirabadi, A. K., Hallman, E., Okutsu, K., Battaglia, M., Magh-

soudimehrabani, M., Amit, A., Hulbert, D., Pereira, R., Weber, S., Handoko, Peristyy, A., Malina, S., Mehkary, M., Aly, R., Reidegeld, F., Dick, A.-K., Friday, C., Singh, M., Shapourian, H., Kim, W., Costa, M., Gurdogan, H., Kumar, H., Ceconello, C., Zhuang, C., Park, H., Carroll, M., Tawfeek, A. R., Steinerberger, S., Aggarwal, D., Kirchhof, M., Dai, L., Kim, E., Ferret, J., Shah, J., Wang, Y., Yan, M., Burdzy, K., Zhang, L., Franca, A., Pham, D. T., Loh, K. Y., Jackson, A., Giordano, P., Petersen, P., Cosma, A., Colino, J., White, C., Votava, J., Vinnikov, V., Delaney, E., Spelda, P., Stritecky, V., Shahid, S. M., Mourrat, J.-C., Vetoshkin, L., Sponselee, K., Bacho, R., Yong, Z.-X., de la Rosa, F., Cho, N., Li, X., Malod, G., Weller, O., Albani, G., Lang, L., Laurendeau, J., Kazakov, D., Adesanya, F., Portier, J., Hollom, L., Souza, V., Zhou, Y. A., Degorre, J., Yalın, Y., Obikoya, G. D., Rai, Bigi, F., Boscá, M. C., Shumar, O., Bacho, K., Recchia, G., Popescu, M., Shulga, N., Tanwie, N. M., Lux, T. C. H., Rank, B., Ni, C., Brooks, M., Yakimchyk, A., Huanxu, Liu, Cavalleri, S., Häggström, O., Verkama, E., Newbould, J., Gundlach, H., Brito-Santana, L., Amaro, B., Vajipey, V., Grover, R., Wang, T., Kratish, Y., Li, W.-D., Gopi, S., Caciolai, A., de Witt, C. S., Hernández-Cámara, P., Rodolà, E., Robins, J., Williamson, D., Cheng, V., Raynor, B., Qi, H., Segev, B., Fan, J., Martinson, S., Wang, E. Y., Hausknecht, K., Brenner, M. P., Mao, M., Demian, C., Kassani, P., Zhang, X., Avagian, D., Scipio, E. J., Ragoler, A., Tan, J., Sims, B., Plecnik, R., Kirtland, A., Bodur, O. F., Shinde, D. P., Labrador, Y. C. L., Adoul, Z., Zekry, M., Karakoc, A., Santos, T. C. B., Shamseldeen, S., Karim, L., Liakhovitskaia, A., Resman, N., Farina, N., Gonzalez, J. C., Maayan, G., Anderson, E., Pena, R. D. O., Kelley, E., Mariji, H., Pouriamanesh, R., Wu, W., Finocchio, R., Alarab, I., Cole, J., Ferreira, D., Johnson, B., Safdari, M., Dai, L., Arthornthurasuk, S., McAlister, I. C., Moyano, A. J., Pronin, A., Fan, J., Ramirez-Trinidad, A., Malysheva, Y., Pottmaier, D., Taheri, O., Stepanic, S., Perry, S., Askew, L., Rodríguez, R. A. H., Minissi, A. M. R., Lorena, R., Iyer, K., Fasiludeen, A. A., Clark, R., Ducey, J., Piza, M., Somrak, M., Vergo, E., Qin, J., Borbás, B., Chu, E., Lindsey, J., Jallon, A., McInnis, I. M. J., Chen, E., Semler, A., Gloor, L., Shah, T., Carauleanu, M., Lauer, P., Huy, T. D., Shahrtash, H., Duc, E., Lewark, L., Brown, A., Albanie, S., Weber, B., Vaz, W. S., Clavier, P., Fan, Y., Silva, G. P. R. e., Long, Lian, Abramovitch, M., Jiang, X., Mendoza, S., Islam, M., Gonzalez, J., Mavroudis, V., Xu, J., Kumar, P., Goswami, L. P., Bugas, D., Heydari, N., Jeanplong, F., Jansen, T., Pinto, A., Apronti, A., Galal, A., Ze-An, N., Singh, A., Jiang, T., Xavier, J. o. A., Agarwal, K. P., Berkani, M., Zhang, G., Du, Z., Junior, B. A. d. O., Malishev, D., Remy, N., Hartman, T. D., Tarver, T., Mensah, S., Loume, G. A., Morak, W., Habibi, F., Hoback, S., Cai, W., Gimenez, J., Montecillo, R. G., Łucki, J., Campbell, R., Sharma, A., Meer, K., Gul, S., Gonzalez, D. E., Alapont, X., Hoover, A., Chhablani, G., Vargus, F., Agarwal, A., Jiang, Y., Patil, D., Outevsky, D., Scaria, K. J., Maheshwari, R., Dendane, A., Shukla, P., Cartwright, A., Bogdanov, S., Mündler, N., Möller, S., Arnaboldi, L., Thaman, K., Siddiqi, M. R., Saxena, P., Gupta, H., Fruhauff, T., Sherman, G., Vincze, M., Usawasutsakorn, S., Ler, D., Radhakrishnan, A., Enyekwe, I., Salauddin, S. M., Muzhen, J., Maksapetyan, A., Rossbach, V., Harjadi, C., Bahaloohoreh, M., Sparrow, C., Sidhu, J., Ali, S., Bian, S., Lai, J., Singer, E., Uro, J. L., Bateman, G., Sayed, M., Menshawy, A., Duclosel, D., Bezzi, D., Jain, Y., Aaron, A., Tiryakioglu, M., Siddh, S., Krenek, K., Shah, I. A., Jin, J., Creighton, S., Peskoff, D., EL-Wasif, Z., P, R., Richmond, M., McGowan, J., Patwardhan, T., Sun, H.-Y., Sun, T., Zubić, N., Sala, S., Ebert, S., Kaddour, J., Schottdorf, M., Wang, D., Petruzella, G., Meiburg, A., Medved, T., ElSheikh, A., Hebbar, S. A., Vaquero, L., Yang, X., Poulos, J., Zouhar, V., Bogdanik, S., Zhang, M., Sanz-Ros, J., Anugraha, D., Dai, Y., Nhu, A. N., Wang, X., Demircali, A. A., Jia, Z., Zhou, Y., Wu, J., He, M., Chandok, N., Sinha, A., Luo, G., Le, L., Noyé, M., Perełkiewicz, M., Pantidis, I., Qi, T., Purohit, S. S., Parcalabescu, L., Nguyen, T.-H., Winata, G. I., Ponti, E. M., Li, H., Dhole, K., Park, J., Abbondanza, D., Wang, Y., Nayak, A., Caetano, D. M., Wong, A. A. W. L., del Rio-Chanona, M., Kondor, D., Francois, P., Chalstrey, E., Zsambok, J., Hoyer, D., Reddish, J., Hauser, J., Rodrigo-Ginés, F.-J., Datta, S., Shepherd, M., Kamphuis, T., Zhang, Q., Kim, H., Sun, R., Yao, J., Dernoncourt, F., Krishna, S., Rismanchian, S., Pu, B., Pinto, F., Wang, Y., Shridhar, K., Overholt, K. J., Briia, G., Nguyen, H., David, Bartomeu, S., Pang, T. C., Wecker, A., Xiong, Y., Li, F., Huber, L. S., Jaeger, J., De Maddalena, R., Lù, X. H., Zhang, Y., Beger, C., Kon, P. T. J., Li, S., Sanker, V., Yin, M., Liang, Y., Zhang, X., Agrawal, A., Yifei, L. S., Zhang, Z., Cai, M., Sonmez, Y., Cozianu, C., Li, C., Slen, A., Yu, S., Park, H. K., Sarti, G., Briański, M., Stolfo, A., Nguyen, T. A., Zhang, M., Perlitz, Y., Hernandez-Orallo, J., Li, R., Shabani, A., Juefei-Xu, F., Dhingra, S., Zohar, O., Nguyen, M. C., Pondaven, A., Yilmaz, A., Zhao, X., Jin, C., Jiang, M., Todoran, S., Han, X., Kreuer, J., Rabern, B., Plassart, A., Maggetti, M., Yap, L., Geirhos, R., Kean, J., Wang, D., Mollaei, S., Sun, C., Yin, Y., Wang, S., Li, R., Chang, Y., Wei, A., Bizeul, A., Wang, X., Arrais, A. O., Mukherjee, K., Chamorro-Padial, J., Liu, J., Qu, X., Guan, J., Bouyamourn, A., Wu, S., Plomecka, M., Chen, J., Tang, M., Deng, J., Subramanian, S., Xi, H., Chen, H., Zhang, W., Ren, Y., Tu, H., Kim, S., Chen, Y., Marjanović, S. V., Ha, J., Luczyna, G., Ma, J. J., Shen, Z., Song, D., Zhang, C. E., Wang, Z., Gendron, G., Xiao, Y., Smucker, L., Weng, E., Lee, K. H., Ye, Z., Ermon, S., Lopez-Miguel, I. D., Knights, T., Gitter, A., Park, N., Wei, B., Chen, H., Pai, K., Elkhanany, A., Lin, H., Siedler, P. D., Fang,

