Benchmark Saturation: A Statistical Analysis of Score Compression Across LLM Leaderboards

Abel Yagubyan
Independent Researcher
abel.yagubyan@gmail.com

February 2026

Abstract

As large language models improve, benchmark score distributions compress toward ceilings, diminishing their ability to meaningfully distinguish between models. While benchmark saturation is widely acknowledged, no systematic quantification exists. We analyze 11,836 model evaluations across 12 benchmarks spanning two generations of the HuggingFace Open LLM Leaderboard (July 2023–March 2025). We introduce the **Benchmark Discriminability Index (BDI)**, an entropy-based metric that quantifies a benchmark's ability to differentiate models over time. We find that (1) V1 benchmarks (HellaSwag, Winogrande, ARC) lost 15–24% of their discriminability within 10 months, with top-10 score gaps compressing to <1 point; (2) logistic saturation curves predict benchmark ceilings for actively saturating benchmarks ($R^2 = 0.99$ for GSM8K); (3) V2 replacement benchmarks (GPQA, MMLU-PRO) retain high discriminability, though IFEval already approaches the saturation zone (ceiling proximity = 0.90); and (4) knowledge-retrieval benchmarks saturate faster than reasoning benchmarks. We provide a saturation timeline for all 12 benchmarks and propose BDI monitoring as a principled criterion for benchmark retirement.

1 Introduction

The evaluation of large language models (LLMs) relies heavily on standardized benchmarks—curated test sets that measure capabilities ranging from factual knowledge to mathematical reasoning. As models improve rapidly, a recurring concern in the community is *benchmark saturation*: the phenomenon where state-of-the-art models achieve near-perfect scores, rendering the benchmark unable to differentiate between them [Kiela et al., 2021].

This concern has driven practical decisions. In June 2024, the HuggingFace Open LLM Leaderboard—the most widely used public evaluation platform for open-weight models—retired its original benchmark suite (ARC, HellaSwag, MMLU, TruthfulQA, Winogrande, GSM8K) and replaced it with harder alternatives (IFEval, BBH, MATH Lvl 5, GPQA, MUSR, MMLU-PRO). The motivation was explicitly that the original benchmarks "were no longer able to differentiate between models" [Hugging Face, 2024].

Despite this widespread recognition, benchmark saturation has not been rigorously quantified. When exactly does a benchmark become saturated? How fast does discriminability decay? Can we predict when a new benchmark will meet the same fate? And critically, is there a principled metric—beyond informal judgment—for determining when a benchmark should be retired?

We address these questions through a large-scale statistical analysis of the Open LLM Leaderboard, spanning both its V1 (July 2023–June 2024) and V2 (June 2024–March 2025) eras. Our

contributions are:

1. We introduce the **Benchmark Discriminability Index (BDI)**, an entropy-based metric that quantifies a benchmark's ability to distinguish models at any point in time (§4).
2. We fit **logistic saturation curves** to maximum score trajectories, enabling prediction of benchmark ceilings and time-to-saturation (§5).
3. We provide the first systematic **saturation timeline** across 12 benchmarks spanning two leaderboard generations, showing that knowledge-retrieval benchmarks saturate faster than reasoning benchmarks (§6).
4. We propose **BDI monitoring** as an actionable criterion for benchmark retirement decisions (§7).

2 Related Work

Benchmark criticism. Concerns about benchmark validity predate LLMs [Bowman and Dahl, 2021]. Kiela et al. [2021] advocated for dynamic benchmarks that adapt to model capabilities. Recht et al. [2019] demonstrated that top-line accuracy improvements on ImageNet did not transfer to new test sets, suggesting benchmark-specific overfitting. More recently, Anonymous [2026a] found systematic disagreements between LLM benchmark rankings and downstream task performance, terming this the "benchmark illusion."

LLM evaluation methodology. The rapid pace of LLM development has strained evaluation practices. Liang et al. [2023] proposed HELM as a comprehensive evaluation framework. Anonymous [2026c] argued for transitioning from model-centric to agent-centric evaluation. NIST [2026] proposed statistical models for more robust AI evaluation. Anonymous [2026b] developed methods for statistically efficient LLM evaluation that reduce the number of test examples needed.

