Comparing the performance of Qwen 3.5 and Gemma 4

Google DeepMind has just launched Gemma 4, its latest open weight family with sizes E2B, E4B, 26B, A4B, and 31B. Alibaba's Qwen 3.5 is already one of the most robust open families on the market, including sizes 2B, 4B, 9B, 27B, 35B-A3B, 122B-A10B, and 397B-A17B.

Google DeepMind has just launched Gemma 4 , its latest open weight family with sizes E2B, E4B, 26B, A4B, and 31B. Alibaba's Qwen 3.5 is already one of the most robust open families on the market, including sizes 2B, 4B, 9B, 27B, 35B-A3B, 122B-A10B, and 397B-A17B.

If you're choosing an open model for local agents, laptop-based inference, or self-hosted production, a useful question isn't "which family has the most outstanding model?" but rather which family wins in terms of the size you can actually run.

Currently, there are no third-party Vals-style comparison charts that include all the sizes of Gemma 4 and Qwen 3.5. Therefore, the most reliable comparison at present must distinguish between two different types of evidence:

1. Evidence of third-party chat preference, in which Google cites Arena AI's open-source text ranking.
2. Formal overlap between models-tags, where we match models by implementation class and only compare benchmark rows that both sides actually publish.

On the currently published official overlap benchmark, the Qwen 3.5 wins more rows in the 2B, 4B, and mid-size MoE classes, while the Gemma 4 is most competitive in the dense ~30B class and performs better for edge audio, multilingual, and some multimodal workloads. However, on Arena AI, both the Gemma 4 31B and Gemma 4 26B A4B rank higher than comparable large Qwen 3.5 models in terms of chat priority.

Quick comparison

class size	Gemma 4 model	Qwen 3.5 model	The best reading material available today.
Edge / mobile	Gemma 4 E2B	Qwen3.5-2B	Currently, there are no reliable third-party rankings; official results suggest **Qwen** has a slight edge.
Class 4B	Gemma 4 E4B	Qwen3.5-4B	Currently, there are no reliable third-party rankings; official results suggest **Qwen** has a slight edge.
Large Dense Models (LDM)	Gemma 4 31B	Qwen3.5-27B	The official overlap is evenly split; Arena AI chat priority leans towards **Gemma 4 31B** .

class size	Gemma 4 model	Qwen 3.5 model	The best reading material available today.
Efficient MoE	Gemma 4 26B A4B	Qwen3.5-35B-A3B	Official overlap benefits **Qwen** ; the Arena AI chat option benefits **Gemma 4 26B A4B**

Qwen also has a higher-end segment that Gemma 4 can't match: 9B, 122B-A10B, and 397B-A17B.

Why might Google's claim still be true?

If you only read the model-tag tables below, you might easily conclude that Qwen 3.5 is far superior to Gemma 4. That's too subjective. Google's launch post is presenting a different kind of evidence.

On Arena AI's open-source ranking page, March 31, 2026:

1. The Gemma 4 31B ranked third among open-ended models with a score of 1452 ± 9.
2. Qwen3.5-397B-A17B ranked 4th with a score of 1449 ± 6
3. Gemma 4 26B A4B ranked 6th with a score of 1441 ± 9
4. Qwen3.5-122B-A10B reached 1416 ± 6
5. Qwen3.5-27B reached 1404 ± 6
6. Qwen3.5-35B-A3B reached 1400 ± 6

This is third-party evidence supporting Gemma, particularly regarding Google's "byte for byte" and "intelligence-per-parameter" approach. This doesn't mean Gemma 4 beats Qwen 3.5 on every performance test. It does mean that on large-scale chat priority rankings, the two larger Gemma 4 models are currently ranking higher than the main open Qwen 3.5 models.

An objective assessment is:

1. Arena AI currently favors Gemma 4 in terms of assistant quality on a large scale.
2. The official model-card overlap shows that Qwen 3.5 leans towards Qwen 3.5 in many aspects such as static inference, programming, and the number of agent rows.
3. Third-party evidence for small models remains limited, so conclusions about 2B and 4B are still more provisional than those for large models.

The bigger trend: Open weighting is improving almost as fast as closed weighting.

A helpful way to view the Gemma 4 versus Qwen 3.5 issue is to chart the weighted best-performing open patterns and best-performing close patterns at each point in time over the past year based on a single third-party index.

In that respect, Arena AI isn't perfect, but it's useful. It measures conversation quality based on human preferences, not the accuracy of static benchmarks. That means it shouldn't replace the underlying model-tag tables. But it puts open and closed models on the same scale, which makes it a good indicator of the growth rate of advanced models.

