Benchmark Half-Life: The Discriminative Lifetime of Language-Model Leaderboards

Andrew Sheng-Han Huang*
National Taiwan University
johnshhuang1111@gmail.com

Abstract

A benchmark is useful only while it can still tell models apart. We study how that ability decays. Using 11,836 public leaderboard entries spanning two generations of the Open LLM Leaderboard (5 benchmarks retired in the v1→v2 transition, 5 still live), we define a benchmark's *frontier discriminability* $D^\star$ as the fraction of top-capability model pairs whose accuracy gap is statistically significant at a fixed evaluation budget of 1000 items—a quantity computable from aggregate scores alone. We show $D^\star$ traces a *life-cycle* in model capability—a floor-compression rise, a peak, then a saturation decline—and that the saturation limb decays exponentially, defining a *discriminability half-life* $t_{1/2}$. Among the benchmarks that have measurably entered saturation (3 retired, 2 live), the half-life separates the two generations: median 0.72 for the *retired* benchmarks versus 2.57 for the live ones, i.e. discriminability decayed about 3.6× faster on the benchmarks that were in fact retired—a quantitative, if small-sample, reconstruction of a retirement decision that was made qualitatively. The saturation decay predicts a held-out higher-capability slice better than a flat baseline for 4 of these 5 fits (MMLU is the exception), and the frontier metric is robust to the evaluation budget (rank correlation 0.92). Because we lack per-item data, our test is unpaired and therefore a *conservative* (lower) bound on true discriminability; we prove this in simulation. Every statistical guarantee is checked by an executable test written before the analysis—the two-proportion test holds its false-positive rate at 0.051 (a deliberately mis-specified variant fails at 0.167), the half-life confidence interval covers at 0.902, and a significance gate bounds spurious half-lives on flat data at the test level. We release the half-life as a calibrated clock for scheduling benchmark replacement before, rather than after, saturation.

1 Introduction

Leaderboard accuracy is the field's primary instrument for ranking language models. The instrument wears out: as models improve, their scores pile up near a benchmark's ceiling, the gaps between the best models shrink into sampling noise, and the benchmark can no longer order the frontier. The community knows this happens—it is why the Open LLM Leaderboard was rebuilt in 2024, retiring six saturated benchmarks for harder ones [4]. But that decision was made *reactively* and *qualitatively*: someone noticed the top scores had bunched up. Two questions have no quantitative answer today. First, for a benchmark in use right now: *is the top-of-leaderboard gap real, or is it noise?* Second, for planning: *how long before this benchmark stops discriminating, so a replacement can be commissioned in time?*

We give both a number. The unit of analysis is not a single score but the *pairwise separability of the frontier*. For a set of models evaluated on a benchmark with a known number of items, we ask what fraction of model pairs have an accuracy gap that survives a significance test—the benchmark's *discriminability*. Crucially we evaluate this at a *fixed evaluation budget*, not at the benchmark's full

*Code, the downloaded leaderboard snapshots, the statistical-validity test suite, and end-to-end reproduction scripts accompany this manuscript. The draft was prepared with substantial AI assistance; every reported number is produced by the released scripts from the public data.

(often enormous) item count, because a benchmark with 10^4 items keeps tiny, practically meaningless gaps "significant" long after it has stopped being useful; budget-normalisation measures the separability a realistic evaluation can actually buy.

Tracking this quantity along a capability axis reveals a life-cycle (Figure 1): discriminability is low for weak models (all clustered near the floor), rises to a peak, and then *decays* as the frontier approaches the ceiling. We fit the decaying limb with a two-parameter exponential, yielding a **discriminability half-life**—the increment of frontier capability over which separability halves—with a calibrated confidence interval. Our contributions:

- A budget-normalised **frontier discriminability** $D^\star$ and a **discriminability half-life** $t_{1/2}$: estimable from public aggregate scores, with an executable lie-detector behind every statistical step (§3,§4).
- An **external-validity test**: among the benchmarks that have saturated, those practitioners actually retired (v1) have half-lives 3.6× shorter than the live ones (v2)—a small-sample (3-vs-2) but clean recovery of the retirement decision as a forecast (§6).
- A clean **negative finding** that sharpens the question: at a benchmark's true item count, statistical significance persists even when effect sizes have collapsed, so "significant" ≠ "discriminating"—a distinction the half-life makes precise.