J., Mishra, R., Zsolnai-Fehér, K., Jiang, X., Khan, S., Yuan, J., Jain, R. K., Lin, X., Peterson, M., Wang, Z., Malusare, A., Tang, M., Gupta, I., Fosin, I., Kang, T., Dworakowska, B., Matsumoto, K., Zheng, G., Sewuster, G., Villanueva, J. P., Rannev, I., Chernyavsky, I., Chen, J., Banik, D., Racz, B., Dong, W., Wang, J., Bashmal, L., Gonçalves, D. V., Hu, W., Bar, K., Bohdal, O., Patlan, A. S., Dhuliawala, S., Geirhos, C., Wist, J., Kansal, Y., Chen, B., Tire, K., Yücel, A. T., Christof, B., Singla, V., Song, Z., Chen, S., Ge, J., Ponkshe, K., Park, I., Shi, T., Ma, M. Q., Mak, J., Lai, S., Moulin, A., Cheng, Z., Zhu, Z., Zhang, Z., Patil, V., Jha, K., Men, Q., Wu, J., Zhang, T., Vieira, B. H., Aji, A. F., Chung, J.-W., Mahfoud, M., Hoang, H. T., Sperzel, M., Hao, W., Meding, K., Xu, S., Kostakos, V., Manini, D., Liu, Y., Toukmaji, C., Paek, J., Yu, E., Demircali, A. E., Sun, Z., Dewerpe, I., Qin, H., Pflugfelder, R., Bailey, J., Morris, J., Heilala, V., Rosset, S., Yu, Z., Chen, P. E., Yeo, W., Jain, E., Yang, R., Chigurupati, S., Chernyavsky, J., Reddy, S. P., Venugopalan, S., Batra, H., Park, C. F., Tran, H., Maximiano, G., Zhang, G., Liang, Y., Shiyu, H., Xu, R., Pan, R., Suresh, S., Liu, Z., Gulati, S., Zhang, S., Turchin, P., Bartlett, C. W., Scotese, C. R., Cao, P. M., Wu, B., Karwowski, J., Scaramuzza, D., Nattanmai, A., McKellips, G., Cheraku, A., Suhail, A., Luo, E., Deng, M., Luo, J., Zhang, A., Jindel, K., Halevy, K., Baranov, A., Liu, M., Avadhanam, A., Zhang, D., Ma, B., Fu, E., Do, L., Lass, J., Yang, H., Sunkari, S., Bharath, V., Ai, V., Leung, J., Agrawal, R., Zhou, A., Chen, K., Kalpathi, T., Xu, Z., Wang, G., Xiao, T., Maung, E., Lee, S., Yue, R., Zhao, B., Yoon, J., Sun, S., Singh, A., Peng, C., Osbey, T., Wang, T., Echeazu, D., Wu, T., Patel, S., Kulkarni, V., Sundarapandiyan, V., Le, A., Nasim, Z., Yalam, S., Kasamsetty, R., Samal, S., Sun, D., Shah, N., Saha, A., Zhang, A., Nguyen, L., Nagumalli, L., Wang, K., Wu, A., Telluri, A., Dillmann, S., Wang, Z., Luo, J., Lunn, H., Gazizov, A., Qiu, H., Hart, A. G., Gabrielsson, R. B., Lukoianov, A., Yue, S., Wang, A., and Hendrycks, D. Humanity's Last Exam, 2025. URL https://arxiv.org/abs/2501.14249. Version Number: 10.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., and Hanna, A. AI and the Everything in the Whole Wide World Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Curran, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf. arXiv: 2111.15366.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL http://aclweb.org/anthology/D16-1264.

Rajpurkar, P., Jia, R., and Liang, P. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2124. URL http://aclweb.org/anthology/P18-2124.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. GPQA: A Graduate-Level Google-Proof Q&A Benchmark, 2023. URL https://arxiv.org/abs/2311.12022. Version Number: 1.

