

As the landscape of large language models evolves at breakneck speed, observers frequently encounter a perplexing notation on leaderboards and timelines: certain highly anticipated models are marked as “not rated.” In places like the LMArena text leaderboard—which offers a rare style-controlled comparison—and datasets such as Hugging Face’s `lmarena-ai/leaderboard-dataset`, this “not rated” status raises questions for researchers, product managers, and enthusiasts alike.

This blog post unpacks why many recently announced gated models—especially those from heavyweights like OpenAI, Anthropic, and Google DeepMind—either never appeared on LMArena or remain unrated despite months passing since announcement. We’ll dissect key themes including the acceleration of release cadences post-2023, diminishing user-perceived gains per iteration, the tension between blind-vote preference testing versus benchmark marketing, and the common confusion between announced versus shipped models.

Accelerating Release Cadence Since 2023: A Double-Edged Sword

In the era before 2023, large language models (LLMs) typically debuted with months — if not years — between major releases. OpenAI, Google DeepMind, and other labs had the bandwidth to meticulously test, benchmark, and ship models before unveiling the next iteration.

Since 2023, however, the release cadence has accelerated to quarterly or even monthly frequencies, spawning a profusion of versions. This rapid cycle supports a dynamic innovation culture but has unintended effects on rating and benchmarking:

- **Shifting Baselines:** Rapid releases often come with minimal publicly available baseline data, making it difficult for benchmarks like LMArena to update ratings consistently.
- **Confusion Between Announced and Shipped:** Companies frequently announce upcoming models well ahead of shipment, creating timelines where models are listed but never fully accessible.
- **Benchmark Fatigue:** Benchmark curators struggle to test every new gated model exhaustively given resource constraints, leading to fewer models rated.

A pertinent example is Anthropic’s series of Claude models. While <https://suprmind.ai/hub/ai-models-index/> updates are announced with excitement, gated access and staggered internal evaluation periods often preclude immediate leaderboard inclusion, prolonging the “not rated” status.

Shrinking User-Perceived Gains Per Release

Early LLM releases exhibited sizable step-changes in capability, making new models easy to evaluate and highly sought after. In 2023 and 2024, the incremental gains per release have typically diminished, reflecting maturation of core architectures and optimizations rather than leaps in fundamental ability.

This dampening effect leads to benchmarking and rating becoming more nuanced:



- Subtle improvements often require fine-grained, large-scale preference testing rather than straightforward benchmark scores.
- Statistical confidence bands widen as gains approach the noise floor, complicating leaderboard ranking.
- End-users perceive improvements more in style, safety, or aligned interaction quality than raw benchmark points.

Consequently, “not rated” labels may also signal that current benchmarks, such as those tracked by LMArena or aggregated in Hugging Face datasets, are ill-equipped to capture incremental user value, leading curators to hold off on ratings until more robust, differentiating data emerges.

Blind-Vote Preference Testing vs Benchmark Marketing

Many model-makers including OpenAI, Anthropic, and Google DeepMind now emphasize blind-vote preference testing as a gold standard for release evaluation. This human-in-the-loop approach contrasts traditional benchmarks by capturing nuanced subjective preferences rather than pure accuracy or knowledge metrics.

However, blind-vote testing has key implications for public rating systems:

1. **Opaque Results:** Preference results are typically proprietary and reported as aggregate metrics, rarely releasing detailed test sets or raw outputs.
2. **Non-Comparability:** Preference scores are model- and context-dependent, complicating direct, apples-to-apples comparisons on text benchmarks.
3. **Delayed Public Scoring:** Because blind preference tests often require months of user interaction and statistical validation, model ratings lag behind announcements.

By contrast, benchmark marketing—highlighting tasks where models perform best—can promote fast leaderboard rankings but often cherry-picks favorable snapshots and ignores confidence intervals or long-term drift. This divergence means gated models can hover in limbo: announced and hyped yet “not rated” by benchmarks waiting for independent, repeatable preference data.

Announced vs Shipped: Clearing Up the Confusion

One of the most common sources of misunderstanding is the conflation of “announced” and “shipped” model versions.

- **Announced Models:** These are versions showcased by companies in blog posts, research papers, or press releases. They may represent internal prototypes, partial rollouts, or plans with projected capabilities rather than finalized, released products.
- **Shipped Models:** Fully deployed and accessible models, either gated behind API keys or publicly available for testing and benchmarking.

Leading companies frequently announce models months—or even quarters—ahead of general availability. For example, Google DeepMind's advances in the Gemini series are often teased to illustrate research direction but do not become immediately accessible for leaderboard evaluation.



This lag means that many timeline entries on LMArena or datasets like Hugging Face’s `lmarena-ai/leaderboard-dataset` read as “not rated.” The model physically exists in announcements and marketing collateral, but no valid public or gated ranking can verify its performance yet.

Gated Models: The Invisible Majority on Leaderboards

An important factor is the prevalence of gated models—high-performing but access-restricted versions controlled by the originating company. These gated models mostly emerge in 2023 and 2024, aligning with new commercial API-driven business models. They are often unavailable for independent benchmarking or user-level testing, reinforcing the “not rated” phenomenon.

Because leaderboard curators and the LLM research community rely on open or at least sandboxed access to evaluate models reliably, access restrictions functionally exclude many gated models from consistent rating pipelines.

Additionally, gated models often see internal updates and versioning invisible externally, so tracking quality drift or improvement becomes nearly impossible without company cooperation. This opacity creates moving leaderboards where the “not rated” entries correspond to largely black-box performance claims.

Practical Takeaways for Buyers and Builders

1. **Don't Equate "New Version" with "Better Version."** New model releases, especially gated ones announced rapidly, aren't guaranteed to outperform previous public benchmarks.
2. **Watch for "Not Rated" as a Signal of Evaluation Incompleteness.** It often flags models yet to be fully vetted in public benchmarks or preference testing, not necessarily poor quality.
3. **Leverage Multiple Data Sources.** Use LMArena's text leaderboard alongside Hugging Face datasets to triangulate model behavior and identify missing baselines or shifting leaderboards.
4. **Demand Transparency in Benchmarking.** Request detailed test set releases, confidence intervals, and strict adherence to stable leaderboards to avoid cherry-picked marketing numbers.
5. **Factor In Access Constraints.** Gated models can offer superior capabilities behind API walls, but evaluating them requires company cooperation or internal adoption rather than public leaderboard trust.

Conclusion

The "not rated" annotations peppered through LMArena timelines and Hugging Face leaderboard datasets are not mere oversights. Instead, they reflect complex market dynamics: an accelerated release cadence, shrinking apparent gains per release, reliance on blind preference testing, and the gap between announced and shipped models—all underscored by the rise of gated, commercialized AI models.

For teams building on top of or buying subscriptions to these AI services, understanding these nuances is critical. "Not rated" doesn't mean "underperforming"—it signals incomplete public evaluation and the need to complement leaderboard snapshots with direct testing, transparency requests, and a mindful eye on release timing.

By applying these principles, organizations can better navigate the noisy, fast-moving AI model ecosystem of 2023 and 2024—and make more confident decisions about which models truly meet their needs.

References and Tools:

- LMArena Text Leaderboard with Style Control
- Hugging Face [lmarena-ai/leaderboard-dataset](https://huggingface.co/arena-ai/leaderboard-dataset)
- OpenAI, Anthropic, Google DeepMind public model release announcements and research blogs