The ATOM Report: Measuring the Open Language Model Ecosystem

Nathan Lambert[1] and Florian Brand[1]
[1]Interconnects AI, atomproject.ai

Abstract

We present a comprehensive adoption snapshot of the leading open language models and who is building them, focusing on the ∼1.5K mainline open models from the likes of Alibaba's Qwen, DeepSeek, Meta's Llama, that are the foundation of an ecosystem crucial to researchers, entrepreneurs, and policy advisors. We document a clear trend where Chinese models overtook their counterparts built in the U.S. in the summer of 2025 and subsequently widened the gap over their western counterparts. We study a mix of Hugging Face downloads and model derivatives, inference market share, performance metrics and more to make a comprehensive picture of the ecosystem.

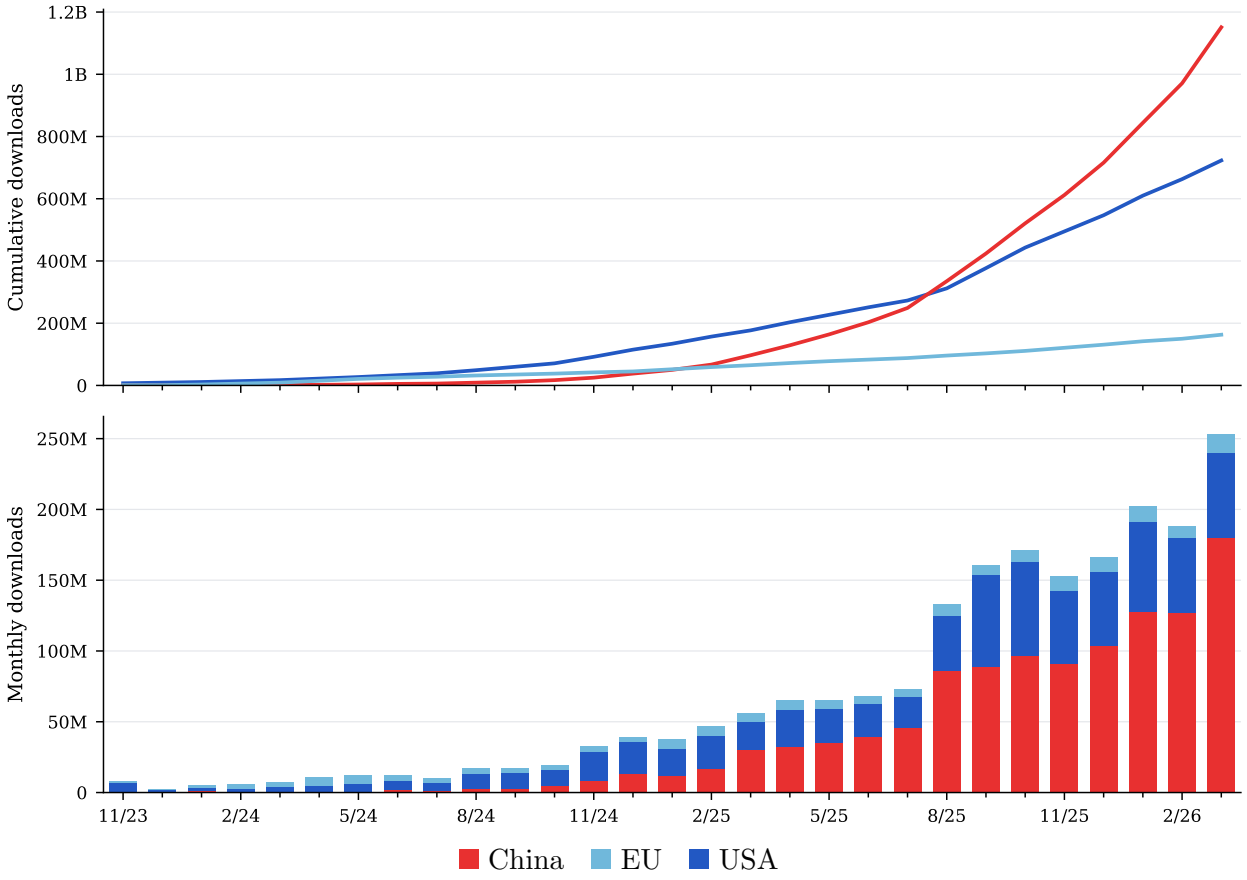

Figure 1: Cumulative open model downloads by region, November 2023 – March 2026. Chinese models overtook American models by August 2025, with the gap widening to >400M downloads.

Contents

1 Introduction

Open weight AI models are becoming foundational infrastructure across research, startups, and governments negotiating their future in understanding, building, and deploying increasingly powerful AI systems. During this time period of rapid advancements in AI capabilities and training methodologies, the role of open models is constantly evolving through different families of models and performance gaps to the best, closed frontier models. Within all of this, measuring the *adoption* of open models in parallel to the substantial work on model performance and quality (including Arena, formerly ChatBotArena (Chiang et al., 2024), Artificial Analysis's Intelligence Indices (Artificial Analysis, 2026), and Epoch AI's benchmarking (Epoch AI, 2025)), is often opaque and hard to draw insights from. In this report we detail our findings across different data sources to show the current key trends across open model usage.

The data for this report is expanded from the initial data methods used to make the case for The ATOM Project (Lambert, 2026), and it is designed to inspire more precise investment and innovation on open language models. This report does not make specific policy recommendations, for those, see The ATOM Project essay at `https://atomproject.ai/`. A key function of The ATOM Report and its broader related initiatives[1], is to focus on the specific question of the development of *open language models* specifically and not all of open-source AI collectively. This involves curating a specific subset of the public data on open models and curating the rough group of models that is defining adoption patterns – e.g. many small, specific open models such as classifiers models can dominate download or fine-tuning numbers, while having minimal impact on technological advancements.

We detail the following key findings in this report:

1. **Chinese open models moved from trailing the U.S. to a clear lead in cumulative adoption**: China overtook the U.S. in late July 2025 and reached 1.15B cumulative downloads versus 723M by March 2026, as shown in Fig. 1.

2. **Qwen is the single most-used open model family overall**: The growth of the Chinese model ecosystem can largely be attributed to Alibaba's Qwen, which is responsible for almost a billion cumulative downloads by March 2026. Other open model families such as Llama, DeepSeek or Kimi lag far behind. The growth of Qwen-based models has even accelerated especially with the updated model series of Qwen3 2507 and Qwen3.5, respectively.

3. **Other adoption metrics back up the trend**: Inference data from OpenRouter shows Chinese models taking a large lead in the summer of 2025, with the gap increasing drastically towards the end of the year. This is largely correlated with Hugging Face downloads and model derivatives. DeepSeek leads in OpenRouter measurements, where the total usage is more split across organizations relative to Hugging Face metrics.

4. **We introduce the Relative Adoption Metric (RAM)**: To showcase the momentum of model adoption beyond just raw download numbers, which are primarily dominated by initial download numbers for small models, we develop a relative metric over time and model size. We find that, among large general models, GPT-OSS 120B and Nemotron 3 Super show exceptional adoption at 10–20× their size-class median, while DeepSeek V3.2 and GLM 4.7 underperformed at well below 1× their respective baselines. This suggests the relative popularity, even long after release, of US-based models.

[1] Such as extended analysis at `interconnects.ai`. Examples are here or here.

2 Methodology

2.1 Data Sources and Collection

We rely on a variety of data sources in this report. The primary data is composed of aggregated, public data on the open model platform Hugging Face, which measures downloads of each model and highlights model derivatives (i.e. which models are used as base models for fine-tuning). We also include data from public benchmarking websites Artificial Analysis (Artificial Analysis, 2026) and Arena (formerly ChatBotArena) (Chiang et al., 2024) to showcase model performance and other adoption data from the popular open model inference platform `OpenRouter.ai`.

Hugging Face Downloads. The primary adoption signal is cumulative download counts over time from the Hugging Face Hub, which is the primary destination for open models and provides publicly accessible data. By taking daily snapshots of the total downloads of every model, we can showcase trends of data not directly on `huggingface.co`. A "download" is defined as any HTTP request to the model file hosting endpoint, including programmatic access via `transformers`, `curl`, and similar tools[2]. We obtained historical monthly download data directly from the Hugging Face team for the leading open model organizations – Meta Llama, Qwen, Mistral AI, DeepSeek, Google, and Microsoft – covering November 2023 through July 10, 2025. Beginning in early July 2025, we operate an independent daily scraper that records total cumulative downloads and other public metadata for all publicly available models. The historical series (through July 10, 2025) uses 2.5× IQR outlier-filtered data (described below). From August 2025 onward, we compute monthly deltas from the unfiltered scraper data and add them to the filtered baseline, preserving continuity at the splice point without re-filtering the full history. Throughout, monthly labels refer to data as of the first of that month: "Aug 25" denotes cumulative downloads through July 31, 2025.

Throughout the report we refer to open models, open language models, and open-weight language models. The final term is the most precise, as we are measuring the adoption of any set of model weights available on the Hugging Face platform, regardless of license or information available about it (which could earn it the classification "open-source").

We track every prominent open language model released since ChatGPT – i.e. models that accept text (and optionally other modalities) as input and produce text as output – while excluding embedding models, text-to-image diffusion models, and other non-generative architectures. The full list of ∼1.5K tracked models is publicly available.[3] In total, these models account for over 3 billion downloads across the study period (November 2023 through March 2026). To mitigate noise on the initial data, we applied 2.5× interquartile range outlier detection on daily download series; models without anomalous spikes are left unfiltered. This substantially improved the historical data, which is prone to large variations and unexplainable features.