Score compression and ceiling effects. Ceiling effects are well-studied in psychometrics [Uttl, 2005] and have been noted anecdotally in NLP benchmarks. Anonymous [2026d] showed that trivial lexical and syntactic variations in evaluation prompts cause significant score fluctuations, suggesting that small score differences at the ceiling may be noise rather than signal. Anonymous [2024] quantified uncertainty in LLM benchmark scores, finding that confidence intervals often overlap between top-ranked models.

Our work differs from prior criticism in that we *quantify* saturation dynamics empirically, propose a formal metric for discriminability, and provide predictive models for benchmark shelf life.

3 Data

We analyze model evaluation results from both generations of the HuggingFace Open LLM Leaderboard:

V1 (July 2023 – June 2024). 7,260 model evaluations across 6 benchmarks: ARC Challenge [Clark et al., 2018], HellaSwag [Zellers et al., 2019], MMLU [Hendrycks et al., 2021b], TruthfulQA [Lin et al., 2022], Winogrande [Sakaguchi et al., 2021], and GSM8K [Cobbe et al., 2021]. Data sourced from the archived `open-llm-leaderboard-old/contents` dataset.

V2 (June 2024 – March 2025). 4,576 model evaluations across 6 benchmarks: IFEval [Zhou et al., 2023], BBH [Suzgun et al., 2023], MATH Lvl 5 [Hendrycks et al., 2021a], GPQA [Rein et al., 2024], MUSR [Sprague et al., 2024], and MMLU-PRO [Wang et al., 2024]. Data sourced from `open-llm-leaderboard/contents`.

Both datasets include submission timestamps, enabling temporal analysis. We aggregate scores into monthly windows, requiring a minimum of 10 model evaluations per window for statistical validity.

4 Metrics

4.1 Benchmark Discriminability Index (BDI)

We define the BDI as the normalized Shannon entropy of the score distribution at time t:

$$\text{BDI}(t) = \frac{H(\mathbf{p}_t)}{H_{\max}} = \frac{-\sum_{i=1}^{B} p_i(t) \log_2 p_i(t)}{\log_2 B} \tag{1}$$

where $\mathbf{p}_t$ is the probability distribution obtained by binning model scores into $B = 20$ equal-width bins over $[0, 100]$, and $H_{\max} = \log_2 B$ is the maximum entropy (uniform distribution).

Interpretation: BDI $\in [0, 1]$. A BDI of 1 indicates maximum discriminability (scores uniformly spread across the range). A BDI approaching 0 indicates all models scoring in the same bin (complete saturation). In practice, we observe BDI values between 0.3 and 0.9.

Sensitivity: BDI depends on the number of bins B. We tested $B \in \{5, 10, 15, 20, 25, 30, 50\}$ and found that while absolute BDI values shift (± 0.07 across bin choices), the *relative ordering* and *temporal trends* remain stable. We use $B = 20$ throughout as a balance between resolution and stability. All reported BDI changes are significant at the 95% level via bootstrap confidence intervals (1,000 resamples).

4.2 Top-K Gap

As a complementary measure focused on the competitive frontier, we compute the mean pairwise gap between adjacent models in the top-K ranking:

$$\text{Gap}_K(t) = \frac{1}{K-1} \sum_{i=1}^{K-1} (s_{(i)}(t) - s_{(i+1)}(t)) \tag{2}$$

where $s_{(i)}(t)$ is the i-th highest score at time t. We report $K = 10$ and $K = 20$.

4.3 Ceiling Proximity

We define ceiling proximity as the ratio of the maximum observed score to the theoretical maximum:

$$\text{CP}(t) = \frac{\max_m s_m(t)}{100} \tag{3}$$

While simple, CP does not capture distributional compression—a benchmark can have high CP but still discriminate well if the score distribution retains variance.