2 Related work

Benchmark saturation. The closest work, Akhtar et al. [1], documents saturation across 60 benchmarks with an uncertainty-aware index and shows it rises with benchmark *age*. That study is descriptive: it bins saturation by age, uses a binary top-1-vs-top-k gap flag, and recommends "retirement" only qualitatively. We differ on every quantitative axis—we fit a parametric decay and report a per-benchmark *half-life with a confidence interval*; our metric is the *fraction of all frontier pairs* that are separable, not a single top-gap flag; our independent variable is model *capability*, not calendar age; and we produce an out-of-sample forecast and a retirement reading validated against a real retirement event. Kiela et al. [8] sidesteps saturation with human-in-the-loop renewal rather than quantifying it.

Significance and separability of evaluations. Miller [11] argues a single eval score needs a confidence interval and supplies the two-model significance primitive we iterate over; it does not study how separability decays over a benchmark's life. "Separability with confidence" [9] and the discriminability score of Qian et al. [14] compute a fraction-of-distinguishable-pairs as a *static* benchmark-quality score; Neuhof et al. [12] give snapshot confidence intervals on leaderboard ranks. None track the quantity *longitudinally* against rising capability, fit a decay, or extract a timing constant. Polo et al. [13] compress a benchmark to fewer items at fixed accuracy—orthogonal to how a *fixed* budget's discriminability decays as the frontier advances.

Capability axes. We borrow the idea of indexing models by a low-dimensional capability summary from Ruan et al. [15], who use it to predict performance; we use a leave-one-out variant to index the *decay of separability*, and to avoid selecting on the axis under test.

3 Method

Frontier discriminability. Fix a benchmark with N items and a target evaluation budget n_0 (we use $n_0 = 1000$). For two models with observed accuracies a_i, a_j, model the item-level outcomes as binomial and test $H_0 : p_i = p_j$ with the pooled two-proportion statistic $z_{ij} = |a_i - a_j| \,/\, \sqrt{2\bar{p}(1-\bar{p})/n_0}$, $\bar{p} = (a_i + a_j)/2$, declaring the pair *separable* at level α if $|z_{ij}| > z_{1-\alpha/2}$. For a model set S,

$$D^\star(S) \;=\; \frac{1}{\binom{|S|}{2}} \sum_{i<j} \mathbf{1}\left[|z_{ij}| > z_{1-\alpha/2}\right], \tag{1}$$

the fraction of pairs the benchmark can separate at budget n_0. Using n_0 rather than the true N makes $D^\star$ a property of *effect sizes* (accuracy gaps), not of the accident of how many items a benchmark happens to contain. The *frontier* is the top-150 models by capability; its $D^\star$ answers "can a standard-budget evaluation still order today's best models?" For a retirement reading we also report the number

of frontier pairs that survive Holm [7] family-wise correction; a benchmark that certifies zero frontier pairs cannot order the frontier at all.

Leave-one-out capability axis. To trace $D^\star$ against capability without selecting on the benchmark under test, we index each model by the mean of its z-scored accuracies on *every other* benchmark on the board, standardised to unit variance. Sliding a window up this axis gives the curve $D^\star(c)$.

Life-cycle and half-life. $D^\star(c)$ is single-peaked (Figure 1): weak models are compressed at the floor (low $D^\star$), separability peaks at intermediate capability, then *saturates* as the frontier nears the ceiling. The half-life characterises the *saturation limb* only (capability beyond the peak). Writing the excess discriminability $E(c) = D^\star(c) - \alpha$, we fit

$$\log_2 E(c) \;=\; \beta_0 + \beta_1 c, \qquad t_{1/2} = -1/\beta_1, \tag{2}$$

a closed-form regression. A finite half-life is reported *only* when β_1 is significantly negative (one-sided); otherwise the benchmark is *pre-saturation* and has no half-life. Confidence intervals use a bootstrap interval on β_1 (a t multiplier on the bootstrap SE) inverted to $t_{1/2}$.