For derivative and fine-tune tracking, we identify derivatives via the Hugging Face `base_model` tag, which records the parent model in the format `base_model:ORG/MODEL`. We restrict to derivatives whose base model appears in our tracked model list, require more than five lifetime downloads, and exclude local-inference re-uploads (e.g., GGUF, MLX).

OpenRouter. Inference token share data is drawn from OpenRouter, which provided historical data on the top 10 open models per month ranked by total tokens served. Only the top 10 are reported for each month, so organizations whose usage is spread across time or many models may

[2] See the Hugging Face documentation for more details

[3] `https://github.com/Interconnects-AI/tracked-models/blob/main/models.csv`

be undercounted. This provides a complementary demand signal, capturing which models people use, which is not reflected in download counts alone.

Arena. Aside from quantitative metrics, we also take into account qualitative metrics to showcase the progress of open models. Community Elo ratings come from Arena (formerly ChatBotArena), which measures general chat quality through blind pairwise comparisons by human evaluators (Chiang et al., 2024). We use the style-controlled overall Elo as the primary performance metric. Scores prior to May 19, 2025 are shifted upward by 59.2 points to account for a platform-wide recalibration when style control was set as default (Li et al., 2024). For each region, we log the highest-scoring open model at each snapshot date to showcase the regional performance frontier; scores are enforced to be monotonically non-decreasing (despite scores fluctuating slightly due to the statistical variance in Arena's platform) so that the frontier only advances.

Artificial Analysis. The Overall Intelligence Index from Artificial Analysis aggregates scores across standard academic benchmarks for leading language models. We extract the top open model per region over time and fit a simple linear regression to highlight performance trends, beginning from models released after April 2024.

2.2 Model Categorization

Size Buckets. As size is one of the most important factors to select the right model, we partition models into seven parameter-count categories: <1B, 1–5B, 7–9B, 10–50B, 50–100B, 100–250B, and 250B+. We seperate the 7–9B range from the broader 1–10B span because it captures a natural concentration of popular model architectures (Llama 8B, Mistral 7B, Qwen 7B). This is also apparent in the data (see Figure 2), where this bucket alone accounts for one third of all downloads, making it the single largest bucket, followed closely by the 1–5B bucket. The remaining buckets represent natural partitions for models which have emerged over time, e.g. there have been standard model sizes of 32B and 70B parameters for multiple years in the ecosystem due to popular hardware configurations. For MoE models, we use total parameter count rather than active parameters per token—e.g., DeepSeek-R1 (671B total, 37B active) is placed in the 250B+ bucket. Finer-grained analysis of MoE model sizes and sparsity relative to dense models is left to future work.

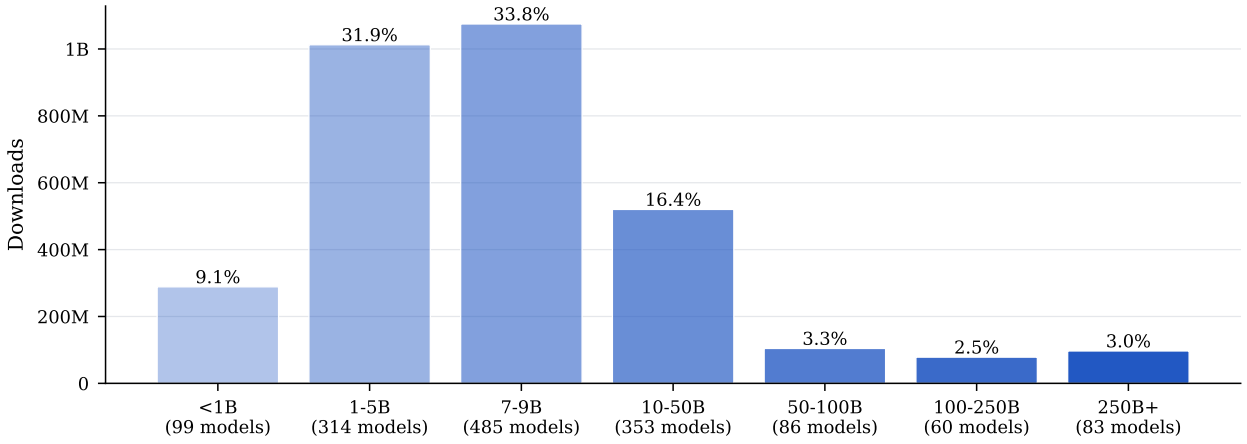

Figure 2: Distribution of our tracked open model downloads by parameter count. The 7–9B range captures 33.8% of all downloads, with sub-10B models accounting for ∼75%.

Regional Classification. Models are attributed to one of three regions—United States, China, or Europe—based on the headquarters of the releasing organization. This is an imperfect proxy, as it does not capture the nationality of individual contributors or even their organizations. The mapping is as follows:

- **United States:** Meta (Llama), Google (Gemma), Microsoft (Phi), NVIDIA (Nemotron), OpenAI (GPT-OSS), Allen AI (Olmo), IBM (Granite), Snowflake, AI21 Labs
- **China:** Alibaba (Qwen), DeepSeek, ByteDance, Baidu, InternLM, Zhipu AI (GLM/ChatGLM), OpenBMB, Inclusion AI, Skywork, Tencent, Xiaomi (MiMo), Moonshot AI (Kimi), MiniMax
- **Europe:** Mistral AI, Hugging Face (SmolLM)

2.3 Limitations

Hugging Face is not the sole distribution channel for open models, and meaningful traffic can flow through other platforms such as ModelScope. Large deployments of open models can also count as just one download if a user downloads the model to local infrastructure and then never relies on the public versions. The latter is also the case for other services, such as clouds or other SaaS products, which might cache model weights for production usage. These factors might result in under reporting the popularity of certain model (families). While this decentralization makes it impossible to pinpoint exact numbers, the larger trends and the relation of the models to each others largely hold outside of this tracked usage. The exclusion of re-packaged models in GGUF or MLX format from derivative counts avoids inflating fine-tune metrics with local-inference re-uploads, which are especially prevalent for small, popular models.

As for the metrics for usage, download counts do not equate to active usage, as automated CI/CD pipelines, bots, and repeated pulls inflate total downloads numbers, particularly for small models. This effect might be stronger for models with "non-standard" architectures, as (small) models of a new architecture might be used in pipelines and tests to represent all current (and future) models of the same architecture.

However, we find that all available usage data is strongly correlated with the imperfect Hugging Face download metrics and show the same trend(s) and similar distribution.

3 Model Adoption by Region

The core measurement of this report is following the balance of influence internationally among open model builders. The overall rate and cumulative total of model downloads, binned per nation/region among the three largest contributors to open models – The United States, China, and Europe generally – is shown in Fig. 1.

We observe three different phases: After the release of ChatGPT, the EU dominated in terms of model adoption, exclusively driven by Mistral 7B (Jiang et al., 2023) and Mixtral 8x7B (Jiang et al., 2024). This was followed by a dominance of US-based models with the release of Llama 3 (Grattafiori et al., 2024), which has seen a lot of adoption due to its permissive license and the range of different sizes. Following the release of DeepSeek V3 (DeepSeek-AI et al., 2025), R1 (Guo et al., 2025) and Qwen3 (Yang et al., 2025), the lead in usage went towards Chinese model builders, which have increased the gap since then. In each of these eras, the respective country was responsible for >50% of downloads and fine-tuned models (see Figure 1 for downloads and Figure 3 for derivatives). Cumulative tracked downloads reached 2.04 billion by March 2026, a 6× year-over-year increase from 339M in March 2025, highlighting the explosion of growth in open models. China grew 11.9× YoY (97M to 1.15B) versus 4.1× for the USA (177M to 723M) and 2.5× for the EU (65M to 163M), widening the total download gap from 23M at August 2025 to 428M by March 2026.

The share of new model derivatives – fine-tunes and adaptations built on base models – follows the same pattern as downloads: China rose from 10% in November 2023 to 70% by February 2026, while the EU fell from a peak of 58% to 4% over the same period.

Open model inference data from OpenRouter shows the same shift with an even bigger magnitude (Figure 4): Chinese models' token share rose from 2.8% to over 70% in 14 months, while US-based models have shrunken considerably.