Sainz, O., García-Ferrero, I., Jacovi, A., Ander Campos, J., Elazar, Y., Agirre, E., Goldberg, Y., Chen, W.-L., Chim, J., Choshen, L., D'Amico-Wong, L., Dell, M., Fan, R.-Z., Golchin, S., Li, Y., Liu, P., Pahwa, B., Prabhu, A., Sharma, S., Silcock, E., Solonko, K., Stap, D., Surdeanu, M., Tseng, Y.-M., Udandarao, V., Wang, Z., Xu, R., and Yang, J. Data Contamination Report from the 2024 CONDA Shared Task. In *Proceedings of the 1st Workshop on Data Contamination (CONDA)*, pp. 41–56, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.conda-1.4. URL https://aclanthology.org/2024.conda-1.4.

Sakaguchi, K., Le Bras, R., Bhagavatula, C., and Choi, Y. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740, April 2020. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v34i05.6399. URL https://ojs.aaai.org/index.php/AAAI/article/view/6399.

Salaudeen, O. E., Reuel, A., Ahmed, A. M., Bedi, S., Robertson, Z., Sundar, S., Domingue, B. W., Wang, A., and Koyejo, S. Measurement to Meaning: A Validity-Centered Framework for AI Evaluation. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025. URL https://openreview.net/forum?id=2Bw6uC49QF.

Sap, M., Rashkin, H., Chen, D., Le Bras, R., and Choi, Y. Social IQa: Commonsense Reasoning about Social Interactions. 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)*, pp. 4462–4472, Hong Kong, China, 2019. Association for Computational

Linguistics. doi: 10.18653/v1/D19-1454. URL https://www.aclweb.org/anthology/D19-1454.

Saxon, M., Holtzman, A., West, P., Wang, W. Y., and Saphra, N. Benchmarks as Microscopes: A Call for Model Metrology, 2024. URL https://arxiv.org/abs/2407.16711. Version Number: 2.

Schaeffer, R. Pretraining on the Test Set Is All You Need, 2023. URL https://arxiv.org/abs/2309.08632. Version Number: 1.

Shi, F., Suzgun, M., Freitag, M., Wang, X., Srivats, S., Vosoughi, S., Chung, H. W., Tay, Y., Ruder, S., Zhou, D., Das, D., and Wei, J. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=fR3wGCk-IXp.

Singh, S., Romanou, A., Fourrier, C., Adelani, D. I., Ngui, J. G., Vila-Suero, D., Limkonchotiwat, P., Marchisio, K., Leong, W. Q., Susanto, Y., Ng, R., Longpre, S., Ruder, S., Ko, W.-Y., Bosselut, A., Oh, A., Martins, A., Choshen, L., Ippolito, D., Ferrante, E., Fadaee, M., Ermis, B., and Hooker, S. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18761–18799, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.919. URL https://aclanthology.org/2025.acl-long.919.

Son, G., Hong, J., Ko, H., and Thorne, J. Linguistic Generalizability of Test-Time Scaling in Mathematical Reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14333–14368, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.699. URL https://aclanthology.org/2025.acl-long.699.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A. S., Andreassen, A. J., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A. M., La, A., Lampinen, A. K., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakaş, A., Roberts, B. R., Loe, B. S., Zoph, B., Bojanowski, B., Özyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B. Y., Howald, B., Orinion, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ferri, C., Singh, C., Rathkopf, C., Meng, C., Baral, C., Wu, C., Callison-Burch, C., Waites, C., Voigt, C., Manning, C. D., Potts, C., Ramirez, C., Rivera, C. E., Siro, C., Raffel, C., Ashcraft, C., Garbacea, C., Sileo, D., Garrette, D., Hendrycks, D., Kilman, D., Roth, D., Freeman, C. D., Khashabi, D., Levy, D., González, D. M., Perszyk, D., Hernandez, D., Chen, D., Ippolito, D., Gilboa, D., Dohan, D., Drakard, D., Jurgens, D., Datta, D., Ganguli, D., Emelin, D., Kleyko, D., Yuret, D., Chen, D., Tam, D., Hupkes, D., Misra, D., Buzan, D., Mollo, D. C., Yang, D., Lee, D.-H., Schrader, D., Shutova, E., Cubuk, E. D., Segal, E., Hagerman, E., Barnes, E., Donoway, E., Pavlick, E., Rodolà, E., Lam, E., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E. A., Dyer, E., Jerzak, E., Kim, E., Manyasi, E. E., Zheltonozhskii, E., Xia, F., Siar, F., Martínez-Plumed, F., Happé, F., Chollet, F., Rong, F., Mishra, G., Winata, G. I., Melo, G. d., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G. X., Jaimovitch-Lopez, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H. F. A., Schuetze, H., Yakura, H., Zhang, H., Wong, H. M., Ng, I., Noble, I., Jumelet, J., Geissinger, J., Kernion, J., Hilton, J., Lee, J., Fisac, J. F., Simon, J. B., Koppel, J., Zheng, J., Zou, J., Kocon, J., Thompson, J., Wingfield, J., Kaplan, J., Radom, J., Sohl-Dickstein, J., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J., Xu, J., Song, J., Tang, J., Waweru, J., Burden, J., Miller, J., Balis, J. U., Batchelder, J., Berant, J., Frohberg, J., Rozen, J., Hernandez-Orallo, J., Boudeman, J., Guerr, J., Jones, J., Tenenbaum, J. B., Rule, J. S., Chua, J., Kanclerz, K., Livescu, K., Krauth, K., Gopalakrishnan, K., Ignatyeva, K., Markert, K., Dhole, K., Gimpel, K., Omondi, K., Mathewson, K. W., Chiafullo, K., Shkaruta, K., Shridhar, K., McDonell, K., Richardson, K., Reynolds, L., Gao, L., Zhang, L., Dugan, L., Qin, L., Contreras-Ochando, L., Morency, L.-P., Moschella, L., Lam, L., Noble, L., Schmidt, L., He, L., Oliveros-Colón, L., Metz, L., Senel, L. K., Bosma, M., Sap, M., Hoeve, M. T., Farooqi, M., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Ramirez-Quintana, M. J., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M. L., Hagen, M., Schubert, M., Baitemirova, M. O., Arnaud, M., McElrath, M., Yee, M. A., Cohen, M., Gu, M., Ivanitskiy, M., Starritt, M., Strube, M., Swędrowski, M., Bevilacqua, M., Yasunaga, M., Kale, M., Cain, M., Xu, M., Suzgun, M., Walker, M., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., T, M. V., Peng, N., Chi, N. A., Lee, N., Krakover, N. G.-A., Cameron, N., Roberts, N., Doiron, N., Martinez, N., Nangia, N., Deckers, N., Muennighoff, N., Keskar, N. S., Iyer, N. S., Constant,

