

Short answer: In the release-cadence figures examined here, Anthropic is fastest, with a median interval of **22.5 days** between qualifying premium-model releases. Qwen follows at **45.5 days**, and OpenAI at **52.5 days**.

That ranking answers one narrow question: how often does a lab ship a new premium model? It does not tell you which model is best, which lab improves most with each release, or whether your team should upgrade. For those decisions, cadence is only the first measure. Verified availability, blind-vote preference, regressions, and cost all matter too.

There is an important caveat. Release-cadence numbers depend on what counts as a release and which dates are recorded. The figures above should be read as a tracked-sample comparison, not a permanent property of each company. A fast cadence can reflect frequent launches, but it can also reflect how a tracker groups model variants and handles preview announcements. Before using the ranking for planning, check the underlying release log and its inclusion rules.

What “fastest” means in this comparison

Counting press releases is a poor way to measure shipping speed. A company can announce a model weeks before customers can use it. It can also introduce several versions under one product name, or publish a model card before opening access broadly. Those are different events.

This comparison treats a release as a qualifying premium model that has a verifiable date of practical availability. “Premium” refers to a model positioned for paid or higher-capability use, rather than every lightweight, open-weight, or experimental variant a lab may publish. The cadence figure is the median number of days between qualifying releases in the tracked sample. A median reduces the influence of one unusually long gap or a cluster of launches.. Exactly.

Lab or model family	Reported median interval	How to interpret it
Anthropic	22.5 days	Fastest cadence in the tracked comparison
Qwen	45.5 days	About twice Anthropic’s interval
OpenAI	52.5 days	Longest interval among these three figures

These numbers are not a forecast. They do not guarantee that Anthropic will always publish a new premium model every 22.5 days, or that OpenAI will wait 52.5 days. They summarize a sample. The comparison also does not rank every AI lab: Qwen is included because it is part of the supplied cadence figures, while the central question here is how the named frontier labs compare in a practical release-tracking workflow.

Announced dates are not usable dates

For buyers, the date that matters most is usually the date the model becomes usable in the product, API, or deployment channel they actually rely on. A model may be announced at an event, exposed to a small test group, added to an API in stages, and made generally available later. Treating all those milestones as the same date makes a lab look faster or slower depending on which announcement a tracker chooses.

Suprmind Data Lab should maintain two dates for each model: the **announced date** and the **verified-availability date**. The first captures when the company publicly described the release. The second records the earliest date supported by evidence that the model was accessible to the relevant users. If access is limited to a preview, mark it as a preview rather than silently treating it as general availability.

A practical date hierarchy

1. **General availability:** The company says the model is available to the intended customer group, or its documentation and product surfaces confirm access.
2. **API or product availability:** A dated model listing, API documentation, or product update shows that users can select or call the model.
3. **Limited preview:** Access is restricted, waitlisted, or explicitly experimental. Record the restriction.
4. **Announcement only:** The lab has described the model but has not provided evidence of practical access.

For a release-cadence ranking, use the earliest verified date that matches a stated availability rule. Then publish that rule. If the rule changes—for example, from API availability to broad consumer access—the historical series should be recalculated, not mixed with the old one.



Cadence is not quality

A lab that ships more often gives customers more opportunities to try new capabilities. But more releases can also mean more migration work, more evaluation cycles, and more chances for regressions. A two-week gap between model launches is useful only if the new model improves the tasks your team cares about and behaves reliably enough to deploy.

That is why the cadence figures should sit beside, not replace, preference and task-based evaluation. One useful public signal is LMArena (formerly LMSYS Chatbot Arena), especially its **text leaderboard with style control**. The arena compares models through blind user votes, which can reveal which answers people prefer in pairwise comparisons. Style control is relevant because response length, tone, and formatting can influence a vote independently of factual accuracy or task success.



Blind-vote preference is not a complete quality measure. Users may favor a polished answer even when it contains an error. A leaderboard result can shift with the prompt mix, sample size, model availability, and the population casting votes. It is a useful signal about perceived response quality—not a substitute for your own evaluation set, safety review, latency tests, or cost analysis.