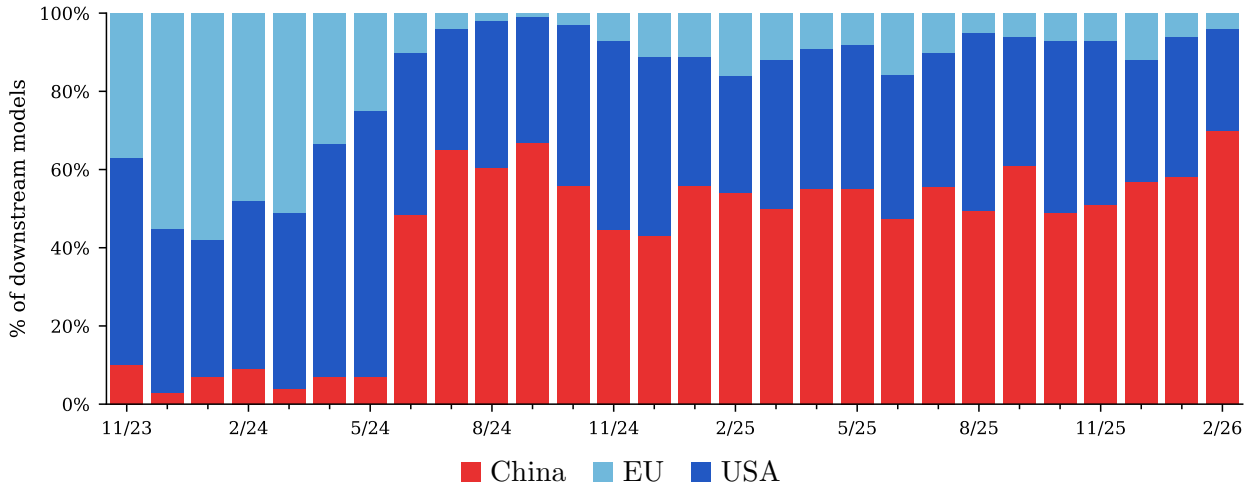

Figure 3: Share of new model derivatives by region per month. EU share fell from 58% in January 2024 to 4% by February 2026, while China rose to 70%.

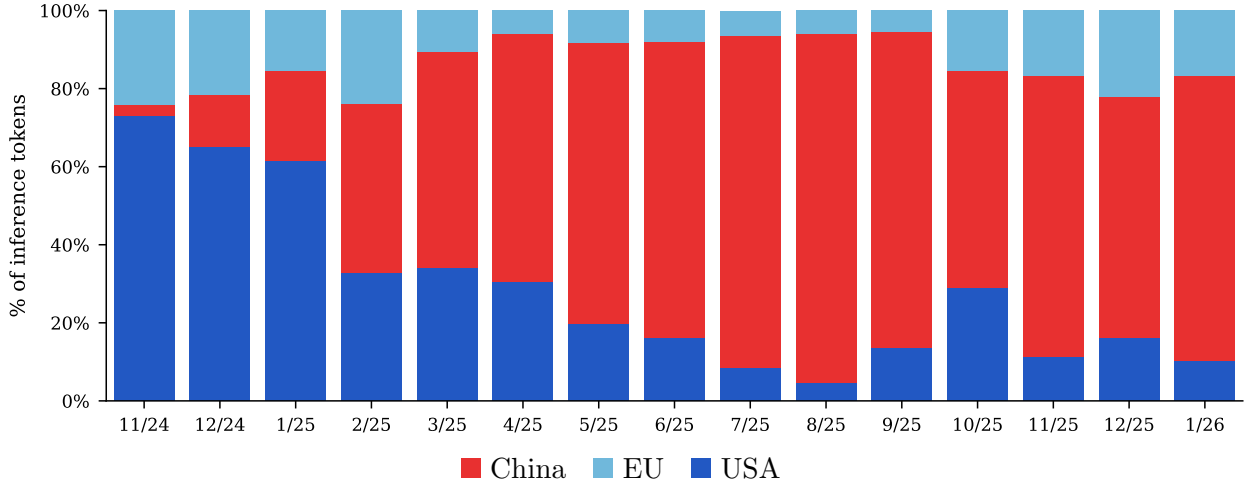

Figure 4: Open model inference token share by region (OpenRouter). China and the US completely inverted positions in 14 months, with China reaching 72.7% by January 2026.

4 Model Performance By Region

To track the qualitative performance of leading open models over time, we rely on existing public benchmarking projects such as Arena (Chiang et al., 2024) and Artificial Analysis's Intelligence Index (Artificial Analysis, 2026). For this, we took the open model subset of these leaderboards over time corresponding and plotted them per region with the categorization from Section 2.2. Both benchmarks show a similar trend to the quantitative metrics, with the leading Chinese models overtaking the *capabilities* of the leading American models late in 2024, holding a lead since. However the lead is less profound as in the quantitative metrics, which also makes sense – a small lead in model performance can create a large lead in adoption, as uses of open models by default opt for the best model.

The evolution of these benchmarks is shown for Arena's overall data in Fig. 5 and for Artificial Analysis's Index in Fig. 6.

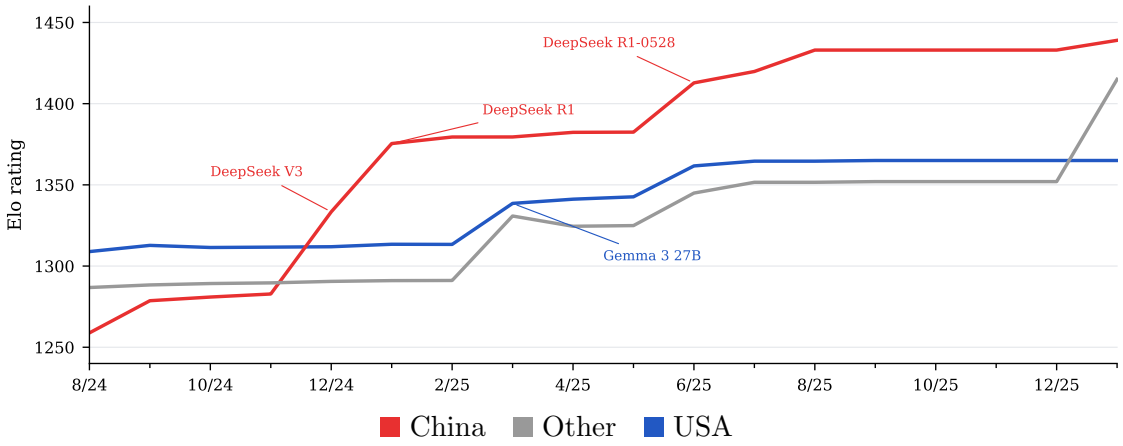

Figure 5: Arena Elo ratings for top open models by region. China surpassed the US in December 2024, driven by DeepSeek V3, and extended its lead through January 2026.

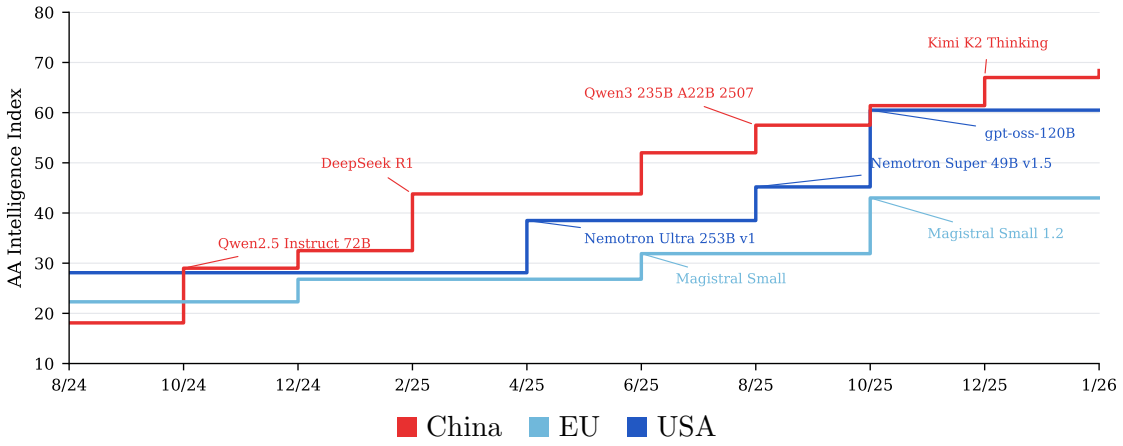

Figure 6: Artificial Analysis Overall Intelligence Index by region. China's best open model score quadrupled in 18 months, reaching 68.4 by January 2026.

5 Model Adoption by Organization

To complement our analysis of open model performance and adoption by region, we also present the data split into individual organizations. We balance the analysis between the evolution of the leading open model organizations, tracking the rise of Qwen, with the latest, emerging labs releasing excellent open models to challenge the incumbents.

5.1 Ecosystem Leaders (Qwen, Llama, DeepSeek, Mistral, and OpenAI)

Figure 7 shows cumulative downloads for five representative model families: Alibaba's Qwen, Meta's Llama, Mistral, DeepSeek, and OpenAI. Qwen surpassed Llama in cumulative downloads in September 2025 (325.4M vs. 323.7M) and by March 2026 reached 942.1M, thus nearly doubling the downloads of the Llama models (476.0M). This change is even more stark in terms of derivatives, i.e., fine-tunes and other adaptations (such as LoRA adapters), where Qwen became the primary choice as a base model as early as June 2024. Qwen's share of new fine-tunes and adaptations rose from 1% in January 2024 to 69% by February 2026 (Figure 8), while Meta's peaked at 44% in August 2024 before falling to 11%.

This data also shows that new entrants can have impactful adoption, e.g. with OpenAI's new GPT-OSS models accumulating more adoption than long-established open model organizations such as Mistral AI.