N., Fiedel, N., Wen, N., Zhang, O., Agha, O., Elbaghdadi, O., Levy, O., Evans, O., Casares, P. A. M., Doshi, P., Fung, P., Liang, P. P., Vicol, P., Alipoormolabashi, P., Liao, P., Liang, P., Chang, P. W., Eckersley, P., Htut, P. M., Hwang, P., Miłkowski, P., Patil, P., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R. E., Gabriel, R., Habacker, R., Risco, R., Millière, R., Garg, R., Barnes, R., Saurous, R. A., Arakawa, R., Raymaekers, R., Frank, R., Sikand, R., Novak, R., Sitelew, R., Bras, R. L., Liu, R., Jacobs, R., Zhang, R., Salakhutdinov, R., Chi, R. A., Lee, S. R., Stovall, R., Teehan, R., Yang, R., Singh, S., Mohammad, S. M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S. R., Schoenholz, S. S., Han, S., Kwatra, S., Rous, S. A., Ghazarian, S., Ghosh, S., Casey, S., Bischoff, S., Gehrmann, S., Schuster, S., Sadeghi, S., Hamdan, S., Zhou, S., Srivastava, S., Shi, S., Singh, S., Asaadi, S., Gu, S. S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Debnath, S. S., Shakeri, S., Thormeyer, S., Melzi, S., Reddy, S., Makini, S. P., Lee, S.-H., Torene, S., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S., Lin, S., Prasad, S., Piantadosi, S., Shieber, S., Misherghi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T., Hashimoto, T., Wu, T.-L., Desbordes, T., Rothschild, T., Phan, T., Wang, T., Nkinyili, T., Schick, T., Kornev, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T., Shaham, U., Misra, V., Demberg, V., Nyamai, V., Raunak, V., Ramasesh, V. V., prabhu, v. u., Padmakumar, V., Srikumar, V., Fedus, W., Saunders, W., Zhang, W., Vossen, W., Ren, X., Tong, X., Zhao, X., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y., Yang, Y., Hao, S., Chen, Y., Belinkov, Y., Hou, Y., Hou, Y., Bai, Y., Seid, Z., Zhao, Z., Wang, Z., Wang, Z. J., Wang, Z., and Wu, Z. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj.

Subramonian, A., Yuan, X., Daumé Iii, H., and Blodgett, S. L. It Takes Two to Tango: Navigating Conceptualizations of NLP Tasks and Measurements of Performance. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 3234–3279, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.202. URL https://aclanthology.org/2023.findings-acl.202.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q., Chi, E., Zhou, D., and Wei, J. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Team, T. T.-B. Terminal-bench: A benchmark for AI agents in terminal environments, Apr 2025. URL https://github.com/laude-institute/terminal-bench.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023. URL https://arxiv.org/abs/2307.09288. Version Number: 2.

Union, E. Article 51: Classification of general-purpose AI models as general-purpose AI models with systemic risk, 2024. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj. [Accessed 12-01-2026].

Vilares, D. and Gómez-Rodríguez, C. HEAD-QA: A Healthcare Dataset for Complex Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 960–966, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1092. URL https://aclanthology.org/P19-1092.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, 2018. URL http://aclweb.org/anthology/W18-5446.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. SuperGLUE: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

Wang, W., Tu, Z., Chen, C., Yuan, Y., Huang, J.-t., Jiao, W., and Lyu, M. All Languages Matter: On the Multilingual Safety of LLMs. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 5865–5877, Bangkok, Thailand and virtual meeting, 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.349. URL https://aclanthology.org/2024.findings-acl.349.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., and Chen, W. MMLU-Pro: a more robust and challenging multi-task language understanding benchmark, 2024b.

Wei, J., Karina, N., Chung, H. W., Jiao, Y. J., Papay, S., Glaese, A., Schulman, J., and Fedus, W. Measuring short-form factuality in large language models, 2024. URL https://arxiv.org/abs/2411.04368. Version Number: 1.

Wei, K., Paskov, P., Dev, S., Byun, M. J., Reuel, A., Roberts-Gaal, X., Calcott, R., Coxon, E., and Deshpande, C. Position: Human baselines in model evaluations need rigor and transparency (With recommendations & reporting checklist). In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J. (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 82265–82325. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/wei25s.html.

White, C., Dooley, S., Roberts, M., Pal, A., Feuer, B., Jain, S., Shwartz-Ziv, R., Jain, N., Saifullah, K., Dey, S., Shubh-Agrawal, Sandha, S., Naidu, S., Hegde, C., LeCun, Y., Goldstein, T., Neiswanger, W., and Goldblum, M. LiveBench: A Challenging, Contamination-Limited LLM Benchmark. In Yue, Y., Garg, A., Peng, N., Sha, F., and Yu, R. (eds.), *International Conference on Learning Representations*, volume 2025, pp. 91595–91631, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/e4a46394ba5378b3f9a186a5b4c650d1-Paper-Conference.pdf.

Xue, Y., Cao, X., Yang, X., Wang, Y., Wang, R., and Li, J. We Need to Talk About Reproducibility in NLP Model Comparison. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9424–9434, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.586. URL https://aclanthology.org/2023.emnlp-main.586.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472.

Zeng, Y., Yang, Y., Zhou, A., Tan, J. Z., Tu, Y., Mai, Y., Klyman, K., Pan, M., Jia, R., Song, D., Liang, P., and Li, B. AIR-Bench 2024: A Safety Benchmark Based on Risk Categories from Regulations and Policies, 2024. URL https://arxiv.org/abs/2407.17436. Version Number: 2.

Zheng, Z., Cheng, Z., Shen, Z., Zhou, S., Liu, K., He, H., Li, D., Wei, S., Hao, H., Yao, J., Sheng, P., Wang, Z., Chai, W., Korolova, A., Henderson, P., Arora, S., Viswanath, P., Shang, J., and Xie, S. LiveCodeBench Pro: How Do Olympiad Medalists Judge LLMs in Competitive Programming? In Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., and Chen, N. (eds.), *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/9712b78386cebdc3db7f1a48c2d20edb-Paper-Datasets_and_Benchmarks_Track.pdf.

Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W., and Duan, N. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2299–2314, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.149. URL https://aclanthology.org/2024.findings-naacl.149.

Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., and Hou, L. Instruction-Following Evaluation for Large Language Models, 2023a. URL https://arxiv.org/abs/2311.07911. Version Number: 1.

Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., Lin, Y., Wen, J.-R., and Han, J. Don't Make Your LLM an Evaluation Benchmark Cheater, 2023b. URL https://arxiv.org/abs/2311.01964. Version Number: 1.

A. Related Work

Benchmark Design & Evolution. The development of AI benchmarks has evolved alongside advances in models, with an increasing focus on broad coverage and rigorous design (Hardy et al., 2024; Liu et al., 2024). Recent efforts emphasize diversity of tasks and continuous updates: for example, the BIG-Bench project crowdsourced hundreds of tasks to test language models' breadth (Srivastava et al., 2023), and dynamic benchmarks like Dynabench introduced adversarial, ongoingly collected test data so that evaluation remains challenging as models improve (Kiela et al., 2021). New benchmark paradigms also expand how performance is measured. The Holistic Evaluation of Language Models initiative, for instance, treats evaluation as a "living" benchmark that is continuously updated and tracks multiple metrics (accuracy, calibration, fairness, etc.) across many scenarios (Liang et al., 2023). Additionally, researchers have proposed principled frameworks for benchmark construction to ensure that datasets, tasks, and metrics truly capture the targeted capabilities (Salaudeen et al., 2025; Liu et al., 2024; Subramonian et al., 2023; Raji et al., 2021).

Issues in AI Evaluations. Despite continual benchmark innovation, significant challenges persist in how we evaluate AI systems. Various works have highlighted fundamental evaluation pitfalls in AI evaluation: Data contamination, i.e., when test content appears in training, can artificially inflate scores. For example, Schaeffer (2023) demonstrated that directly fine-tuning on a test set yields nearly perfect accuracy. Gamability of benchmarks is another concern: models often exploit spurious correlations or annotation artifacts to get high accuracy without genuine understanding. For instance, McCoy et al. (2019) have shown that models rely on superficial cues (*e.g.,* lexical overlap or keyword hints) instead of robust reasoning, achieving "right for the wrong reason" performance that fails on stress tests. Finally, reproducibility remains a challenge in AI evaluation: Seemingly superior results frequently vanish under minor experimental changes (*i.e.,* simply altering random seeds or dataset splits can yield statistically significant performance fluctuations and inconsistent evaluation protocols or opaque reporting have further complicated fair comparison of models (Xue et al., 2023; Habba et al., 2025; Ashury-Tahan et al., 2026b)). In parallel to this work Ashury-Tahan et al. (2026a), shows that this brittleness is highly reduced with saturation. Finally, Ott et al. (2022) show that benchmark saturation is a common occurrence, potentially making them misleading indicators of progress once models overfit to test quirks rather than achieve substantive gains. Yet, Ott et al. (2022) neither quantitatively define benchmark saturation nor do the authors analyze the causes of such saturation, two gaps we fill in this work. Relatedly, there is increasing awareness that aggregate metrics such as accuracy, F1, or single-scale scores often fail to capture nuanced model behavior (Murahari et al., 2024). This coarseness can create a misleading sense of benchmark saturation – models may reach near-ceiling aggregate scores while still exhibit substantial variation across subskills or input types. This apparent saturation, driven by the insensitivity of aggregate metrics, obscures remaining weaknesses and limits the diagnostic value of benchmarks. Consequently, several works advocate for holistic evaluation frameworks, including hybrid scoring schemes (Liang et al., 2023) or even a new discipline of model metrology to formalize rigorous, fine-grained measurement practices (Saxon et al., 2024).

B. Hypotheses

We investigate five hypotheses about factors driving benchmark saturation, grounded in prior literature and design challenges. To support this analysis, we annotated 60 LLM benchmarks with related properties such as task format, data curation, and known quality issues. These annotations, detailed in Sec. 3, enable empirical testing of the hypotheses (Sec. 4).

(H1) Data Access and Test Set Exposure: *Public benchmarks saturate faster than private benchmarks with held-out test sets.* When test questions are public, models often memorize or leak this content from their training corpora, yielding artificially high scores that do not reflect true generalization: Zhou et al. (2023b) demonstrate that if an LLM's pre-training data contains examples from an evaluation benchmark, the model's score on that benchmark is significantly boosted. Likewise, Balloccu et al. (2024) conducted a large-scale analysis of GPT-3.5 and GPT-4 and found they were exposed to approximately 4.7 million benchmark samples during training, which may explain why these models quickly achieve near-perfect scores on popular public tests. Deng et al. (2024) devised a protocol to probe contamination on knowledge benchmarks and found that GPT-4 and Claude could fill in missing parts of real test questions with unnaturally high accuracy, implying the models had internalized those test items. These findings support H1: because public benchmarks are easily scraped or overfit, top model scores on them often reflect memorization.

(H2) Language Coverage: *English-only benchmarks saturate faster than multilingual or mixed-language benchmarks.* English dominates the pre-training corpora of most models (often >85–90% of tokens) (Touvron et al., 2023; OpenAI, 2020). Wang et al. (2024a) observe that an LLM's ranking across languages correlates strongly with the proportion of that language

in its training data, where models consistently excel at English and a few other high-resource languages, but struggle as one moves to less-seen languages. Consequently, an English-only task can hit a performance ceiling quicker because the model's exposure to English makes the task easier in distribution. In contrast, a multilingual benchmark introduces linguistic diversity that challenges the model's weaker languages and forces more robust generalization (Hu et al., 2020).

(H3) Data Curation Strategy: *Human-authored benchmarks are more resistant to performance saturation than synthetic or hybrid ones.* Human-curated evaluations typically span a richer diversity of problems and deeper conceptual challenges, often including intentionally difficult or adversarially crafted questions that thwart simple pattern-matching: Das et al. (2024) found that LLM outputs risk repetitive formats and missing corner-case reasoning. Diversity and deliberate complexity introduced by humans make it harder for models to "solve" benchmark tasks by exploiting superficial regularities (Gill et al., 2025). By contrast, LLM-generated (synthetic) benchmarks tend to exhibit hidden structural patterns or stylistic biases that models quickly learn to exploit, yielding artificially high scores without commensurate gains in real capability (Gill et al., 2025).