Figure 1: Maximum Score Trajectories Over Time

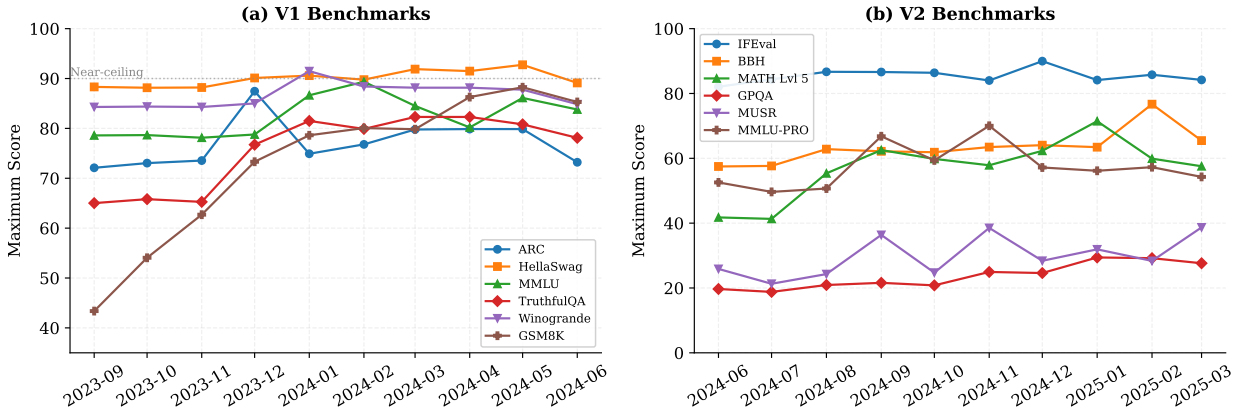

Figure 1: Maximum score trajectories for V1 (left) and V2 (right) benchmarks. V1 benchmarks cluster near the ceiling (>85) by early 2024, while V2 benchmarks show greater spread and headroom.

5 Saturation Curve Analysis

We model the trajectory of maximum scores over time using a logistic growth function:

$$s_{\max}(t) = \frac{L}{1 + e^{-k(t-t_0)}} \tag{4}$$

where L is the projected ceiling, k is the growth rate, and t_0 is the inflection point. We fit this curve using nonlinear least squares (scipy `curve_fit`) with bounds $L \in [\max(s) - 1, 100]$, $k \in [0.01, 10]$.

From the fitted parameters, we derive:

- **Projected ceiling** (L): the estimated maximum achievable score.
- **Time to 90% ceiling** ($t_{90} = t_0 + \frac{\ln 9}{k}$): when models reach 90% of the ceiling.
- **Remaining headroom**: $L - \max(s_{\text{observed}})$.

Fit quality caveat. The logistic model fits well for benchmarks undergoing active saturation (GSM8K: $R^2 = 0.986$; TruthfulQA: $R^2 = 0.804$; GPQA: $R^2 = 0.860$) but poorly for benchmarks that were already near their ceiling at the start of tracking (HellaSwag: $R^2 = 0.243$; Winogrande: $R^2 = -0.300$). This is expected—when only the plateau of the logistic is observed, the fit is underdetermined. We report ceiling projections only for benchmarks with $R^2 > 0.5$.

6 Results

6.1 V1 Benchmarks: Saturation Confirmed

Table 1 summarizes the saturation status of V1 benchmarks (see also Figure 1a for score trajectories and Figure 2a for BDI trends). All six benchmarks show ceiling proximity above 73%, with HellaSwag (92.8%) and Winogrande (91.5%) most saturated.