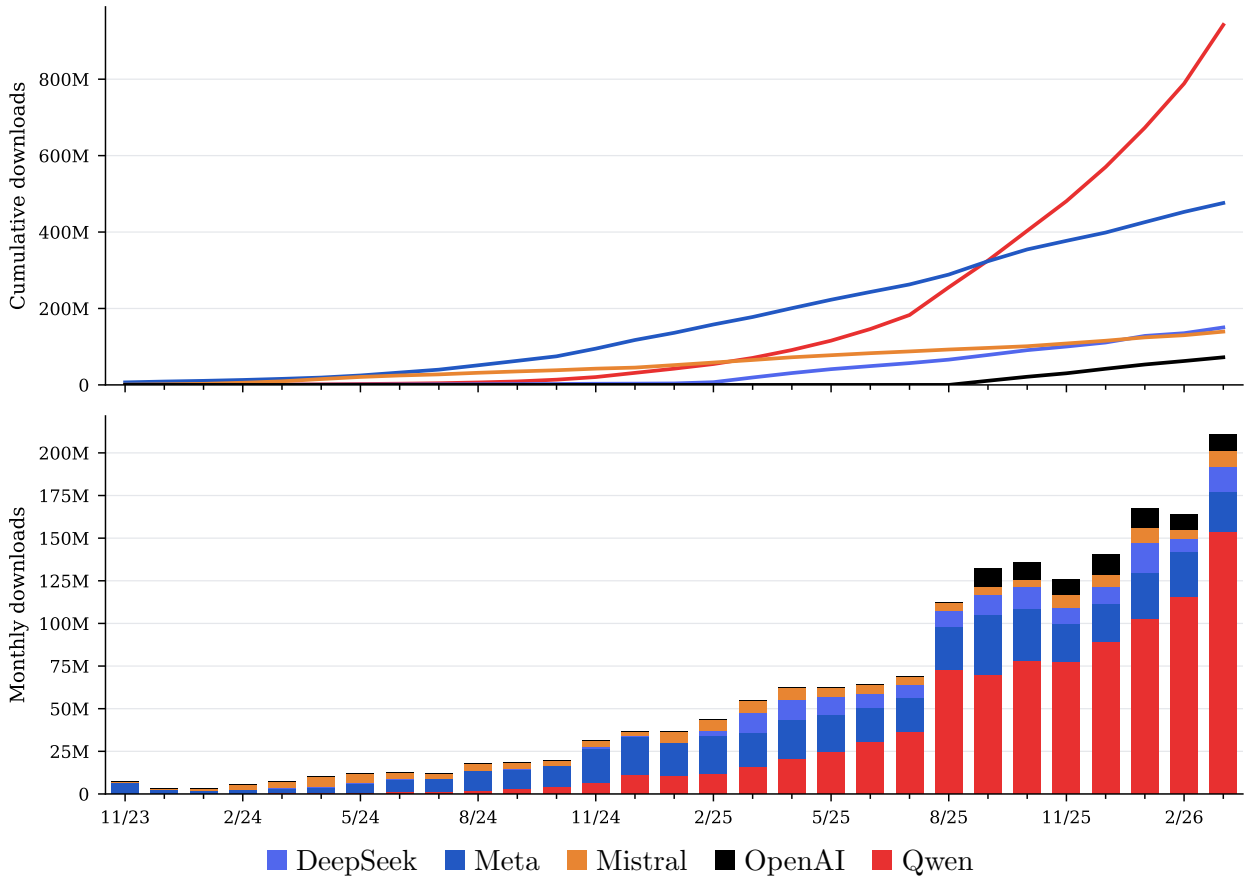

Figure 7: Cumulative downloads for leading open model families. Qwen surpassed Llama in September 2025 and reached 942.1M downloads by March 2026.

Qwen's path to leading the ecosystem. The data of both model derivatives and downloads over time show how the ecosystem has evolved through multiple eras of model defaults. The first two iterations of Meta's Llamas and Mistral's early models were the first platforms of the ecosystem through most of 2023. Later, Meta's initial Llama 3 models were released in April of 2024, with the more popular 3.1 variants coming in July – spurring Meta to an all-time lead in model derivatives in the second half of 2024. These models carried Meta to an all-time downloads lead in Q1 of 2025, but Qwen was accelerating its rate of adoption. A pivotal moment was the release of Qwen 2.5 (Qwen et al., 2025) in September of 2024, which has started the rise of Qwen in derivative models, which took substantial market share from Meta's Llama models and Mistral's early successful models (7B Dense and Mixtral 8x7B). The outlier in June 2024 is mostly due to spam, which persisted through our filtering of models with at least 5 life-time downloads.

Other prominent releases across the ecosystem, from Google's Gemma 2 (Gemma Team et al., 2024) in July of 2024, Gemma 3 (Gemma Team et al., 2025) in March of 2025, and many continued Mistral releases left their adoption metrics stagnant relative to Qwen's growth. Qwen 3 in April 2025 (Yang et al., 2025) (the same month as Llama 4's release) continued Qwen's acceleration. The early adoption numbers of Qwen3.5 (Qwen Team, 2026) since its release in February 2026 are an indication that the dominance of Qwen relative to its peers will continue (see Section 6 and specifically Fig. 15 for early data on Qwen 3.5's adoption).

Over time, an equilibrium from May 2025 through March 2026 settled with Qwen having a base of 40% or more of the derivative share, growing slowly, and the remainder being split predominantly between Meta, Google, Mistral, DeepSeek, and a long-tail of smaller labs. Despite few model releases over the last year, Meta's Llama and Google's Gemma models maintained about 10% share of derivative models uploaded to Hugging Face each, substantially more than any others.

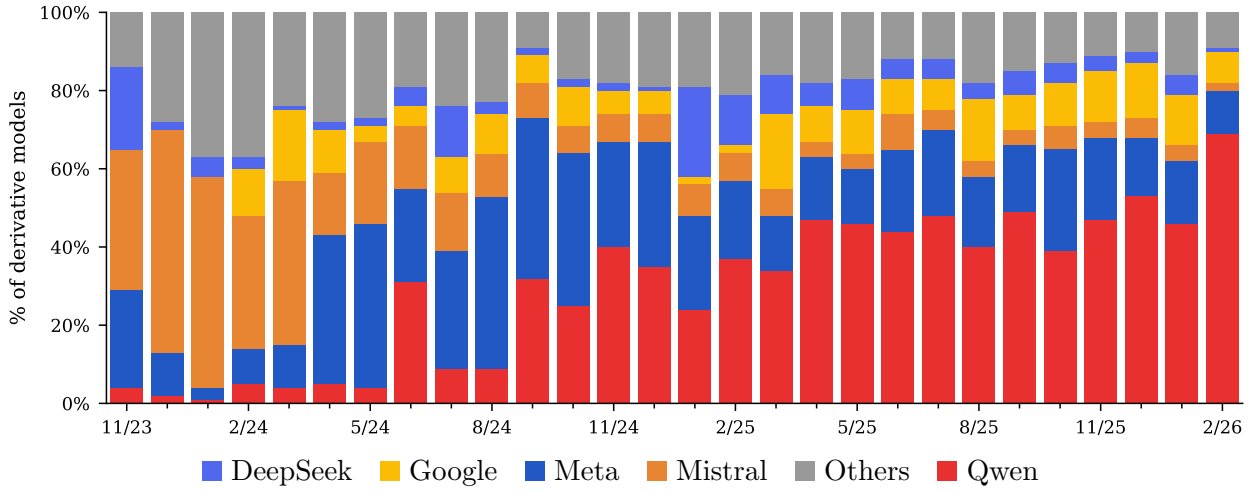

Figure 8: Monthly share of new model derivatives by organization. Qwen's derivative share reached 69% by February 2026, while Meta's Llama fell from 25% in November 2023 to 11%.

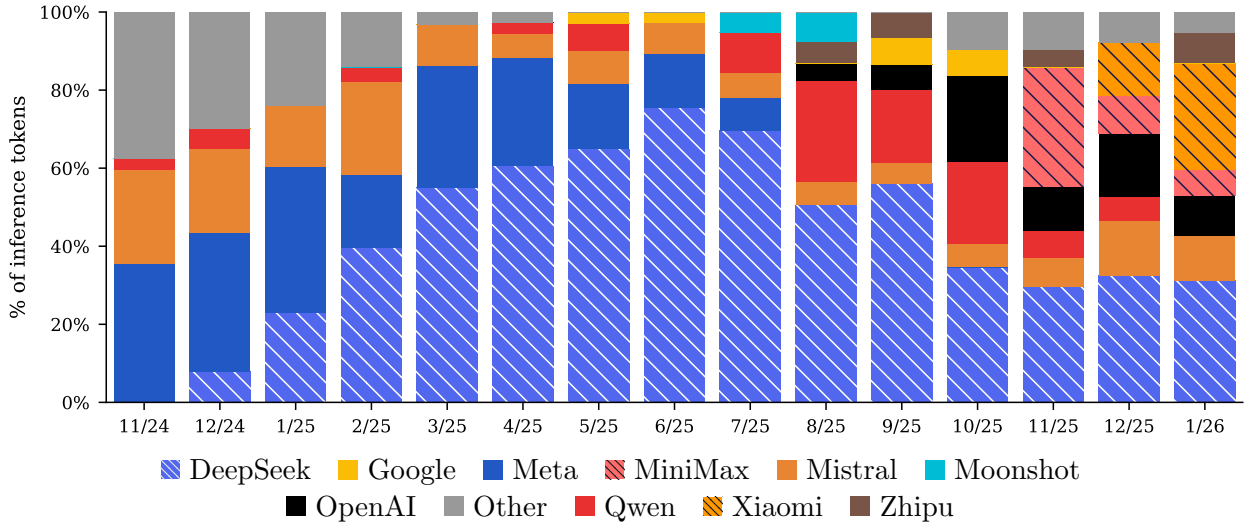

Figure 9: Open model inference token share by organization (OpenRouter). Meta fell from a 37.4% peak in January 2025 to zero, replaced by DeepSeek (31.1%) and Xiaomi (27.2%) by January 2026.