(H4) Task Output Format: *Benchmarks that use a closed-ended response format (e.g. multiple-choice, true/false) tend to saturate faster than those requiring open-ended generation.* Closed-ended tasks constrain the output space, making it easier for models to guess or recognize the correct answer without full understanding. The underlying mechanism is that closed formats convert complex tasks into simpler classification problems: the model's job is reduced to selecting one of N options, a setup amenable to elimination strategies, frequency biases, or even memorized question-option pairs. Moreover, closed-ended benchmarks typically have an inherent guessing baseline (e.g. 25% for 4-choice questions), so even an uninformed model starts at a higher performance floor. Recent work demonstrated that some MCQ benchmarks enable overfitting of models such that they pick the right option without even reading the question (Chandak et al., 2025). By contrast, open-ended prompts (where the model must generate a free-form answer, explanation, or output) vastly expand the solution space and typically require a deeper grasp of the problem.

(H5) Benchmark Maturity and Popularity: *Benchmarks that are older and more widely adopted saturate faster than newer or less-used benchmarks.* As benchmarks mature and become widely adopted by the research community, they are repeatedly used for model development, hyperparameter tuning, prompt engineering, and evaluation, increasing optimization pressure against the benchmark itself. Prior work has noted that performance on popular benchmarks often improves rapidly shortly after release and then plateaus as models converge on similar solutions (Ott et al., 2022). Moreover, widely adopted benchmarks are more likely to be included—directly or indirectly—in training data or evaluation pipelines, further accelerating score convergence. As a result, benchmark age and popularity may jointly contribute to saturation by increasing exposure and targeted optimization, even when absolute task difficulty remains unchanged.

(H6) Template vs Non-Template: *Non-templated benchmarks are more resistant to performance saturation than templated benchmarks.* Templated benchmarks generate data samples using predefined patterns, structures, or parameterized templates, often resulting in repeated surface forms with limited variation. While such designs enable scalability and controlled coverage, they can introduce regularities that models quickly learn to exploit. In contrast, non-templated benchmarks consist of more diverse, free-form instances that are less constrained by fixed generation patterns. We therefore hypothesize that templated benchmarks, due to their structural regularities and reduced diversity, are more prone to faster saturation compared to non-templated, free-form benchmarks.

C. Semantic Scholar Benchmark Collection

We retrieved all benchmarks appearing in the most-cited research papers between 2022 and November 2025 using Semantic Scholar API and the following queries (50 per keyword): `language model evaluation`, `LLM benchmark`, `foundation model benchmark`, `language model benchmark` and `language model evaluation benchmark`. The Semantic Scholar API retrieves the most relevant papers within a given time period. We first retrieve 200 relevant papers per keyword and select the top 50 cited papers. After merging all retrieved papers and deduplicating, we identified 186 papers using these keywords. We excluded non-text-based benchmarks. This keyword-based search yielded 2 additional benchmarks that were previously absent from our collection.

Table 3. Benchmarks included in our analysis (N=60)

Benchmark	Reference	Benchmark	Reference
AGIEval	(Zhong et al., 2024)	MGSM	(Shi et al., 2023)
AIME 2025	–	MMLU	(Hendrycks et al., 2020)
AIR-Bench	(Zeng et al., 2024)	MMLU-Indic	(AI, 2024b)
ANLI	(Nie et al., 2020)	MMLU-Pro	(Wang et al., 2024b)
ARC-AGI	(Chollet, 2019; Chollet et al., 2025)	MMLU-Redux	(Gema et al., 2025)
Arena-Hard	(Li∗et al., 2024)	MMMLU	(OpenAI, 2024)
Belebele	(Bandarkar et al., 2024)	MultiPL-E	(Cassano et al., 2022)
BIG-Bench Hard	(Suzgun et al., 2023)	Natural Questions	(Kwiatkowski et al., 2019)
BoolQ	(Clark et al., 2019)	OpenBookQA	(Mihaylov et al., 2018)
C-Eval	(Huang et al., 2023)	PIQA	(Bisk et al., 2020)
CommonsenseQA	(Talmor et al., 2019)	QuAC	(Choi et al., 2018)
DROP	(Dua et al., 2019)	QuALITY	(Pang et al., 2022)
FACTS Grounding	(Jacovi et al., 2024; 2025)	RACE	(Lai et al., 2017)
Flores-101	(Goyal et al., 2022)	RewardBench	(Lambert et al., 2025)
Global-MMLU	(Singh et al., 2025)	SIB-200	(Adelani et al., 2024)
GLUE	(Wang et al., 2018)	SimpleQA	(Haas et al., 2025; Wei et al., 2024)
GPQA	(Rein et al., 2023)	SIQA	(Sap et al., 2019)
MedQA	(Jin et al., 2021)	SQuAD v2	(Rajpurkar et al., 2016; 2018)
GSM8K	(Cobbe et al., 2021)	SuperGLUE	(Wang et al., 2019)
GSM8K-Indic	(AI, 2024a)	SWE-bench	(Jimenez et al., 2024)
HEAD-QA	(Vilares & Gómez-Rodríguez, 2019)	τ-Bench	(Barres et al., 2025)
HellaSwag	(Zellers et al., 2019)	TerminalBench	(Team, 2025)
HumanEval	(Chen et al., 2021)	TerminalBench 2.0	(Merrill et al., 2026)
Humanity's Last Exam	(Phan et al., 2025)	TriviaQA	(Joshi et al., 2017)
IFEval	(Zhou et al., 2023a)	TruthfulQA	(Lin et al., 2022)
LAMBADA	(Paperno et al., 2016)	TyDiQA	(Clark et al., 2020)
LegalBench	(Guha et al., 2023)	MCLM	(Son et al., 2025)
LiveBench	(White et al., 2025)	Winogrande	(Sakaguchi et al., 2020)
LiveCodeBench	(Zheng et al., 2025; Jain et al., 2025)	WMT	(Kocmi et al., 2022; 2023; 2024)
MATH-500	(Hendrycks et al., 2021)	FrontierMath	(Glazer et al., 2024)

D. Field Definitions for Annotation and Examples

This appendix provides detailed tables describing the annotation schema and benchmark metadata used in our analysis (Table 4), along with example rows illustrating the collected saturation metrics (Table 6) and dataset properties (Table 5).

Table 4. Benchmark Annotation Schema. Each benchmark in our dataset is annotated with the following fields to enable systematic analysis of saturation dynamics.

