

The fast-paced release of large language models (LLMs) over the past few years has led to an increasingly complex ecosystem of versions, announcements, and evaluations. Anyone tracking the evolution of models like OpenAI's GPT series, Anthropic's Claude, or Google's Gemini can be forgiven for noticing that some releases on timeline charts are marked as "*not rated*" or shown in *gray*. What exactly does this mean? Why are certain models excluded from public leaderboards or preference tests? And how should consumers and developers interpret these distinctions in the midst of growing release cadence, fluctuating performance metrics, and evolving evaluation tools?

The Anatomy of a Model Release: Announcements vs Verified Availability

One of the most persistent sources of confusion is the difference between **announcement date** and **verified public availability**. In the AI world:

- **Announcement Date** marks when a company publicly declares a new model version, often accompanied by marketing claims or limited previews.
- **Verified Release Date** is when the model becomes broadly accessible to external users and independent evaluators—usually via API or user interface—with sufficient transparency for performance measurement.

This distinction matters because many models are announced months before they become generally available. For example, a vendor might unveil *GPT-5.2* as the next evolution with claimed enhancements, but if the model is bundled into a gated environment or limited beta, it may not be testable by benchmarking communities.

Public timelines that track model progress tend to gray out or label these gated or unreleased versions as "*not rated*", since no objective, standardized performance data exists. This avoids conflating marketing announcements with verifiable, comparable measurements.

Case Study: GPT-5.2 Cost Implications

Consider GPT-5.2, which was recently reported on aifire.co to have roughly **40% higher inference costs compared to GPT-5.1**. Such pricing signals can be important for enterprise users balancing performance gains against operational budgets.

However, since GPT-5.2 often appears behind gated access and has not appeared on public leaderboards like LMArena's text challenge (more on this shortly), objective comparisons on speed, accuracy, or alignment remain elusive. Thus, it's marked gray and "*not rated*" in many timelines despite its announcement and reported pricing.

Evaluations and Leaderboards: Preference Testing vs Benchmarks

Raw accuracy on benchmarks isn't the only metric driving perceptions of AI progress, especially given the complex, open-ended nature of language tasks. Let's distinguish two evaluation paradigms:

1. **Blind-Vote Preference Testing** — exemplified by the LMArena text leaderboard — prioritizes human raters conducting pairwise comparisons of model outputs without knowing the model identities. This reduces bias and focuses on user preferences, conversational style, and overall helpfulness beyond raw accuracy.
2. **Task-Specific Benchmarks** — standardized datasets with ground truth answers, used to measure exact accuracy or F1 scores on well-defined tasks (e.g., reading comprehension, code generation).

LMarena’s leaderboard, which is open to models **publicly available as of August 2024**, reports user preferences, often incorporating style controls influencing response tone or directness. It includes models like Claude, ChatGPT, Gemini, Grok, and Perplexity within a shared thread, enabling side-by-side human comparisons.

However, many newly announced or gated models never make it onto LMarena’s leaderboard. Without the public availability or integration with multi-model workflows—such as those offered by Suprmind that stitch various models into a single thread for user convenience—they remain unrated or gray on timelines.

Accelerating Release Cadence Since 2023

The AI model release cadence has notably accelerated since 2023, moving from large, spaced-out version bumps to rapid, incremental updates:

- Early models were released at intervals of many months, often with substantial qualitative improvements.
- Since 2023, vendors push out minor version updates every few weeks or even days, many being experimental or incremental.
- This accelerated cadence means some releases are tightly gated or deployed behind enterprise-level restrictions before public API rollout.

Such frequency complicates evaluation. By the time a model reaches public leaderboards, newer versions may already be announced but unavailable. This dynamic increases the number of “not rated” or gray releases <https://suprmind.ai/hub/ai-models-index/> on the timeline, reflecting the gap between hype cycles and measurable availability.

Shrinking Gains per Release and Rising Regressions

Early in the AI boom, each new large language model release often showed dramatic leaps in capabilities. However, as the field matures, improvements have become more incremental, sometimes accompanied by undesirable regressions—performance drops on specific tasks or modes of interaction.



“Not rated” releases on timeline charts often correspond to gated updates focusing on internal tuning (e.g., safety, cost optimization) rather than outright quality improvements. Without public benchmarks, it is difficult to confirm that these releases represent strict gains rather than sideways or downward shifts.

The rising tolerance for regressions also reflects increasing emphasis on preferences and alignment trade-offs. For instance, a new model version might reduce hallucinations but appear less verbose or less accurate on narrow benchmarks, leading to varied user preferences captured only through blind vote tests like LMArena.

Summary of Key Factors Influencing "Not Rated" or Gray Listings

Factor	Explanation	Example/Impact
Gated Models	Limited access models not publicly testable	GPT-5.2 behind higher cost paywall, not on LMArena
Announcement vs Release Date	Announcements precede public availability	Models gray until API access confirmed
Evaluation Methodology	Preference-based tests vs benchmarks	Some models preferred but no accuracy data
Release Cadence	Faster releases with smaller incremental changes	Many minor updates—not all rated yet
Shrinking Gains and Regressions	Improvements less dramatic, sometimes regression	Some gated patches decrease specific metrics

Integrating Multiple Models in a Single Workflow: Suprmind's Multi-Model Threads

One way to mitigate the confusion around gated vs public accessibility is the emergence of multi-model workflow tools like Suprmind. Suprmind allows users to simultaneously engage Claude, ChatGPT, Gemini, Grok, and Perplexity models within a single conversational thread.

This integration aids users in making side-by-side qualitative comparisons and compensates for the lack of uniform evaluations across all models independently. However, models locked behind exclusive access still limit this harmonization and thus remain unrated on public leaderboards.

Conclusion

In summary, when you see AI model releases marked as *not rated* or shaded gray in release timelines, it's a signal to look deeper. It often reflects a gap between exciting announcements and verified, public access that supports standardized evaluation. Key contributing factors include gated access restricting public testing, faster incremental release cadences, the nuanced difference between preference-based judgment and benchmark scores, and the gradual slowing of performance leaps with possible regressions.

As of August 2024, public leaderboard platforms like LMArena provide transparent, blind-vote preference tests for only those models made publicly available, keeping the bar high for what qualifies as a "rated" release. Tools such as Suprmind's multi-model threads enrich user experience but cannot fully close the visibility gap created by gated models.

For decision-makers and developers, distinguishing between announcement hype and verified ratings is critical to making accurate assessments of AI progress in an increasingly crowded and complex landscape.



Notes:

- GPT-5.2 pricing data referenced from aifire.co, citing ~40% higher inference costs than GPT-5.1.
- LMArena leaderboard publicly started accepting models in August 2024, focusing on blind preference testing rather than raw benchmark scores.