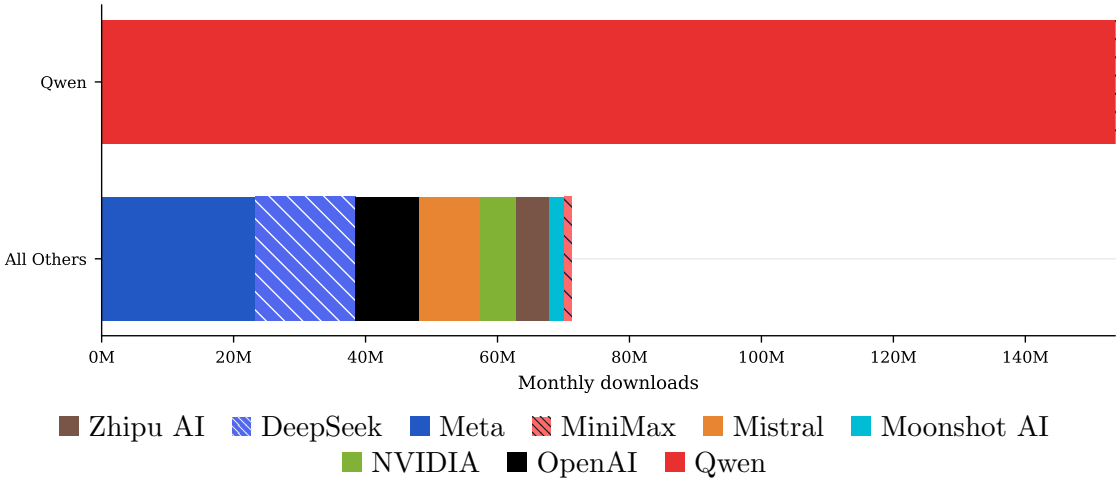

Figure 10: Monthly Qwen downloads compared to combined downloads from all other major organizations. In February 2026, Qwen alone generated 153.6M downloads, more than double the combined 71.2M from the next eight major organizations.

DeepSeek and Qwen's complementary roles. Alibaba's Qwen and DeepSeek are the two most prominent open model families built in China, each having built critical acclaim through different styles of building and releasing models. Qwen is the adoption leader by releasing numerous models across a range of sizes and abilities, where DeepSeek releases the most used, large MoE models.

Figure 10 shows that in February 2026, Qwen alone generated 153.6M monthly downloads – more than double the combined 71.2M from the 8 other, leading open model builders. This effect might be exaggerated in its magnitude as Qwen3.5 (Qwen Team, 2026) was released in February 2026, but the comparison remains accurate for other timeframes (e.g. Qwen had 87.5M downloads in December 2025, relative to the 61.3M of Meta Llama, DeepSeek, OpenAI, Mistral, Nvidia, Z.ai,

Kimi, and MiniMax combined). The driver of this gap is Qwen's smaller models – just six small models of the Qwen3 series (out of 66 Qwen3 models in the complete family), i.e., Qwen3-0.6B – Qwen3 8B have as many monthly downloads as six leading model organizations *combined*, namely Zhipu AI, MiniMax, Mistral, Moonshot, NVIDIA and OpenAI (Figure 11).

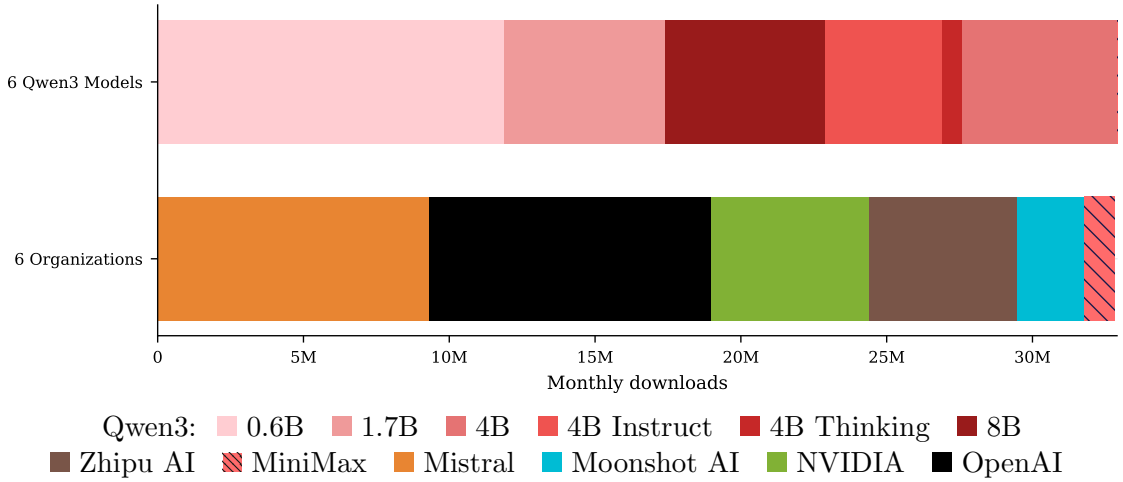

Figure 11: Monthly downloads for selected Qwen 3 models, including their quantized variants (0.6B–8B) versus competing organizations. In February 2026, six Qwen 3 models combined for 32.9M downloads, roughly the same amount as the 32.8M generated by six competing organizations.

DeepSeek, by contrast, has a major adoption lead relative to Qwen in the lifetime use of the largest models. Figure 12 shows that DeepSeek captures 47% of total tracked downloads in the 250B+ segment, while Qwen leads sub-10B models with 44%. These large model sizes represent the only model class where Qwen does not have a significant lead in adoption metrics.

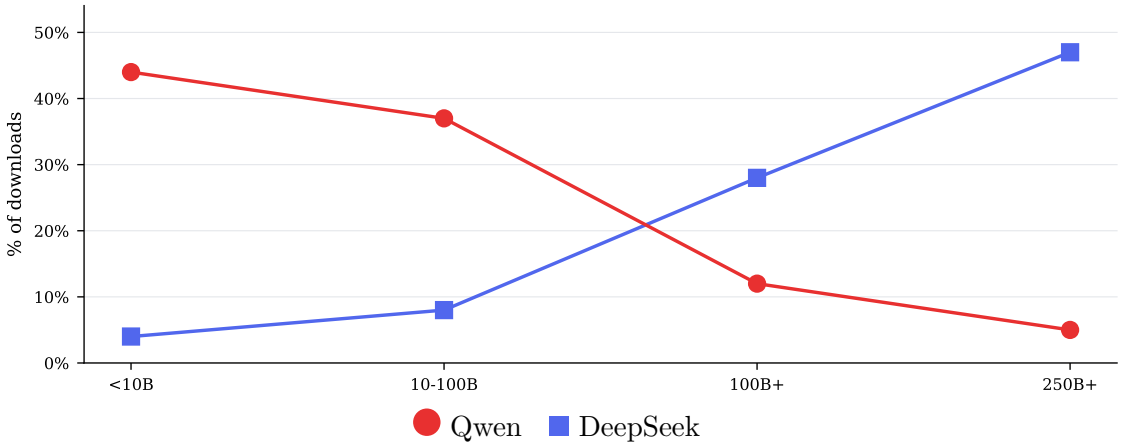

Figure 12: Download share by model size for DeepSeek vs. Qwen. DeepSeek dominates the 250B+ segment (47%) while Qwen leads sub-10B (44%), showing complementary specialization.

Summary. Beyond Qwen's clear dominance in adoption metrics, the remaining lead model families each tell a distinct story:

- **Mistral** was the earliest leader of the open language model ecosystem, highlighted by derivative

share, peaking at 57% in December 2023 on the strength of Mistral 7B (Jiang et al., 2023) and Mixtral (Jiang et al., 2024), but cumulative downloads of the models never accelerated to match Qwen or Llama's success.

- **DeepSeek** overtook Mistral in cumulative downloads in January 2026 (128.2M vs. 124.3M). Its impact is larger than downloads suggest: on OpenRouter, DeepSeek V3 (DeepSeek-AI et al., 2025) and R1 (Guo et al., 2025) accounted for up to 75.6% of inference tokens in June 2025 and still held 31.1% by January 2026 (Figure 9), reflecting heavy usage of a small number of flagship models rather than a broad derivative ecosystem.
- **OpenAI** is the most recent entrant, releasing GPT-OSS (OpenAI et al., 2025) models starting in September 2025. By spring 2026, OpenAI's monthly downloads from a handful of models have surpassed those of Mistral's entire portfolio of historical models.
- **Meta** remains second in cumulative downloads (476.0M) but following the release of Llama 4 has collapsed on inference platforms and plateaued in adoption metrics, falling from a 37.4% inference token share peak in January 2025 to zero by August 2025, replaced by a rotating cast of Chinese model providers.

5.2 New Entrants (Nvidia, Moonshot AI, MiniMax, Z.ai, ...)

Crucially, many of the labs known for their frontier-level, large MoE models, such as MiniMax, Moonshot AI, or Z.ai, are far from dominating the adoption metrics of open models. These organizations are competing at a fraction of the adoption level of the leaders, but they have meaningful growth that is worth watching.