On this index, the growth rate of the open-weighted model increased from 1398 to 1456 over the past year (+58), while the growth rate of the closed-weighted model increased from 1448 to 1504 (+56). Thus, the growth rate of the open-weighted model improved almost as fast as the closed-weighted model on this ranking, but the absolute gap remained almost unchanged: 50 points at the top of the window compared to 48 points at the bottom.

Key milestones behind the chart

Track	Become a pioneer	Model	Arena AI Score
open weight	Starting on Windows (April 2, 2025)	DeepSeek-R1	1398 ± 5
open weight	August 21, 2025	DeepSeek-V3.1	1418 ± 6
open weight	December 22, 2025	GLM-4.7	1443 ± 6
open weight	January 27, 2026	Kimi K2.5 Thinking	1453 ± 5
open weight	February 11, 2026	GLM-5	1456 ± 6
Closed weight	Starting on Windows (April 2, 2025)	Gemini 2.5 Pro Experimental	1448 ± 3
Closed weight	August 5, 2025	Claude Opus 4.1 Thinking	1449 ± 3
Closed weight	November 12, 2025	GPT-5.1 High	1455 ± 4
Closed weight	November 24, 2025	Claude Opus 4.5 Thinking	1474 ± ??4
Closed weight	February 5, 2026	Claude Opus 4.6 Thinking	1504 ± 6

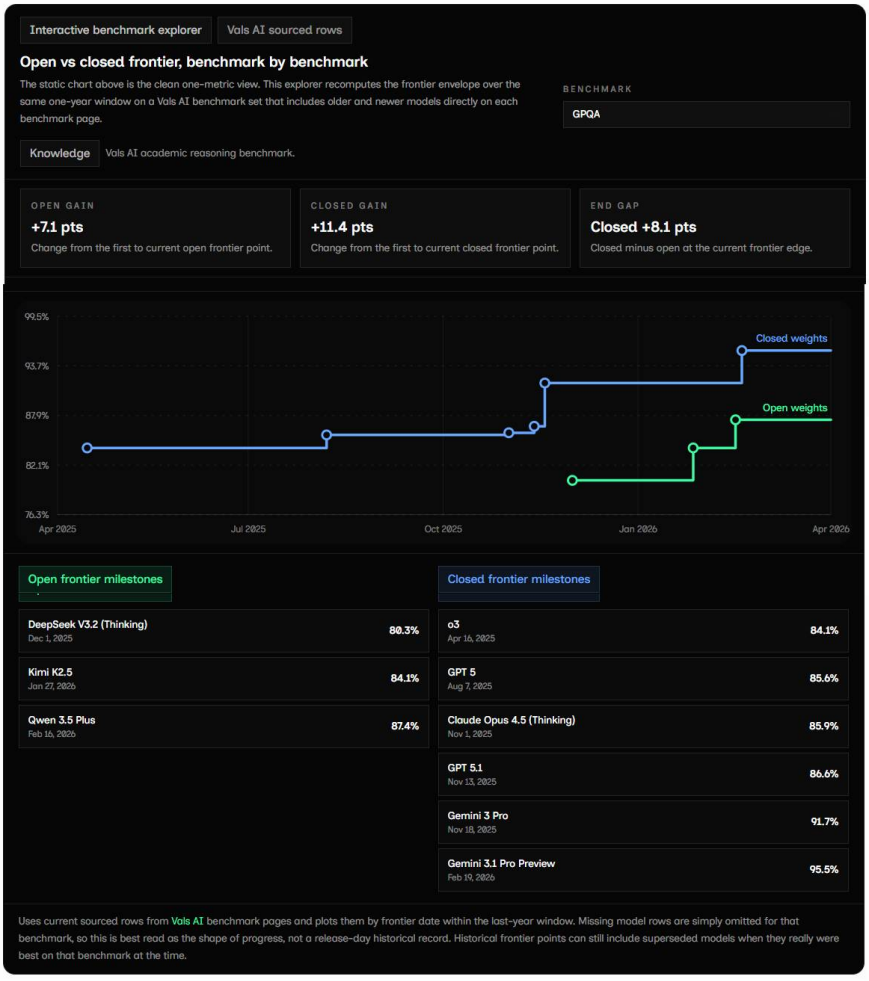
This chart clarifies three things:

1. The rate of increase in open-weighted models is no longer so slow. On this third-party chat index, the slope is now very similar.
2. The gap remains. Open-weighted models have gained nearly the same number of points, but they haven't significantly erased the lead held by closed-weighted models over the past year.
3. Much of the catch-up in open-weighted models comes from a few major releases, rather than steady monthly growth. The same is true for closed-weighted models, particularly around Claude Opus 4.5 and Claude Opus 4.6 .