Figure 2: Benchmark Discriminability Index Over Time

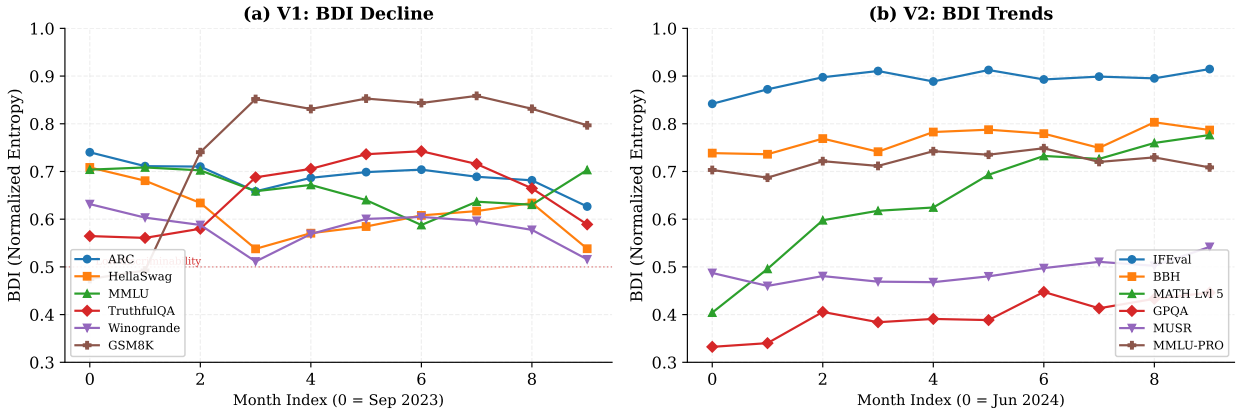

Figure 2: Benchmark Discriminability Index over time. V1 benchmarks (left) show declining BDI, indicating loss of discriminative power. V2 benchmarks (right) show stable or increasing BDI, with GPQA and MUSR retaining the most room for growth.

Table 1: V1 benchmark saturation summary (July 2023 – June 2024).

Benchmark	Max Score	CP	$\text{BDI}_{\text{first}}$	BDI_{last}	ΔBDI	Ceiling (L)	R^2
ARC	87.5	0.88	0.740	0.627	−0.113	—	−0.52
HellaSwag	92.8	0.93	0.708	0.538	−0.170	—	0.24
MMLU	89.4	0.89	0.704	0.703	−0.001	—	0.19
TruthfulQA	82.3	0.82	0.565	0.589	+0.024	82.8	0.80
Winogrande	91.5	0.92	0.632	0.516	−0.116	—	−0.30
GSM8K	88.2	0.88	0.475	0.797	+0.322	87.2	0.99

Finding 1: HellaSwag and Winogrande were already saturated. Both benchmarks had top-10 gaps below 1 point from the earliest observations (Figure 5a). HellaSwag showed a BDI decline of 0.170 (24% relative) over 10 months.

Finding 2: GSM8K was the exception. Unlike other V1 benchmarks, GSM8K showed *increasing* discriminability ($\Delta\text{BDI} = +0.322$). This is because models entered the mid-range rapidly—the max score rose from 43.4 to 88.2 over 10 months, with a near-perfect logistic fit ($R^2 = 0.986$, Figure 3). GSM8K was the last V1 benchmark to approach saturation, consistent with reasoning tasks being harder to saturate than knowledge-retrieval tasks.

6.2 V2 Benchmarks: Early Signs

Finding 3: V2 benchmarks retain high discriminability. All V2 benchmarks show stable or increasing BDI (Figure 2b), indicating they are still in the "useful" phase. GPQA (CP = 0.29) and MUSR (CP = 0.39) have the most remaining headroom (Figure 4).

Finding 4: IFEval shows early saturation risk. Despite being a V2 benchmark, IFEval already has CP = 0.90, comparable to late-stage V1 benchmarks. Its logistic fit projects a ceiling of 89.0, suggesting it may be the first V2 benchmark to require retirement.