Field	Description
Identification & Temporal	
Benchmark	Name of the benchmark being analyzed
Released On	Publication date of benchmark paper or public release on platforms like HuggingFace/GitHub
Citations	Citation count for the benchmark paper
Saturation Measurement	
Saturation Metadata	Top-5 model names and scores used to determine saturation status
Recent Models Evaluated	Whether frontier models released in 2025 (e.g., Gemini-2.5, Qwen-3) have been evaluated
SOTA in Paper	Best-performing model and score reported in the original benchmark paper
Data Quality	
Dataset Issues	Known post-release issues: contamination, biases, mislabeling/noise, or other data problems
Issue Sources	Follow-up papers or reports documenting identified dataset issues
Task Structure	
Input Format	Task input type: QA (MCQ), Instruction (open-ended), Coding (unit-test evaluated), or Agentic (multi-turn/tool-use)
Output Format	Expected response format: MCQ (select option) or Free-form (open generation)
Metric	Primary evaluation metric: Accuracy, BLEU, LLM-as-judge, or task-specific
Dataset Properties	
Curation Method	How data was created: expert human, crowdsourced, LLM-generated, or programmatically scraped
Curation Notes	Additional details on data collection methodology
Languages	Languages included in the benchmark
Sample Count	Number of evaluation examples in the benchmark
Availability	Whether benchmark and ground-truth labels are publicly accessible
Literal Diversity	Whether prompts use templated structures (e.g., "What is the capital of ___?") vs. natural variation

Table 5. Benchmark Dataset Properties (Example Rows)

Benchmark	Input	Output	Curation	Lang.	Samples	Citations	Avail.	Templated
Math-500	Instruction	Free-form	Expert human	EN	500	2398	Public	No
GPQA Diamond	QA/MCQ	MCQ	Expert human	EN	564	1180	Public	No

Table 6. Benchmark Saturation Analysis (Example Rows)

Benchmark	Released	SOTA (Paper)	Recent Eval	Top-5 Models & Scores	Issues
Math-500	Mar 2021	6.9 (GPT-2)	Yes	o3: 99.2; Grok-4: 99.0; DeepSeek R1: 98.3; GLM-4.5: 98.2; Claude Opus 4: 98.2	Contam.
GPQA Diamond	Nov 2023	38.8 (GPT-4)	Yes	Grok4: 87.7; GPT-5: 85.4; Gemini-2.5: 84.4; Claude-4.5: 83.4; GLM 4.6: 82.9	None

E. Benchmark-Level Saturation - Overview and Case Studies

To complement our analysis, we provide benchmark-level case studies in Table 7 illustrating how the saturation index behaves across different benchmarks. These examples highlight how score compression, evaluation uncertainty, and dataset properties jointly determine whether a benchmark is saturated, stagnated, or remains discriminative. These examples show a

range of saturation indices from fully saturated benchmarks, where evaluation noise obscures all meaningful differences, to unsaturated benchmarks that retain strong discriminative power.

Math-500 (very high saturation, $S_{\text{index}} = 0.92$). The Math-500 leaderboard shows that top-performing models are tightly clustered within a 1.0-point range (98.2–99.2), which lies within the estimated evaluation uncertainty ($SE_{\Delta} = 0.0338$). This results in a low normalized range ($R_{\text{norm}} = 0.30$), indicating that performance differences are not statistically meaningful and the benchmark has lost discriminative power.

LiveBench (very high saturation, $S_{\text{index}} = 0.99$). Although designed to mitigate contamination through regular updates, LiveBench shows high score compression (range = 1.09) relative to its uncertainty ($SE_{\Delta} = 0.1028$), resulting a very low $R_{\text{norm}} = 0.11$. Notably, this occurs at moderate performance levels (79%), suggesting model-level stagnation rather than task completion.

LiveCodeBench (high saturation, $S_{\text{index}} = 0.77$). LiveCodeBench shows stronger separation among top models (performance range = 3.9), which results in a higher normalized range ($R_{\text{norm}} = 0.51$). While still showing score compression, it demonstrates that also dynamically constructed benchmarks can saturate when evaluation resolution is limited.

TruthfulQA (moderate saturation, $S_{\text{index}} = 0.55$). The TruthfulQA leaderbaord shows a wider spread among top models (range = 6.7), which exceeds evaluation uncertainty. This leads to meaningful differentiation ($R_{\text{norm}} = 0.78$), but partial clustering indicates early signs of convergence, which is consistent with the benchmarks age and exposure.

Humanity's Last Exam (low saturation, $S_{\text{index}} = 0.22$). This benchmark shows substantial separation among top models (range = 11.4), which exceeds uncertainty ($R_{\text{norm}} = 1.23$). Combined with its large test set and recent release, it retains strong discriminative power and shows now clear sign of saturation.

Table 7. Representative benchmarks illustrating different saturation regimes.

Benchmark	n	Range	SE_{Δ}	R_{norm}	S_{index}	Level
Math-500	500	1.00	0.0338	0.2955	0.9164	Very high
LiveBench	1000	1.09	0.1028	0.1060	0.9888	Very high
LiveCodeBench	1000	3.90	0.0761	0.5124	0.7691	High
TruthfulQA	817	6.70	0.0863	0.7766	0.5471	Moderate
Humanity's Last Exam	2500	11.40	0.0926	1.2309	0.2198	Low

F. Further Saturation Analysis

This appendix presents additional results from the joint regression analysis, including posterior coefficient estimates (Figure 5) and model performance, to provide a more detailed view of the factors associated with benchmark saturation (Figure 6).