Methodology Note : The chart uses Arena AI's overall text ranking as of March 31, 2026, as the sole general scale, then plots a bounding envelope by the model's release date. This is helpful for comparing speeds, but it remains an indicator based on chat preferences, not a lab-based benchmark. Small changes in scores within the stated uncertainty range should be viewed with caution. GLM-5 uses the first public launch date on Hugging Face (February 11, 2026) as the timeline because details about Z.ai's public launch are not accessible in the machine-readable post when this article is updated.

Interactions on benchmarks

The static chart above is the simplest example because it keeps the open and closed states on the same Arena AI chat priority scale. But if you want to check the durability of the results on task benchmarks, the explorer below recalculates the open and closed weights over the same 1-year period using rows taken from Vals AI, whose benchmark pages also keep rows from the older model visible.



Note : This chart uses current Vals AI benchmark pages instead of model-tag snapshots from the release date. This makes it useful for comparing the shape of progress across benchmarks, but it should not be read as a historical record from the release date. If a model is missing a row taken from a given benchmark, that milestone will not appear for that metric. Historical milestones may still include outdated models, even if they were actually the best benchmark models at the time.

How to match the models

There are two important points to note before reviewing the tables:

1. Gemma's small models use efficient parameters. Gemma 4 E2B has an efficient parameter of 2.3B / parameter loaded with embeddings is 5.1B, and Gemma 4 E4B has an efficient parameter of 4.5B / parameter loaded is 8B. These parameters are understood to be the Qwen 2B and 4B corresponding parameters at the implementation level, rather than their exact raw weighted corresponding parameters.
2. Qwen publishes different benchmark modes by size. Qwen3.5-4B, Qwen3.5-27B, and Qwen3.5-35B-A3B run in thinking mode by default on their model cards. Qwen3.5-2B publishes separate scores for thinking and non-thinking mode; this post uses the thinking mode score whenever the card displays thinking/non-thinking results.

Remove any benchmarks that are too methodologically sensitive to be considered clear, direct comparisons here, such as:

1. AIME 2026 lacks tools, as Gemma announced, but the corresponding size of Qwen is not.
2. SWE-bench verified, as Qwen announced this for larger models but not for Gemma 4.
3. CodeForces, as Qwen notes, measures its CodeForces results on its own query set, making it a less effective direct comparison to the Codeforces ELO scores published by Google.

This means the tables below should be read as reliable but limited: They are good for comparing published duplication, but not the whole story about the overall quality of the assistant.

In the tables below, read the following rows:

1. MMLU-Pro: General Knowledge and Reasoning
2. GPQA Diamond: Expert Scientific Reasoning
3. LiveCodeBench v6: Programming
4. Tau2 / TAU2-Bench: Tool/Agent Usage Behavior
5. MMMLU: Multilingual Inference
6. MMMU-Pro: Multimodal Inference

Match the size

Deployment layer	Gemma 4	Qwen 3.5	Why is this a good combination?
Edge / mobile	E2B	2B	Smallest local models
Class 4B	E4B	4B	Small laptop / edge-plus level
Large Dense Models (LDM)	31B dense	27B dense	The largest non-inference model in each family
Efficient MoE	26B A4B	35B-A3B	The nearest medium-sized MoE class, with approximately 4 billion versus approximately 3 billion operational parameters.

Should I choose Qwen 3.5 or Gemma 4?

For 2B edge implementations: Qwen3.5-2B if you care about text quality, tool usage, and long context; Gemma 4 E2B if audio and Google-edge integration are more important.

For 4B: Qwen3.5-4B. This is the easiest option in this article.

For large, non-inferential models: Qwen3.5-27B for text-prioritizing assistants and agents, Gemma 4 31B for more balanced multilingual and multimodal workloads.

For medium-sized MoEs: Qwen 3.5-35B-A3B unless your core KPI is heavy programming-intensive multilingual work and you want to specifically test Gemma 4 26B A4B.

The important thing is that the priorities at the team level are now clear enough to guide the shortlisting. But these priorities aren't one-sided: If you're interested in static task benchmarks, published overlaps often point to Qwen 3.5. If you're interested in assistant-style chat quality, the strongest signals from third parties currently indicate that Gemma 4 is in the lead.

However, the final quality still depends on your questions, context length, tool calls, and latency tolerances. That's why real-world implementation teams should treat public performance tests as a filter, not the final answer.