Finding 5: MATH Lvl 5 follows the GSM8K pattern. Like GSM8K, MATH Lvl 5 shows rapidly increasing BDI (+0.372) as models populate the mid-range. The logistic fit projects

Figure 3: Logistic Saturation Curves (V1 Benchmarks)

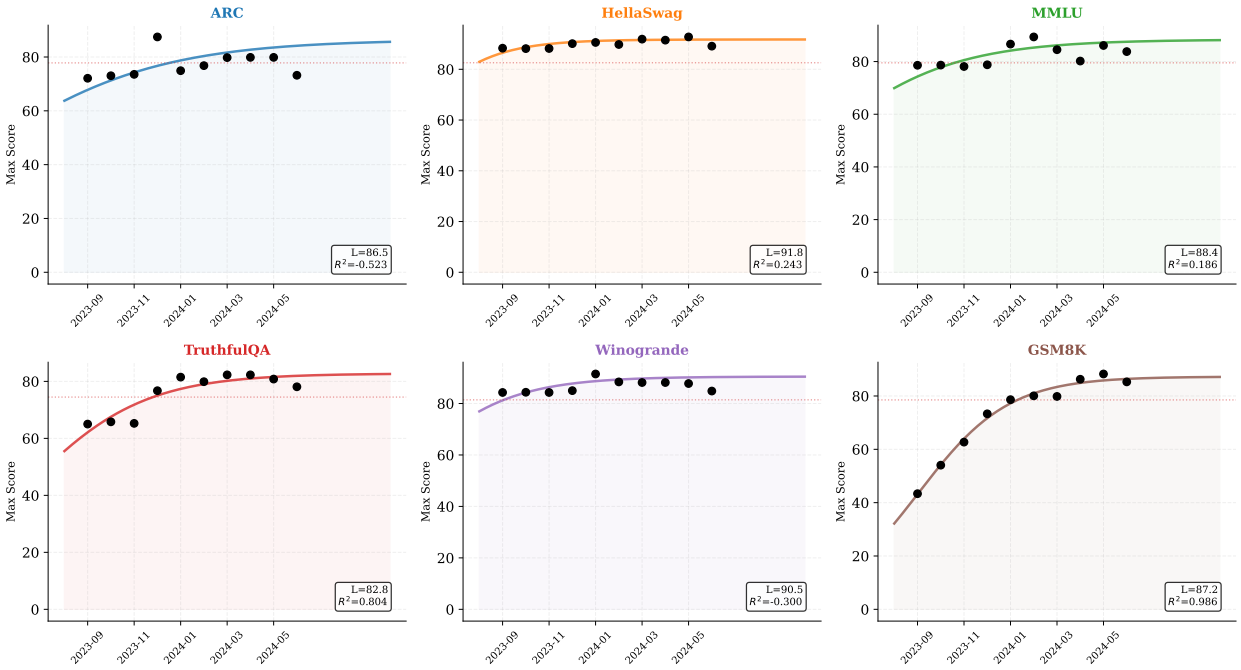

Figure 3: Logistic saturation curve fits for V1 benchmarks. GSM8K exhibits a near-perfect logistic trajectory ($R^2 = 0.986$), while HellaSwag and Winogrande were already near their ceilings at the start of tracking. L denotes the projected ceiling.

a ceiling of 70.5, which the observed maximum (71.5) already exceeds—indicating rapid saturation that outpaces even our curve-fitting projections.

6.3 Knowledge vs. Reasoning Benchmarks

Across both leaderboard generations, we observe a consistent pattern: **knowledge-retrieval benchmarks saturate faster than reasoning benchmarks.**

V1 knowledge benchmarks (ARC, HellaSwag, MMLU, Winogrande) showed ceiling proximity above 85% from the earliest observations, while GSM8K—the sole V1 reasoning benchmark—started at 43% and took 7+ months to approach saturation. In V2, the pattern repeats: IFEval (instruction following, largely format compliance) is approaching saturation at CP = 0.90, while GPQA (graduate-level reasoning, CP = 0.29) and BBH (compositional reasoning, CP = 0.77) retain substantial headroom.

We note that this comparison is limited by the small number of reasoning-focused benchmarks in V1 (only GSM8K). However, the consistency across both leaderboard generations supports the intuition that knowledge can be memorized from training data, while reasoning capabilities improve more gradually.

7 Discussion

7.1 When Should a Benchmark Be Retired?

We propose a three-criteria retirement test based on our metrics:

Table 2: V2 benchmark status (June 2024 – March 2025).

