The Benchmark Saturation & Validity Crisis

Why Current LLM Evaluation is Failing

Yousef Khattari

University of Go¨ttingen

Why Do We Need Benchmarks?

- Compare models objectively
- Track progress over time
- Identify strengths and weaknesses
- Guide research and deployment

Benchmarks are the measurement tools of AI

Benchmarks Shape the Entire Field

- What researchers optimize
- What gets published
- What companies deploy

What we measure → what we build

The Illusion of Progress

- Leaderboards show rapid improvements every year
- New models constantly claim SOTA

But do higher scores really mean better intelligence?

Core Assumption

High benchmark score

$\Downarrow$

High real-world capability

Our Claim

Benchmark Validity Crisis
Current LLM evaluation is becoming unreliable.

Scores may not reflect real capability.

Three Failure Modes

1. Benchmark Saturation
2. Data Contamination
3. Construct Validity Failure

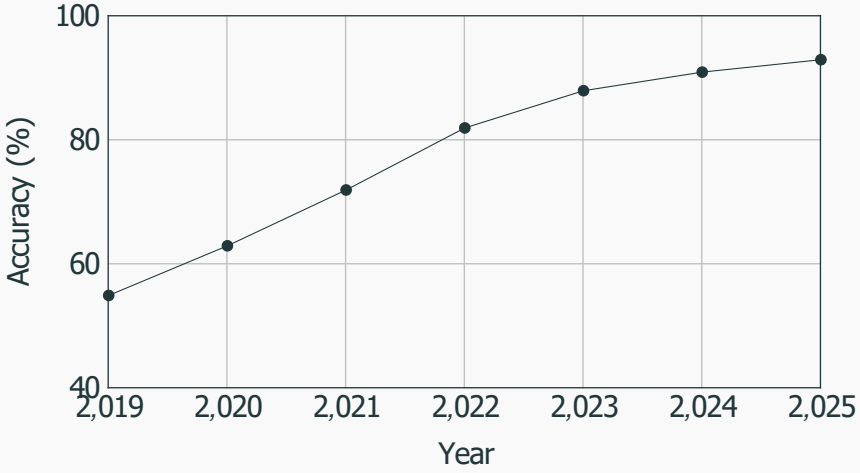
These combine to create misleading progress

Failure #1: Benchmark Saturation

- Many models achieve near-ceiling scores
- Tiny differences dominate leaderboards
- Hard to distinguish real improvements

The test becomes too easy

Example: Rapid Benchmark Saturation



Ceiling performance leaves little differentiation

Impact of Saturation

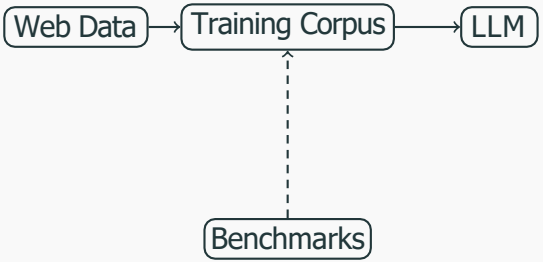
- Statistical noise looks like breakthroughs
- Encourages benchmark gaming
- Reduces evaluation resolution

Failure #2: Data Contamination

- Benchmarks are public
- Training data is web-scale
- Test questions may appear in training

$$\text{Train} \not\approx \text{Test} \text{ no longer holds}$$

How Contamination Happens



Benchmark data can leak into training

How It Happens

- Web scraping
- Dataset reuse
- Fine-tuning leakage

Impact of Contamination

- Memorization instead of reasoning
- Inflated scores
- False sense of generalization

Failure #3: Construct Validity

- Do benchmarks measure what they claim?
- Many rely on narrow proxy tasks

We may be measuring the wrong abilities

Proxy Problem

- Multiple-choice ≠ real reasoning
- Refusal tests ≠ true safety
- Static QA ≠ real deployment

Impact of Weak Validity

- Scores don’t predict real-world success
- Progress becomes misleading

How These Failures Interact

- Contamination accelerates saturation
- Saturation hides leakage
- Weak tasks mask both

A systemic measurement crisis

Benchmark Family Comparison

Benchmark	Saturation	Contamination Risk	Validity
MMLU	High	High	Medium
GSM8K	Medium	Medium	Narrow
BIG-Bench	Medium	Unknown	Mixed
HELM	Lower	Lower	Broader

Different benchmarks fail in different ways

Feedback Loop

Contamination → Higher scores
→ Saturation
→ Misleading benchmarks
→ More optimization

What Should Replace Current Benchmarks?

- Dynamic tasks
- Private/held-out test sets
- Domain-specific evaluation
- Continuous monitoring

Better Evaluation Philosophy

Dynamic
Realistic
Hard to game
Continuous

Key Takeaways

- Benchmarks drive AI progress
- Current benchmarks are breaking
- We need better measurement to ensure real progress

References

[1] T. Ivanov and V. Penchev, “AI Benchmarks and Datasets for LLM Evaluation,” 2024.

[2] C. Chengxu, S. Guan, D. Greene, and M.-T. Kechadi, “Benchmark Data Contamination of Large Language Models: A Survey.”

[3] P. Liang et al., “Holistic Evaluation of Language Models,” Transactions on Machine Learning Research, 2023.

[4] BIG-bench Collaboration, “Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models.”

[5] H. Zhou et al., “Lost in Benchmarks? Rethinking Large Language Model Benchmarking with Item Response Theory.”

[6] D. Hendrycks et al., “Measuring Massive Multitask Language Understanding,” ICLR, 2021.

[7] A. M. Bean et al., “Measuring What Matters: Construct Validity in Large Language Model Benchmarks.”

We cannot improve what we cannot measure correctly.

Thank you — Questions?