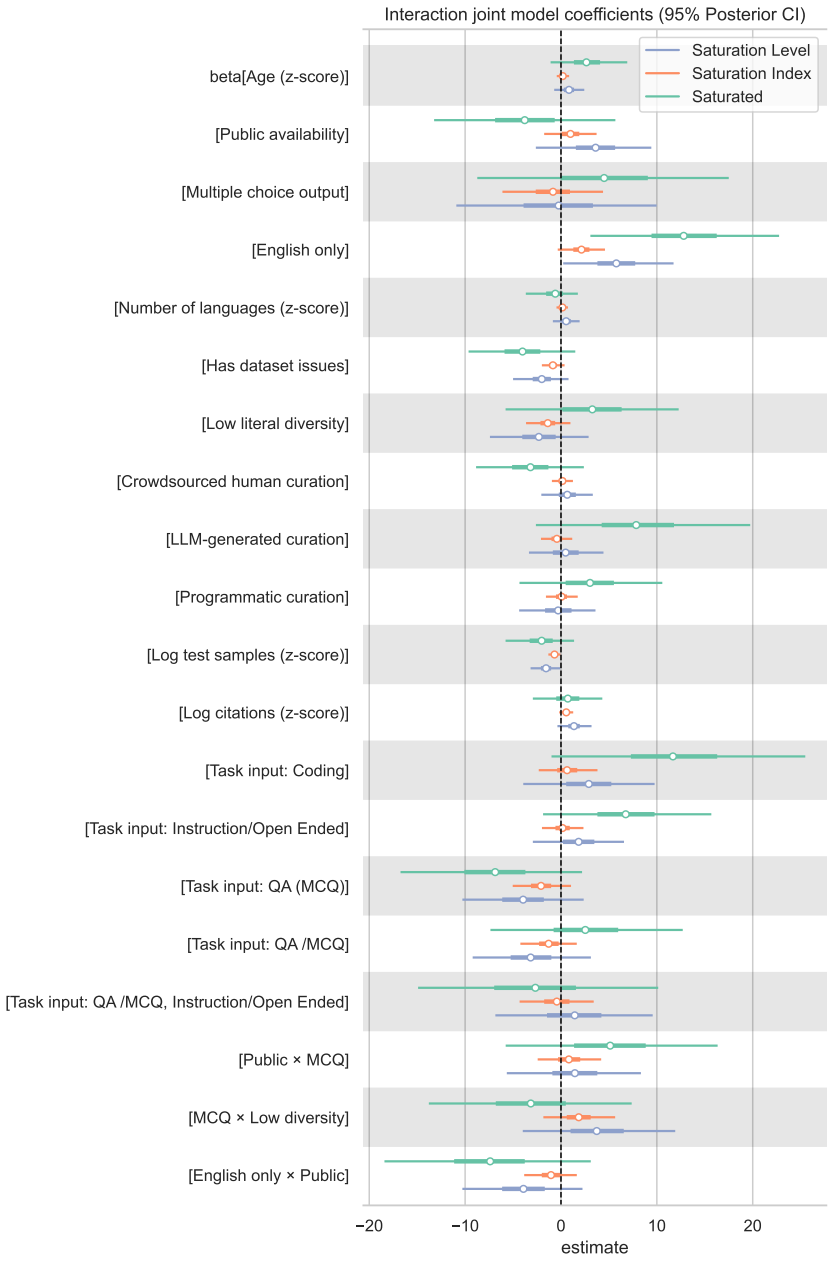

Figure 5. Forest plot of posterior regression coefficients from the joint interaction model predicting benchmark saturation. Points denote posterior means, inner line segments show 50% highest posterior density intervals, and outer segments indicate 95% credible intervals. Benchmark age and test set size exhibit the most consistent effects on saturation, while task format, literal diversity (templating), and their interactions show no strong effects after controlling for confounders.

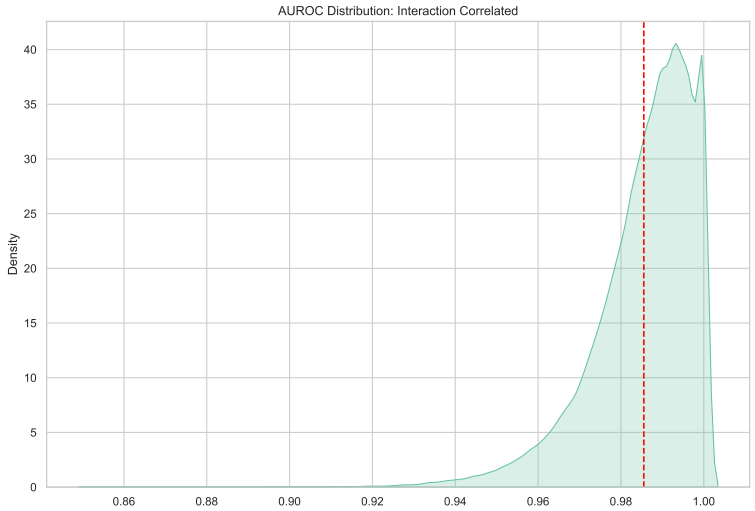

Figure 6. Posterior distribution of the AUROC for the interaction model predicting benchmark saturation. The distribution is tightly concentrated near high values (median approx. 0.98), indicating that the model distinguishes saturated from non-saturated benchmarks across posterior samples.

G. Author Contribution Statement

Conceptualization	M. Akhtar, A. Reuel
Data Curation	M. Akhtar, P. Soni, S. Ahuja, P. Ammanamanchi, R. Rawal, S. Yadav, C. Whitehouse, D. Ki, J. Mickel, M. Šuppa, J. Batzner, J. Chim, J. Sania, Y. Long, H. Rahmani, C. Knight, Y. Nan, J. Raj, Y. Fan, S. Singh, S. Sahoo, E. Habba, S. Pawar, R. Scholz, A. Subramanian, J. Ni, L. Struppek, A. Ghosh
Investigation	M. Akhtar, P. Soni, S. Ahuja, P. Ammanamanchi, R. Rawal, V. Zouhar, S. Yadav, C. Whitehouse, D. Ki, L. Ibrahim, J. Raj, Y. Fan, L. Struppek, U. Gohar, J. Mickel
Methodology	M. Akhtar, A. Reuel, P. Soni, S. Ahuja, P. Ammanamanchi, R. Rawal, V. Zouhar, S. Yadav, C. Whitehouse, D. Ki
Software	M. Akhtar, P. Soni, S. Ahuja, P. Ammanamanchi, R. Rawal, V. Zouhar, S. Yadav, C. Whitehouse, D. Ki, J. Raj, M. Šuppa
Formal Analysis	M. Akhtar, P. Soni, S. Ahuja, P. Ammanamanchi, R. Rawal, V. Zouhar, S. Yadav, C. Whitehouse, D. Ki, Y. Long, J. Chim, J. Sania, M. Šuppa, Y. Nan
Writing (Original Draft)	M. Akhtar, A. Reuel, P. Soni, S. Ahuja, P. Ammanamanchi, R. Rawal, V. Zouhar, S. Yadav, C. Whitehouse, D. Ki, R. Scholz, L. Ibrahim, Y. Fan
Writing (Review & Editing)	M. Akhtar, A. Reuel, P. Soni, S. Ahuja, P. Ammanamanchi, R. Rawal, V. Zouhar, S. Yadav, C. Whitehouse, D. Ki, J. Mickel, L. Choshen, M. Šuppa, J. Batzner, J. Chim, J. Sania, Y. Long, H. Rahmani, C. Knight, Y. Nan, J. Raj, Y. Fan, S. Singh, S. Sahoo, E. Habba, U. Gohar, S. Pawar, R. Scholz, A. Subramanian, J. Ni, L. Struppek, L. Ibrahim, M. Kochenderfer, S. Koyejo, M. Sachan, S. Biderman, Z. Talat, A. Ghosh, I. Solaiman
Visualization	M. Akhtar, J. Raj, V. Zouhar, M. Šuppa, C. Whitehouse
Supervision	M. Akhtar, A. Reuel, L. Choshen, M. Kochenderfer, S. Koyejo, M. Sachan, S. Biderman, Z. Talat, A. Ghosh, I. Solaiman