Benchmark	Max Score	CP	$\text{BDI}_{\text{first}}$	BDI_{last}	ΔBDI	Ceiling (L)	R^2
IFEval	90.0	0.90	0.842	0.915	+0.073	—	−0.76
BBH	76.7	0.77	0.739	0.787	+0.048	>100	0.55
MATH Lvl 5	71.5	0.72	0.404	0.776	+0.372	70.5†	0.62
GPQA	29.4	0.29	0.332	0.446	+0.114	70.5	0.86
MUSR	38.7	0.39	0.487	0.542	+0.055	37.7†	0.33
MMLU-PRO	70.0	0.70	0.703	0.709	+0.005	—	0.03

†Observed maximum already exceeds projected ceiling, indicating the logistic model underestimates the true ceiling for these benchmarks.

1. **Ceiling proximity** > 0.90: maximum scores are within 10% of the theoretical ceiling.
2. **Top-10 gap** < 1.0: the top models are indistinguishable within noise margins.
3. **BDI decline** $> 15\%$ from peak: the benchmark has lost significant discriminative power.

When all three criteria are met, a benchmark should be considered for retirement or supplementation. By these criteria, HellaSwag and Winogrande were due for retirement by late 2023—roughly 6 months before HuggingFace actually retired them. We note that these thresholds are proposed heuristics calibrated against this case study; future work should validate them across additional leaderboards.

7.2 BDI Monitoring

We recommend that leaderboard operators compute and publish BDI alongside benchmark scores. This provides early warning of saturation and enables data-driven decisions about benchmark replacement. A simple dashboard showing BDI trends over time would allow the community to anticipate—rather than react to—benchmark obsolescence.

7.3 Limitations

Population shifts. BDI measures the score distribution of *submitted* models, which changes over time as the community shifts toward larger, more capable architectures. A BDI decline could partially reflect population homogenization (everyone submitting similar models) rather than pure benchmark saturation. Disentangling these effects requires controlling for model capability, which we leave to future work.

Open-weight models only. The Open LLM Leaderboard evaluates only open-weight models. Proprietary models (GPT-4, Claude, Gemini) are excluded. Including these would likely show even higher ceiling proximity on V1 benchmarks.

Score inflation. Some models may achieve high scores through benchmark contamination (training on test data) rather than genuine capability improvement. We do not attempt to distinguish genuine from inflated scores; our analysis reflects the leaderboard as published.

Model family overlap. The leaderboard contains many fine-tuned variants of the same base model (e.g., 96 Qwen2.5 variants in V2). While we count 4,039 and 2,174 unique model families in V1 and V2 respectively, correlated scores within families may inflate apparent distributional patterns. BDI trends, however, are robust to this since they track *temporal changes* in the same population.