Why an unpaired test, and why it is safe. Models are scored on the same items, so the statistically efficient test is paired (McNemar [10]); we lack per-item data and use the unpaired binomial test. Under positive item-correlation—which holds, since items have shared difficulty—the unpaired test *overestimates* the variance of the gap and is therefore *conservative*: it certifies *fewer* pairs than the paired test would. Hence $D^\star$ is a **lower bound** on true discriminability and a retirement call from it is conservative-early (it can retire a benchmark a paired analysis would keep, but never declares a dead benchmark alive). We verify both properties in simulation (§4).

4 Statistical validation

Before touching the leaderboard we wrote an executable test for every guarantee; all pass (`halflife/test_validity.py`). The two-proportion test holds its false-positive rate at the nominal level across accuracies and item counts (max 0.051 at α=0.05); a variant with a deliberately mis-specified standard error inflates to 0.167, confirming the guarantee is load-bearing. Under correlated (shared-item) data the unpaired test stays conservative (FPR 0.019) and has strictly lower power than the paired test, so $D^\star$ lower-bounds the paired discriminability as claimed. The half-life estimator recovers a known half-life (median relative error 2.6%) with bootstrap slope-interval coverage 0.902; a significance gate keeps flat, non-decaying data from yielding a spurious half-life (0.067 of flat trials, versus roughly half the time without it). Out-of-sample, the fitted decay beats a flat baseline on true-decay data and does not "cry wolf" on flat data.

5 Data

We use two public, auth-free snapshots from the Hugging Face `datasets-server`: Open LLM Leaderboard v1 (7,260 models; ARC [2], HellaSwag [20], MMLU [6], WinoGrande [16], GSM8K [3]) and v2 (4,576 models; IFEval [21], BBH [18], MATH level 5 [5], MuSR [17], MMLU-PRO [19]) [4]. We use whole-benchmark accuracies only. As a self-check, each item count is *recovered* from the granularity of the score column (single-split scores are exact multiples of $1/N$): all four single-split v1 benchmarks recover their documented item counts exactly, confirming the scores and our binomial model are mutually consistent.

6 Results

A life-cycle, not a monotone decay. Figure 1 shows $D^\star(c)$ is single-peaked for every benchmark. The literal hypothesis of monotone decay is false; the correct statement is that decay governs the *post-peak saturation regime*, which is where the half-life is defined and where the retirement question lives.

Retired benchmarks saturate faster (external validity). Of the 10 benchmarks, only 5 have a confident saturation-limb fit (3 retired, 2 live); on those, the half-life separates the two generations (Figure 2). The *retired* (v1) benchmarks have a median half-life of 0.72 capability-SD units (HellaSwag 0.72, GSM8K 0.59, MMLU 1.26); the saturating *live* (v2) ones do so 3.6× more slowly