The summer of 2025 was characterized by a wave of very capable open models from a variety of Chinese organizations. This series of impressive, Chinese models prompted the writing of the original The ATOM Project memo in August 2025. Since, we tracked American labs as they attempted to catch up with more useful open models. Figure 13 tracks the newer and smaller American entrants via cumulative downloads since August 2025: NVIDIA leads with 30.7M cumulative downloads driven by its Nemotron family (NVIDIA et al., 2024, 2025), followed by the Allen Institute for AI (Ai2) at 14.8M with OLMo (Team Olmo et al., 2025) and IBM at 8.6M with Granite (Granite Team, 2024). The scale gap remains large – all US entrants combined account for roughly 56M downloads versus Qwen's 942.1M – but their growth trajectories show sustained progress.

Figure 14 captures a broader set of organizations competing with DeepSeek, OpenAI, and Mistral as potential new organizations with top 5 overall downloads, plotted since July 2025, when we expanded our data collection tooling to encompass every organization. Examples here include Hugging Face (SmolLM), MiniMax (M Series), and Moonshot AI (Kimi models).

Inference data from OpenRouter reveals additional entrants not captured in download metrics: Xiaomi's MiMo-V2-Flash (Core Team et al., 2026), a 309B-parameter MoE released in December 2025, surged from zero to 27.2% of inference token share by January 2026, illustrating how quickly popular models can appear on inference platforms without a corresponding footprint in Hugging Face downloads.

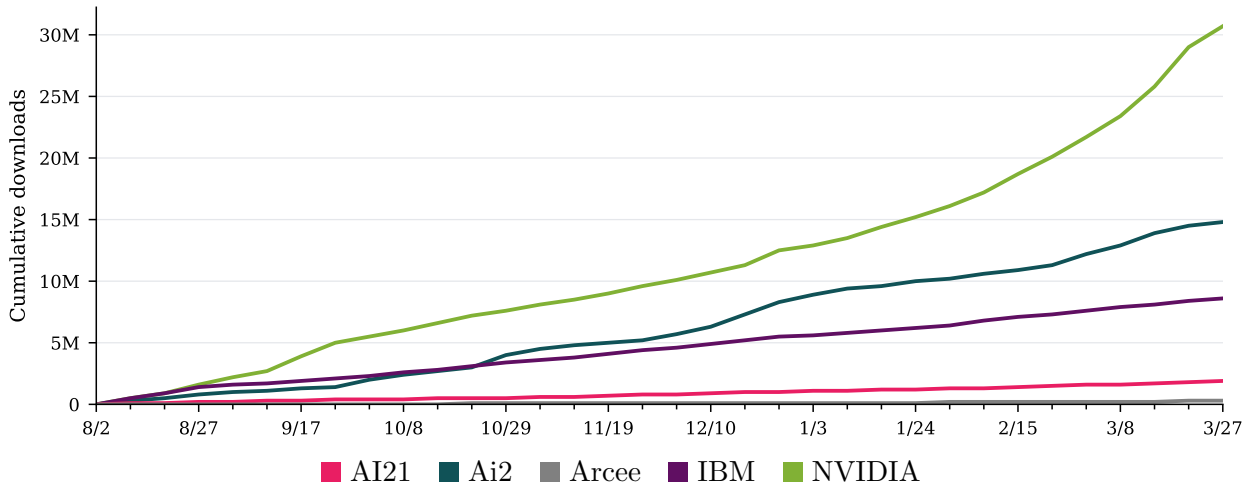

Figure 13: Cumulative downloads from new American open model entrants since August 2025. By March 27, 2026, NVIDIA reached 30.7M downloads, Ai2 14.8M, and IBM 8.6M, still far behind Qwen's 942.1M cumulative downloads.

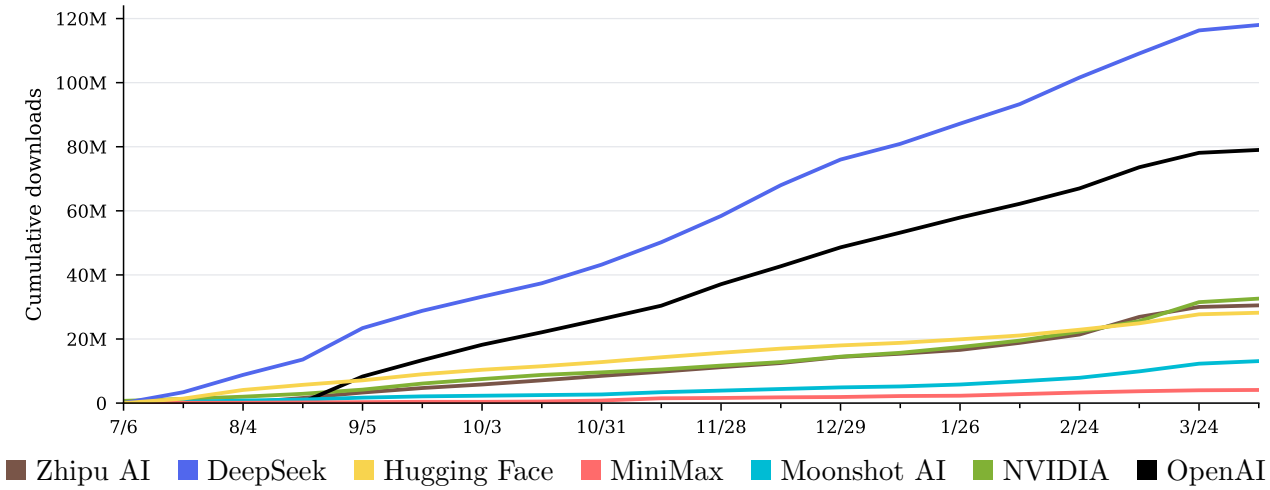

Figure 14: Cumulative downloads since July 2025 for organizations competing with DeepSeek. DeepSeek leads with 118.0M downloads, ahead of OpenAI at 79.0M.

6 The Relative Adoption Metric (RAM)

6.1 Method

Standard practice for measuring the impact of open models is to compare total and monthly download numbers to those of other models. These rough metrics don't account for the variable levels of adoption across model size categories and relative adoption of a new model as the overall adoption of open models grows rapidly. We set out to design a new adoption metric, building on a finer-grained version of the same download data, to better assess models across size categories and time.

For example, it is intuitively simple that a 1.5B model routinely accumulates 10–50× more downloads than a 400B model simply because it is cheaper to run, easier to integrate into CI/CD pipelines, and more frequently loaded in automated testing. We introduce the **Relative Adoption Metric (RAM)** to normalize adoption trajectories within size cohorts. RAM is particularly useful for medium-to-large models, where download numbers can be more precisely contextualized against a small set of peer models.

Definition. For a given model m in size bucket b (same as in Sec. 2.2, with more information in Fig. 2), at milestone t days post-release:

$$\mathrm{RAM}(m,t) = \frac{D(m,t)}{\widetilde{D}_{10}(b,t)} \tag{1}$$

where $D(m,t)$ is the cumulative download count for model m at t days after release, and $\widetilde{D}_{10}(b,t)$ is the median cumulative downloads of the top 10 most-downloaded models in bucket b at the same milestone (the top models used for the initial analysis are documented in Appendix B). A RAM score of 1.0× means the model is tracking the top-10 median for its size class; values above 1.0× indicate out-performance.

Milestones. Scores are computed at seven fixed post-release milestones: 7, 14, 30, 60, 90, 180, and 365 days. Early milestones (7–30 days) capture the initial launch momentum, while later milestones reflect sustained community interest and adoption.

Statistical Design. The reference set for each of the seven size buckets comprises the 10 most-downloaded models in that bucket. We report the median and interquartile range (IQR, 25th–75th percentile) rather than the mean and standard deviation. This choice is motivated by extreme right-skew: in the 1–5B bucket, a single breakout model can pull the mean upward by roughly an order of magnitude relative to the median. Using the median produces reference values that are robust to such outliers and that give an intuitive baseline for interpreting new model performance. Full reference statistics are provided in the supplementary materials; representative reference curves are shown in Figure 17, and case studies for recent models are presented in Figure 15.

Each RAM score is tied to a specific snapshot corresponding to when the top-10 models per size category were pulled. We will periodically update our reference groupings as we study RAM on new models.

Additional details and data on RAM are included in Appendix C.