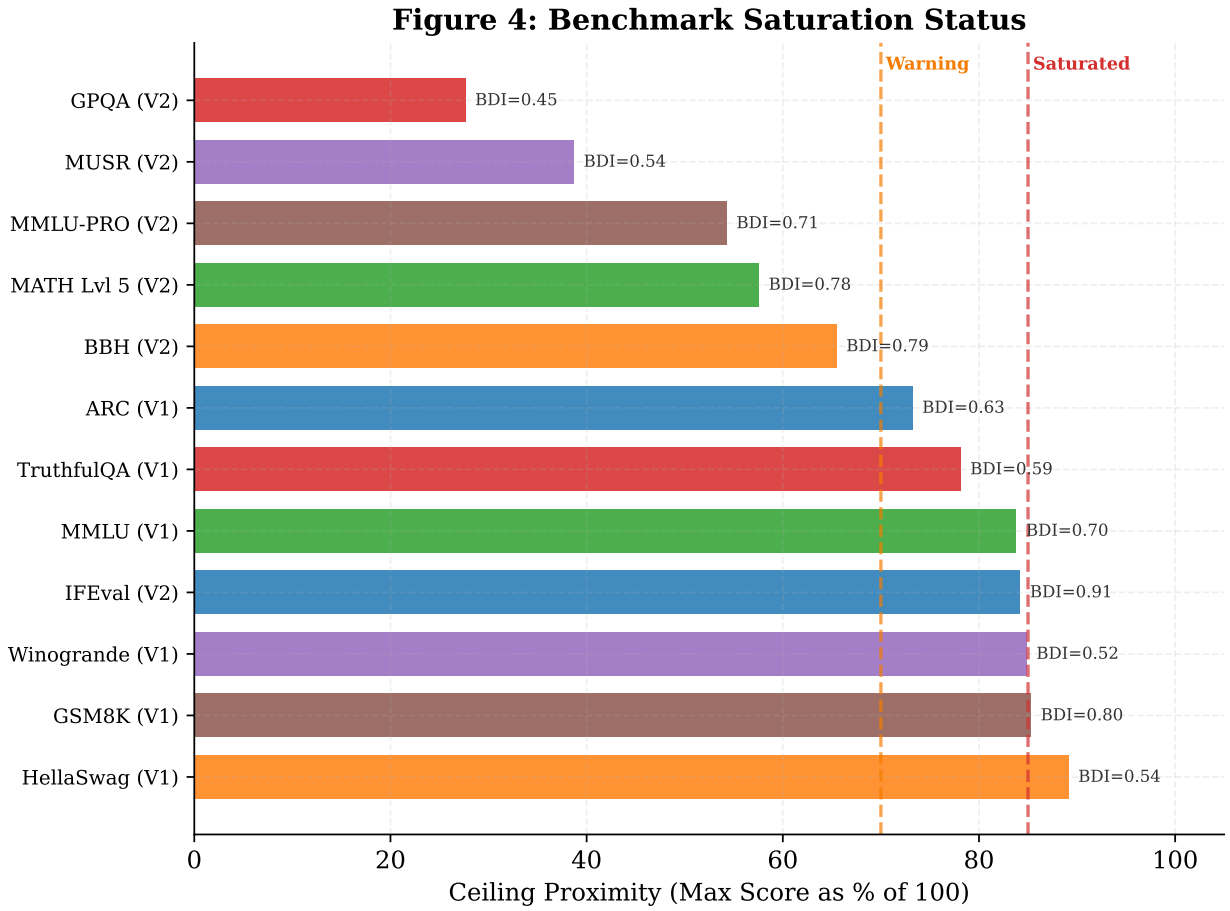

Figure 4: Saturation status of all 12 benchmarks (latest month). Bars show ceiling proximity; BDI values annotated. Red dashed line marks the saturation threshold (85%); orange marks the warning zone (70%).

V2 limited time window. The V2 leaderboard has only 10 months of data. Saturation curves for V2 benchmarks should be treated as preliminary projections.

Single leaderboard. We analyze only the HuggingFace Open LLM Leaderboard. Other evaluation platforms (LMSYS Chatbot Arena, Scale AI SEAL) may show different dynamics.

8 Conclusion and Future Work

We present the first systematic quantification of benchmark saturation across LLM leaderboards. Our analysis of 11,836 evaluations across 12 benchmarks reveals that saturation is predictable and quantifiable. The Benchmark Discriminability Index provides a principled, entropy-based metric for monitoring benchmark health. Knowledge-retrieval benchmarks consistently saturate faster than reasoning benchmarks, suggesting that evaluation suites should increasingly emphasize reasoning tasks. We recommend that leaderboard operators adopt BDI monitoring as standard practice to enable proactive—rather than reactive—benchmark management.