Launch benchmarks have a different limitation. A company may publish scores on selected tasks that show where a model performs well, but those scores may not match your workload. Benchmark versions, prompting rules, tool access, and evaluation methods can differ. Comparing headline scores across announcements without checking the setup invites false conclusions.

Use LMArena data carefully

Teams that want to inspect historical arena results can use the Hugging Face dataset `lmarena-ai/leaderboard-dataset`. It can help researchers examine leaderboard snapshots and related data, but it should not be treated as a ready-made release calendar. A model appearing in a dataset does not, by itself, prove the date it became generally available.

For a reproducible comparison, record the dataset version or snapshot date, the leaderboard category, whether style control is enabled, and the date on which each observation was collected. Keep the arena result separate from the release log. The first describes a preference signal; the second describes product availability.

What to record for each release

- Model name and the exact variant evaluated.
- Announcement date, preview date, and verified-availability date, where applicable.
- Availability channel: consumer product, API, enterprise deployment, or open release.
- Price and usage limits as published on the date of evaluation.
- Relevant LMArena text leaderboard result, including style-control setting and snapshot date.
- Results from the team’s own task set, including failures and regressions.

OpenAI, Anthropic, and the upgrade decision

The cadence figures suggest Anthropic has been shipping qualifying premium models more frequently than OpenAI in the tracked sample. That may matter if your team wants frequent opportunities to reassess capability. It does not mean every Anthropic release is a better upgrade, or that a longer OpenAI gap signals weaker progress. A lab may ship fewer public models while investing in a larger change, offering a different mix of product updates, or releasing capabilities through staged access.

Take the reported pricing comparison between GPT-5.2 and GPT-5.1 as an example. GPT-5.2 has been **reported as about 40% higher cost than GPT-5.1**. Treat that as a **lab claim reported by a secondary source**, not as an independently verified universal price comparison. Actual cost can vary by input and output tokens, service tier, discounts, region, and API terms. A buyer should confirm current pricing directly in the provider's official documentation before forecasting spend.

Even if the price difference is accurate for a given comparison, higher cost is not automatically a reason to reject an upgrade. The relevant question is whether the model produces enough improvement on your workload to justify the extra spend. A model that reduces human review time or handles a previously impossible task may be worth more per token. Last month, I was working with a client who wished they had known this beforehand.. A model that scores higher on a public benchmark but makes your production workflow slower may not be.

Test for regressions, not just improvements

Every upgrade evaluation should look for regressions alongside gains. A new model might write better summaries but follow structured-output instructions less consistently. It might improve reasoning on difficult prompts while becoming more verbose, slower, or less reliable at tool calls. Those differences are easy to miss if a team tests only a handful of impressive examples.

Build a small, stable evaluation set from real work. Include routine prompts, edge cases, known failure modes, and examples where the current model performs well. Compare old and new versions under the same prompts and settings. [AI model release tracker](#) Track task success, factual errors, refusal behavior, latency, formatting, and cost. Where possible, review outputs blind so evaluators do not know which model produced them.

1. Choose the tasks that matter to the business, not just tasks that appear in public benchmarks.
2. Run the current and candidate models with identical inputs, tools, and scoring rules.
3. Measure both wins and regressions, including output consistency and operational behavior.
4. Calculate cost per successful task, rather than comparing token prices alone.
5. Roll out gradually and keep a rollback path for workflows with meaningful risk.

So, which lab ships fastest?

On the reported cadence figures, **Anthropic leads at 22.5 days**, followed by **Qwen at 45.5 days** and **OpenAI at 52.5 days**. That is a useful snapshot for teams tracking how often new premium models arrive. It is not a complete scorecard for model quality or product value.

The most dependable way to use a cadence ranking is to pair it with a careful release log and a repeatable evaluation process. Separate announced dates from verified-availability dates. Treat blind-vote preference on LMArena as one signal, not a verdict. Treat launch benchmarks as claims to inspect, not proof of improvement. And before upgrading, test whether the new model improves your daily work without introducing regressions that cost more than the gains.

That is the practical answer for 2026: track who ships first, but decide who earns the upgrade by measuring what changed in daily use.