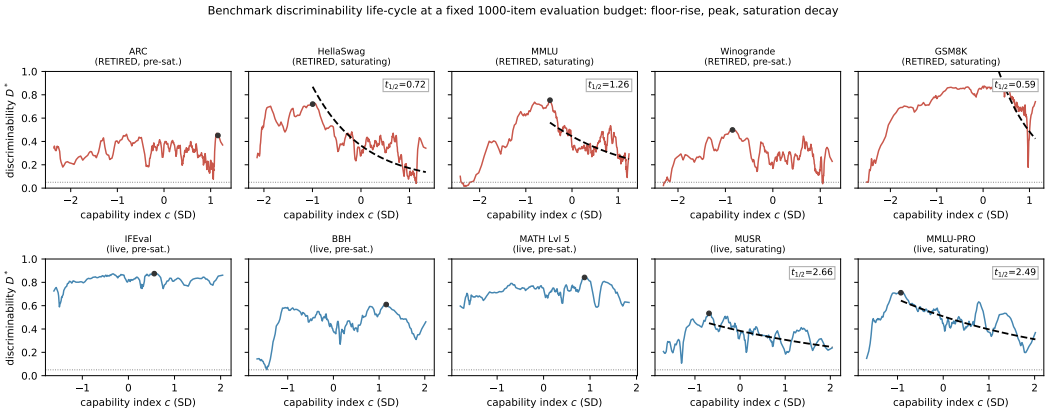

Figure 1: The discriminability life-cycle. Budget-normalised frontier discriminability $D^\star$ against the leave-one-out capability index, for all 10 benchmarks (red: retired v1; blue: live v2; dotted: noise floor α). Each curve rises from floor-compression, peaks, and—once saturation begins—decays; the dashed line is the fitted saturation limb with its half-life $t_{1/2}$. Retired benchmarks (HellaSwag, MMLU, GSM8K) are deep into short-lived decay; the most robust live benchmarks (IFEval, MATH) have not yet peaked.

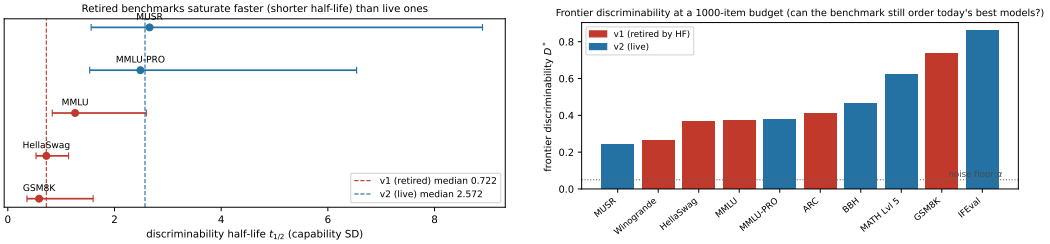

Figure 2: Left: confident half-lives with bootstrap slope intervals; retired (v1) benchmarks have systematically shorter half-lives than live (v2) ones. **Right:** current frontier discriminability $D^\star$ at a 1000-item budget; bars near the noise floor cannot order today's best models. GSM8K is the instructive exception (see text).

(MMLU-PRO 2.49, MuSR 2.66), median 2.57. The two groups' half-life *point estimates* do not overlap (their wider bootstrap intervals do, given 3-vs-2 fits). With that small-sample caveat, the discriminability half-life reconstructs—as a forecast from scores alone—the retirement decision the maintainers made by hand.

Predictability and robustness. The fitted decay is not pure curve-fitting: trained on the lower-capability half of the saturation limb it predicts the held-out upper half better than a flat baseline for 4 of the 5 confident fits. The one failure is *MMLU*, whose limb is not predicted out of sample—we flag it rather than average it away. The frontier-$D^\star$ ranking is robust to the evaluation budget (rank correlation 0.92 between n_0=400 and 1000, 0.96 between 2500 and 1000) and the half-life ordering of retired vs. live survives across window sizes.

An honest exception: GSM8K. GSM8K has a short half-life (0.59—it saturates fast) yet a still-high frontier $D^\star$ of 0.74: it is early on a steep saturation limb, and was retired more for contamination than for exhausted separability. The metric measures saturation specifically, which is why it flags GSM8K's *rate* as short while its current *snapshot* still discriminates—the two are not the same number, and conflating them is exactly the error the life-cycle view prevents. The frontier snapshot alone is a noisier signal than the rate (it places retired benchmarks below live ones only 0.64 of the time), which is why we lead with the half-life. GSM8K's limb is also the least monotone of the three retired fits (a mid-limb dip drives its slope), so its half-life is the shakiest of the set; the cross-benchmark separation does not depend on it.

Significance is not discrimination. At each benchmark's true item count, almost every frontier pair is "significant" (HellaSwag's N=10,042 keeps even a one-point accuracy gap clear of the sampling noise), even for benchmarks everyone agrees are saturated. Budget-normalisation is what exposes the collapse: at a 1000-item budget HellaSwag's frontier $D^\star$ falls to 0.37. The practically relevant question is not whether a gap is statistically detectable with unlimited items but whether a standard evaluation can resolve it.