Figure 5: Score Gap Compression Among Top Models

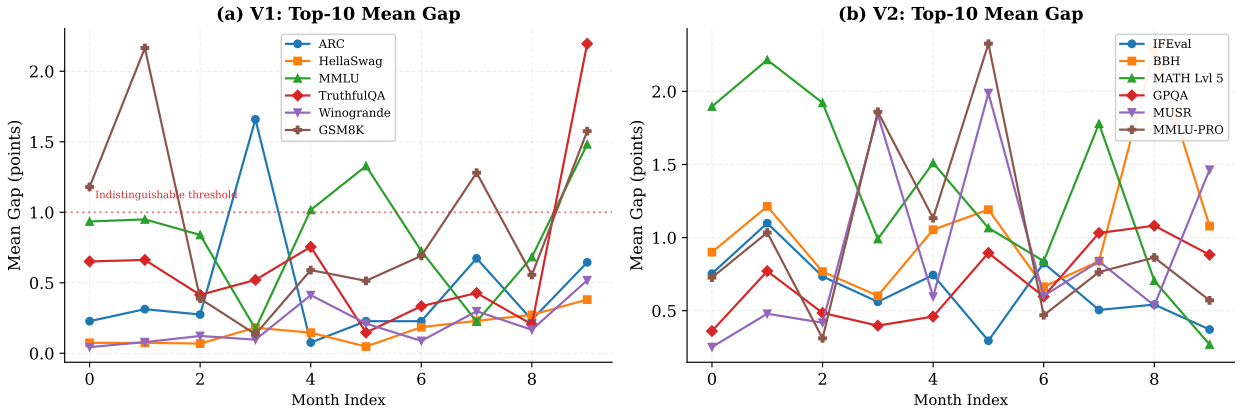

Figure 5: Mean score gap between adjacent top-10 models over time. Gaps below 1 point (red dotted line) indicate that the benchmark can no longer reliably distinguish top models.

Future work. Several extensions are natural: (1) incorporating data from LMSYS Chatbot Arena and closed-source model evaluations to test generalizability; (2) developing a BDI variant that controls for model population shifts; (3) studying whether benchmark contamination accelerates saturation; and (4) connecting BDI to downstream utility—at what BDI threshold does a benchmark lose practical value for model selection?

Reproducibility

All data used in this study is publicly available on the HuggingFace Hub: `open-llm-leaderboard-old/contents` (V1) and `open-llm-leaderboard/contents` (V2). Analysis code and generated datasets will be released upon publication.

References

Anonymous. Towards reproducible LLM evaluation: Quantifying uncertainty in LLM benchmark scores. *arXiv:2410.03492*, 2024.

Anonymous. Benchmark illusion: Disagreement among LLMs and its scientific consequences. *arXiv:2602.11898*, 2026a.

Anonymous. Efficient evaluation of LLM performance with statistical guarantees. *arXiv:2601.20251*, 2026b.

Anonymous. Towards more standardized AI evaluation: From models to agents. *arXiv:2602.18029*, 2026c.

Anonymous. Same meaning, different scores: Lexical and syntactic sensitivity in LLM evaluation. *arXiv:2602.17316*, 2026d.

Samuel R Bowman and George E Dahl. What will it take to fix benchmarking in natural language understanding? In *NAACL*, 2021.

Peter Clark et al. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv:1803.05457*, 2018.

Karl Cobbe et al. Training verifiers to solve math word problems. *arXiv:2110.14168*, 2021.

Dan Hendrycks et al. Measuring mathematical problem solving with the MATH dataset. *NeurIPS*, 2021a.

Dan Hendrycks et al. Measuring massive multitask language understanding. In *ICLR*, 2021b.

Hugging Face. Open LLM leaderboard v2, 2024. URL `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`.

Douwe Kiela et al. Dynabench: Rethinking benchmarking in NLP. In *NAACL*, 2021.

Percy Liang et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *ACL*, 2022.

NIST. Expanding the AI evaluation toolbox with statistical models. Technical report, National Institute of Standards and Technology, 2026.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? *ICML*, 2019.

David Rein et al. GPQA: A graduate-level google-proof q&a benchmark. *arXiv:2311.12022*, 2024.

Keisuke Sakaguchi et al. WinoGrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 2021.

Zayne Sprague et al. MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. *arXiv:2310.16049*, 2024.

Mirac Suzgun et al. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. 2023.

Bob Uttl. North American adult reading test: Age norms, reliability, and validity. *Journal of Clinical and Experimental Neuropsychology*, 2005.

Yubo Wang et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *arXiv:2406.01574*, 2024.

Rowan Zellers et al. HellaSwag: Can a machine really finish your sentence? In *ACL*, 2019.

Jeffrey Zhou et al. Instruction-following evaluation for large language models. *arXiv:2311.07911*, 2023.