EvalEval Coalition

When AI Benchmarks Plateau: A Systematic Study of Benchmark Saturation

Akhtar,* Reuel,* et al. (ICML 2026)

ICML International Conference On Machine Learning

EvalEval

Saturation - When benchmarks stop measuring progress

- Benchmarks drive high-stakes AI governance decisions (EU AI Act Article 51, AISI Inspect, etc.)
- Regulatory processes require stable measurement tools
- Benchmarks lose discriminative power over time
- We lack a clear way to define and measure saturation
- Our study: Which design choices predict faster or slower saturation?

AIM

Need to understand what controls benchmark longevity.

→ *Saturation formalization*
→ *Impacting factors*

How to Quantify Saturation?

Definition

Saturation = top-performing models can no longer be statistically distinguished, and performance approaches the empirical ceiling.

vs. Stagnation

Distinct from stagnation, where models cluster but a ceiling hasn't been reached — stagnation can resolve after architectural breakthroughs (e.g. ARC-AGI V1).

Our Index

Compares the spread among top-*k* models against estimated measurement uncertainty. Model-relative, metric-agnostic, reproducible from public leaderboard snapshots.

Saturation is widely spread

29

out of 60 benchmarks

show high or very high saturation

~48% OF BENCHMARKS

- Applied to 60 text-based LLM benchmarks from major developer technical reports
- Annotated across 14 benchmark design properties: task design, data construction, and evaluation format
- Saturation is widely spread across benchmark families

Myth Busting: What does *not* prevent saturation

✕ Private test sets

No protective effect once distributional characteristics become known — public and private benchmarks show similar saturation.

✕ Open-ended formats

Do not preserve discriminative power over multiple choice formats.

✕ Multilinguality

No significant effect after controlling for age. Multilingual benchmarks are simply younger than English-only ones.

✕ Popularity / citations

Not significant once age is accounted for. Cumulative exposure, not visibility, drives convergence.

Implications & Recommendations

IMPLICATIONS

- Static benchmark lists in governance may be especially vulnerable
- They often cannot be updated at the pace benchmarks go stale
- Saturation on benchmarks with known validity issues is the worst case — it creates false confidence that a capability question is resolved

RECOMMENDATIONS

- Require reporting of uncertainty-aware statistics and score compression, not just peak performance
- Support stratified sub-skill reporting and multi-metric analysis to reveal hidden variation
- Future vision: A live saturation monitoring index — key limitation is dependency on leaderboard data quality

Benchmark Saturation - Main Takeaways

Saturation is structural, not accidental

Common safeguards are insufficient

Design choices that matter

From static benchmarks to lifecycle management