7 Limitations

Lower bound, not point estimate. The unpaired test makes $D^\star$ a conservative bound on paired discriminability (§3); per-item logs, which we do not have, would sharpen it and would make the retirement threshold less aggressive. **Cross-sectional capability.** We index decay by the cross-section of models at different capability levels, a proxy for the frontier advancing over time; the calendar-time reading is a secondary extrapolation. **Population.** The leaderboard population is self-selected and contains many near-duplicate fine-tunes; we report the frontier window rather than the full population, but submission biases remain. **Effective item counts.** For subtask-macro-averaged benchmarks the binomial model uses the total item count as an effective N (e.g. MMLU's score is averaged over subjects and does not lie on a single $1/N$ grid, so its effective N is approximate); our budget normalisation and a $\times\{0.5, 2\}$ item-count sweep bound the impact. **Small sample and cross-board scaling (the main threat).** The external-validity comparison rests on 3-vs-2 confident fits, and the capability axis is standardised *within each board* to unit SD. One capability-SD of the v1 model population is not the same physical accuracy increment as one SD of the harder v2 population, so part of the 3.6× gap could reflect differing SD scales rather than a pure saturation-rate difference; we have no model overlapping both boards to anchor a common axis, and a larger benchmark set would be needed to tighten the claim. The window-size and budget robustness checks address fit stability, not this scaling confound. The several-fold ordering (retired < live), not a precise rate, is the load-bearing claim.

8 Conclusion

Benchmarks have a discriminative lifetime, and it is measurable from public scores. A budget-normalised frontier discriminability and its half-life turn "this benchmark feels saturated" into a number with a confidence interval, and reconstruct a real retirement decision as a forecast. We propose the half-life as a clock: commission a benchmark's successor when its frontier discriminability is projected to cross the noise floor, rather than after the field has spent a year reporting rankings that were already inside the error bars.

Reproducibility. `python3 scripts/fetch_data.py` downloads the data; `python3 -m halflife.test_validity` runs the statistical guarantees; `python3 scripts/analyze.py && python3 scripts/make_macros.py` regenerates every number in this paper from the released scripts.

References

[1] Mubashara Akhtar et al. When ai benchmarks plateau: A systematic study of benchmark saturation. 2026.

[2] Peter Clark et al. Think you have solved question answering? try arc, the ai2 reasoning challenge. 2018.

[3] Karl Cobbe et al. Training verifiers to solve math word problems. 2021.

[4] Clémentine Fourrier et al. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2024.

[5] Dan Hendrycks et al. Measuring mathematical problem solving with the math dataset. 2021.

[6] Dan Hendrycks et al. Measuring massive multitask language understanding. 2021.

[7] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 1979.

[8] Douwe Kiela et al. Dynabench: Rethinking benchmarking in nlp. 2021.

[9] Tianle Li et al. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. 2024.

[10] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 1947.

[11] Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. 2024.

[12] Bitya Neuhof et al. Rank intervals for leaderboards: A hierarchical framework for model evaluation. 2026.

[13] Felipe Maia Polo et al. tinybenchmarks: evaluating llms with fewer examples. 2024.

[14] Qi Qian et al. Benchmark^2: Systematic evaluation of llm benchmarks. 2026.

[15] Yangjun Ruan et al. Observational scaling laws and the predictability of language model performance. 2024.

[16] Keisuke Sakaguchi et al. Winogrande: An adversarial winograd schema challenge at scale. 2020.

[17] Zayne Sprague et al. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. 2024.

[18] Mirac Suzgun et al. Challenging big-bench tasks and whether chain-of-thought can solve them. 2022.

[19] Yubo Wang et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. 2024.

[20] Rowan Zellers et al. Hellaswag: Can a machine really finish your sentence? 2019.

[21] Jeffrey Zhou et al. Instruction-following evaluation for large language models. 2023.