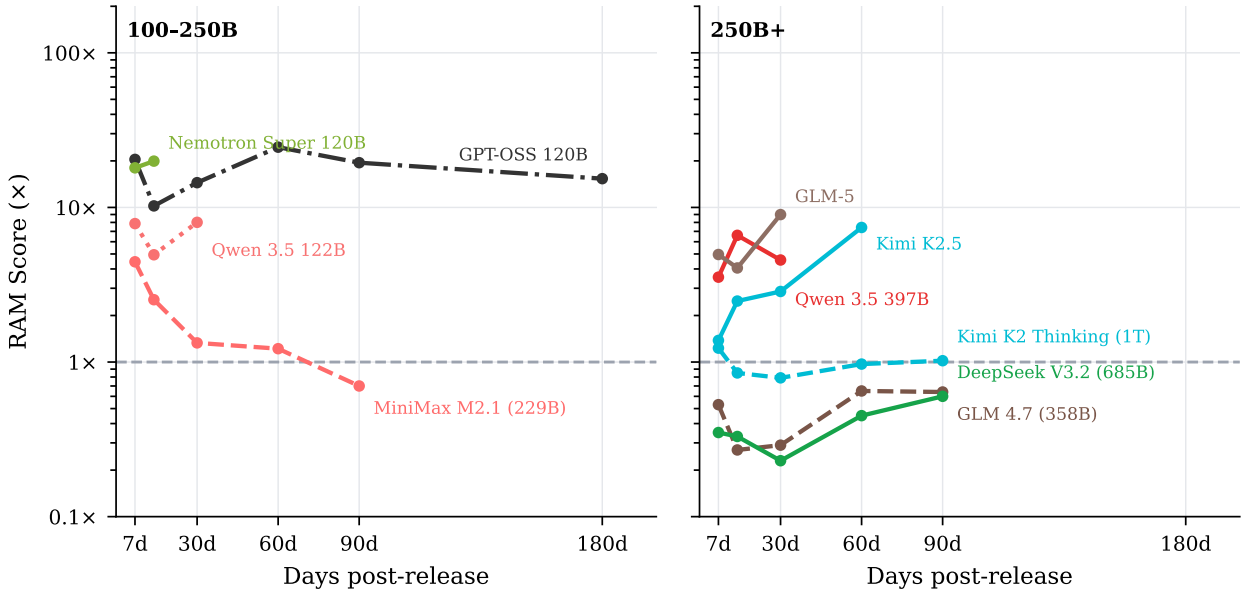

Figure 15: RAM trajectories for notable recent model releases. Each line shows the RAM multiplier over cumulative downloads relative to the top-10 median for that size bucket; 1.0× is the bucket median. Where official Hugging Face releases include multiple published weight-format variants, downloads are aggregated across those variants. Reference date: 2026-04-02.

6.2 Using RAM to Measure New Models

RAM shows a clear ranking of recent models over time within as little as 30 days following the model release. We tested RAM on a variety of open models released in the last 6 months to showcase models that are over- or under-performing on adoption metrics relative to community excitement.

For example, using an April 2, 2026 snapshot of Hugging Face data, the RAM framework makes the February 2026 Qwen3.5 rollout easier to compare across scales. In the 1–5B bucket, shown in Fig. 16, `Qwen3.5-4B` reached 3.45×, 5.29×, and 3.27× the bucket median at 7, 14, and 30 days. This is a strong launch, but not a category outlier because small-model baselines are already high: the 1–5B median was 48K downloads at 7 days, 142K at 14 days, and 722K at 30 days. For comparison, DeepSeek OCR (3B) still launched higher within the same bucket, reaching 6.3× and 9.42× at 7 and 14 days. The clearest breakout model was `Qwen3.5-35B-A3B`, which reached 16.13× at 7 days, 11.10× at 14 days, and 4.54× at 30 days, placing it among the hottest launches in our reference set. In the 100–250B bucket, shown in Fig. 15, `Qwen3.5-122B-A10B` reached 7.86×, 4.94×, and 8.01× at 7, 14, and 30 days, a notably stronger 30-day profile than its early launch suggested. At the giant end, `Qwen3.5-397B-A17B` reached 3.54×, 6.60×, and 4.57× at 7, 14, and 30 days.

Outside of Qwen 3.5 and the aforementioned models, there are clear leading and lagging models released in recent months. In 100–250B, GPT-OSS 120B and Nemotron Super 120B are clear outliers following their launches, approaching levels to be some of the most-downloaded models of all time: GPT-OSS reached 20.45× at 7 days and was still at 15.35× at 180 days, while Nemotron Super 120B opened at 18.03× and 19.93× at 7 and 14 days. A counter example is MiniMax M2.1, which launched strongly but faded back toward the bucket median, moving from 4.45× at 7 days to 0.70× at 90 days.

In 250B+, GLM-5 was a strong recent launches, reaching 4.96×, 4.06×, and 8.99× scores at 7, 14, and 30 days, while GLM 4.7 and DeepSeek V3.2 stayed below median throughout their

post-release monitoring window. Other models include Kimi K2 Thinking, which mostly tracked around the bucket median, or Kimi K2.5, which accelerated from 1.38× at 7 days to 7.42× at 60 days.

Further data for this section is included in Appendix C and specifically Table 1.

7 Conclusion

The ATOM Report is a first step in creating a focused set of tools for understanding the relative adoption of the leading open language models. Core to the methodology is a goal to bring both finer grained measurement and historical analysis to an area that traditionally has been defined by very crude and noisy measurements. The data reinforces many key trends across the ecosystem, from Qwen's growing lead to Llama's stagnation and the role of many new entrants. Crucially, this data is all one small step towards a clearer picture, and the ecosystem needs to continue to refine its data sharing and transparency, in order to better showcase potential new entrants on the value of releasing a new type of open-weights model.

Acknowledgments. We thank Hugging Face and OpenRouter for sharing private data that made this analysis possible, and Artificial Analysis, Arena, and Epoch AI for making valuable data publicly available. Thank you to Caithrin Rintoul for invaluable support for this project.

References

Adem Ait, Javier Luis Cánovas Izquierdo, and Jordi Cabot. On the suitability of hugging face hub for empirical studies. *Empirical Software Engineering*, 30(2), 2025. doi: 10.1007/s10664-024-10608-8. URL https://doi.org/10.1007/s10664-024-10608-8.

Artificial Analysis. Intelligence benchmarking methodology, 2026. URL https://artificialanalysis.ai/methodology/intelligence-benchmarking. Accessed 2026-03-17.

Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. The foundation model transparency index. *arXiv preprint arXiv:2310.12941*, 2023. URL https://arxiv.org/abs/2310.12941.

Rishi Bommasani, Kevin Klyman, Shayne Longpre, Betty Xiong, Sayash Kapoor, Nestor Maslej, Arvind Narayanan, and Percy Liang. Foundation model transparency reports. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:181–195, 2024. doi: 10.1609/aies.v7i1.31628. URL https://doi.org/10.1609/AIES.V7I1.31628.

Rishi Bommasani, Kevin Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej, and Percy Liang. The 2024 foundation model transparency index. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=38cwP8xVxD.

Joel Castaño, Silverio Martínez-Fernández, Xavier Franch, and Justus Bogner. Analyzing the evolution and maintenance of ML models on Hugging Face. In *Proceedings of the 21st International Conference on Mining Software Repositories*, pages 607–618, 2024. doi: 10.1145/3643991.3644898. URL https://doi.org/10.1145/3643991.3644898.

Wei-Lin Chiang et al. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 8359–8388. PMLR, 2024. URL https://proceedings.mlr.press/v235/chiang24b.html.

Core Team et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026. URL https://arxiv.org/abs/2601.02780.

DeepSeek-AI et al. Deepseek-v3 technical report. 2025. URL https://arxiv.org/abs/2412.19437.

Epoch AI. Epoch capabilities index, 2025. URL https://epoch.ai/benchmarks/eci. Accessed 2026-03-17.

Gemma Team et al. Gemma 2: Improving open language models at a practical size. 2024. URL https://arxiv.org/abs/2408.00118.

Gemma Team et al. Gemma 3 technical report. 2025. URL https://arxiv.org/abs/2503.19786.

Avijit Ghosh, Lucie-Aimée Kaffee, Yacine Jernite, and Irene Solaiman. State of open source on hugging face: Spring 2026, 2026. URL https://huggingface.co/blog/huggingface/state-of-os-hf-spring-2026.

Granite Team. Granite 3.0 language models, 2024. URL https://github.com/ibm-granite/granite-3.0-language-models/blob/main/paper.pdf.

Aaron Grattafiori et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Daya Guo et al. Deepseek-r1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z. arXiv preprint: https://arxiv.org/abs/2501.12948.

Albert Q. Jiang et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Albert Q. Jiang et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. URL https://arxiv.org/abs/2401.04088.

Nathan Lambert. The atom project, 2026. URL https://www.atomproject.ai/. Project website. Accessed 2026-03-17.

Tianle Li, Anastasios Angelopoulos, and Wei-Lin Chiang. Does style matter? Disentangling style and substance in Chatbot Arena. https://lmsys.org/blog/2024-08-28-style-control/, August 2024. LMSYS Org blog post, published August 29, 2024.

Shayne Longpre et al. The data provenance initiative: A large scale audit of dataset licensing & attribution in AI. *arXiv preprint arXiv:2310.16787*, 2023. URL https://arxiv.org/abs/2310.16787.

Shayne Longpre et al. Consent in crisis: The rapid decline of the AI data commons. *arXiv preprint arXiv:2407.14933*, 2024. URL https://arxiv.org/abs/2407.14933.

Shayne Longpre, Christopher Akiki, Campbell Lund, Atharva Kulkarni, Emily Chen, Irene Solaiman, Avijit Ghosh, Yacine Jernite, and Lucie-Aimée Kaffee. Economies of open intelligence: Tracing power & participation in the model ecosystem. *arXiv preprint arXiv:2512.03073*, 2025. URL https://arxiv.org/abs/2512.03073.

Nestor Maslej et al. Artificial intelligence index report 2024. *arXiv preprint arXiv:2405.19522*, 2024. URL https://arxiv.org/abs/2405.19522.

Nestor Maslej et al. Artificial intelligence index report 2025. *arXiv preprint arXiv:2504.07139*, 2025. URL https://arxiv.org/abs/2504.07139.

NVIDIA et al. Nemotron-4 340b technical report. *arXiv preprint arXiv:2406.11704*, 2024. URL https://arxiv.org/abs/2406.11704.

NVIDIA et al. Nemotron 3 nano: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning, 2025. URL https://arxiv.org/abs/2512.20848.

OpenAI et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025. URL https://arxiv.org/abs/2508.10925.

Qwen et al. Qwen2.5 technical report. 2025. URL https://arxiv.org/abs/2412.15115.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Team Olmo et al. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

Alexander Wan, Kevin Klyman, Sayash Kapoor, Nestor Maslej, Shayne Longpre, Betty Xiong, Percy Liang, and Rishi Bommasani. The 2025 foundation model transparency index. *arXiv preprint arXiv:2512.10169*, 2025. URL https://arxiv.org/abs/2512.10169.

An Yang et al. Qwen3 technical report. 2025. URL https://arxiv.org/abs/2505.09388.

A Related Work

A handful of other groups have done work collecting and analyzing subsets of the data we discuss in this report. In the closest related work, Longpre et al. (2025) measure the participation dynamics across the entire ecosystem, i.e. not solely focused on key language models since ChatGPT, and reaches many overlapping conclusions as this report.[4] The AI Index provides annual ecosystem-level indicators for releases, capabilities, and deployment trends (Maslej et al., 2024, 2025). Transparency-centered efforts evaluate disclosure practices rather than adoption outcomes, including the Foundation Model Transparency Index and related reporting frameworks (Bommasani et al., 2023, 2025; Wan et al., 2025; Bommasani et al., 2024). Platform-focused studies of Hugging Face characterize repository growth, maintenance behavior, and data quality of the platform itself (Ait et al., 2025; Castaño et al., 2024). Governance and data-provenance work adds complementary evidence on licensing, attribution, and the contraction of the open data commons (Longpre et al., 2023, 2024).

B Top 10 Downloaded Models by Size Category

We include the top 10 models per size category as used for the construction of the Relative Adoption Metric (RAM), as documented in Sec. 6. Download counts are as of late March 2026.

Under 1B

1. Qwen3-0.6B (72.8M)
2. Qwen2.5-0.5B-Instruct (32.3M)
3. Florence-2-large (19.4M)
4. t5gemma-b-b-prefixlm (17.5M)
5. Qwen2.5-0.5B (17.1M)
6. Qwen2.5-Coder-0.5B-Instruct (13.5M)
7. Qwen2-0.5B (10.7M)
8. SmolLM2-135M (10.5M)
9. Florence-2-base (8.8M)
10. llava-onevision-qwen2-0.5b-ov-hf (8.4M)

1-5B

1. Qwen2.5-1.5B-Instruct (150.6M)
2. Qwen2.5-VL-3B-Instruct (70.7M)
3. Qwen2.5-3B-Instruct (70.1M)
4. Llama-3.2-1B-Instruct (55.7M)
5. Llama-3.2-1B (49.2M)
6. Llama-3.2-3B-Instruct (37M)
7. gemma-3-1b-it (34.9M)
8. Qwen2-VL-2B-Instruct (30M)
9. Qwen3-4B (29.7M)
10. Qwen3-1.7B (29.2M)

7-9B

1. Llama-3.1-8B-Instruct (133M)
2. Qwen2.5-7B-Instruct (109M)
3. Mistral-7B-Instruct-v0.2 (53.5M)
4. Qwen2.5-VL-7B-Instruct (51.2M)
5. Qwen3-8B (42.5M)
6. Meta-Llama-3-8B-Instruct (38.9M)
7. Meta-Llama-3-8B (36.6M)
8. Llama-2-7b-chat-hf (29.2M)
9. Llama-2-7b-hf (28.4M)
10. falcon-7b-instruct (26.7M)

[4] Also see related, recurring work in this direction from Hugging Face directly (Ghosh et al., 2026).

10-50B

1. gpt-oss-20b (54M)
2. Qwen2.5-14B-Instruct (33.3M)
3. Qwen3-32B (24.6M)
4. DeepSeek-R1-Distill-Qwen-32B (23M)
5. Mixtral-8x7B-Instruct-v0.1 (20M)
6. Qwen2.5-32B-Instruct (18.5M)
7. Llama-3.2-11B-Vision-Instruct (17.6M)
8. Llama-2-13b-chat-hf (15.2M)
9. Qwen3-VL-30B-A3B-Instruct (13.2M)
10. gemma-3-27b-it (12.3M)

50-100B

1. Llama-3.1-70B-Instruct (20.2M)
2. Qwen3-Next-80B-A3B-Instruct (14.6M)
3. Llama-3.3-70B-Instruct (10.3M)
4. InternVL3-78B (6.2M)
5. Meta-Llama-3-70B-Instruct (5.9M)
6. Qwen2.5-VL-72B-Instruct (5.7M)
7. Qwen2.5-72B-Instruct (5.4M)
8. Llama-2-70b-chat-hf (4.6M)
9. DeepSeek-R1-Distill-Llama-70B (4.3M)
10. Meta-Llama-3-70B (3.2M)

100-250B

1. gpt-oss-120b (29.2M)
2. Mixtral-8x22B-Instruct-v0.1 (6M)
3. Mistral-Large-Instruct-2407 (5M)
4. Mistral-Large-Instruct-2411 (4.9M)
5. Mixtral-8x22B-v0.1 (4.8M)
6. InternVL3_5-241B-A28B-Instruct (4.1M)
7. Qwen3-235B-A22B (3.4M)
8. Qwen3-VL-235B-A22B-Thinking (3.3M)
9. Qwen3-235B-A22B-Instruct-2507-FP8 (2.7M)
10. MiniMax-M2 (1.9M)

250B+

1. Llama-3.1-405B (20.3M)
2. DeepSeek-R1 (16.7M)
3. DeepSeek-V3 (14.3M)
4. DeepSeek-R1-0528 (5.9M)
5. Kimi-K2.5 (5.4M)
6. GLM-5-FP8 (4.9M)
7. DeepSeek-V3-0324 (4M)
8. Llama-3.1-405B-Instruct (3.4M)
9. Qwen3.5-397B-A17B (2.1M)
10. Kimi-K2-Instruct (1.8M)

C Additional RAM Details

The top-10 model download counts over time, used to compute the RAM scores, is shown in Figure 17. This shows that among the top few models in each size category, the median of top-10 downloads over the first 180 days is remarkably similar across buckets. The smallest models have larger outliers, as shown in Appendix B, where models such as Qwen3-0.6B with 72.8M downloads, Qwen2.5-1.5B-Instruct with 150.6M, or Llama-3.1-8B-Instruct with 133M have substantially more downloads than the 10th top model in those categories (llava-onevision-qwen2-0.5b-ov-hf with 8.4M, Qwen3-1.7B with 29.2M, and falcon-7b-instruct with 26.7M).

Figure 16 shows RAM trajectories for a few small and medium models (1–5B and 10–50B), complementing the 100–250B and 250B+ panels in the main text. Table 1 reports exact RAM scores and cumulative downloads at each milestone for all case-study models.

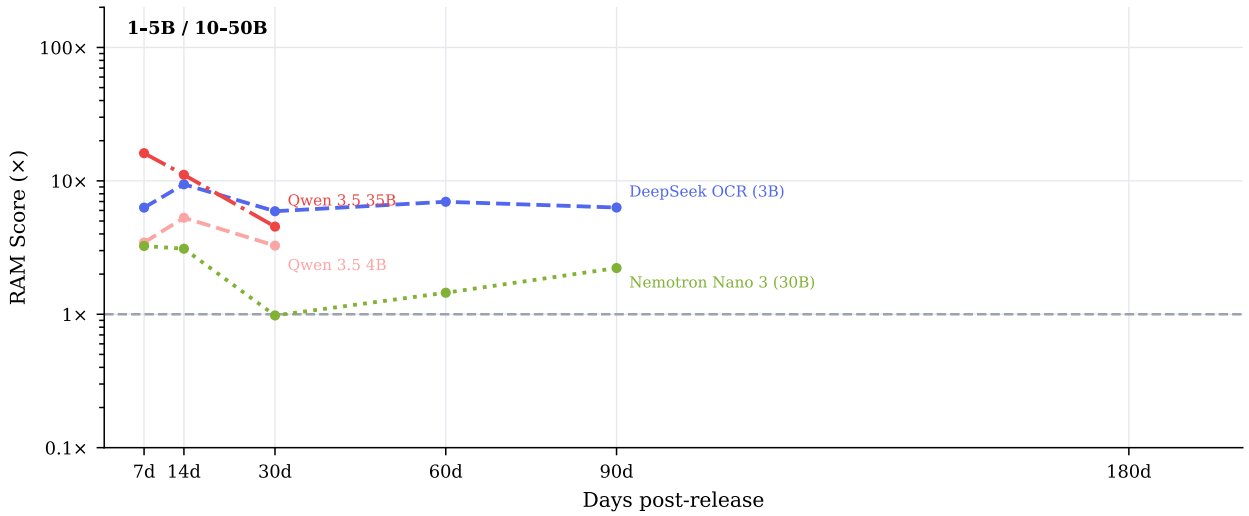

Figure 16: RAM trajectories for small and medium models (1–5B and 10–50B buckets). DeepSeek OCR (3B) sustained 6–9× the bucket median through 90 days, while Qwen 3.5 35B had the hottest launch at 16.1× at 7 days before normalizing. Reference date: 2026-04-02.

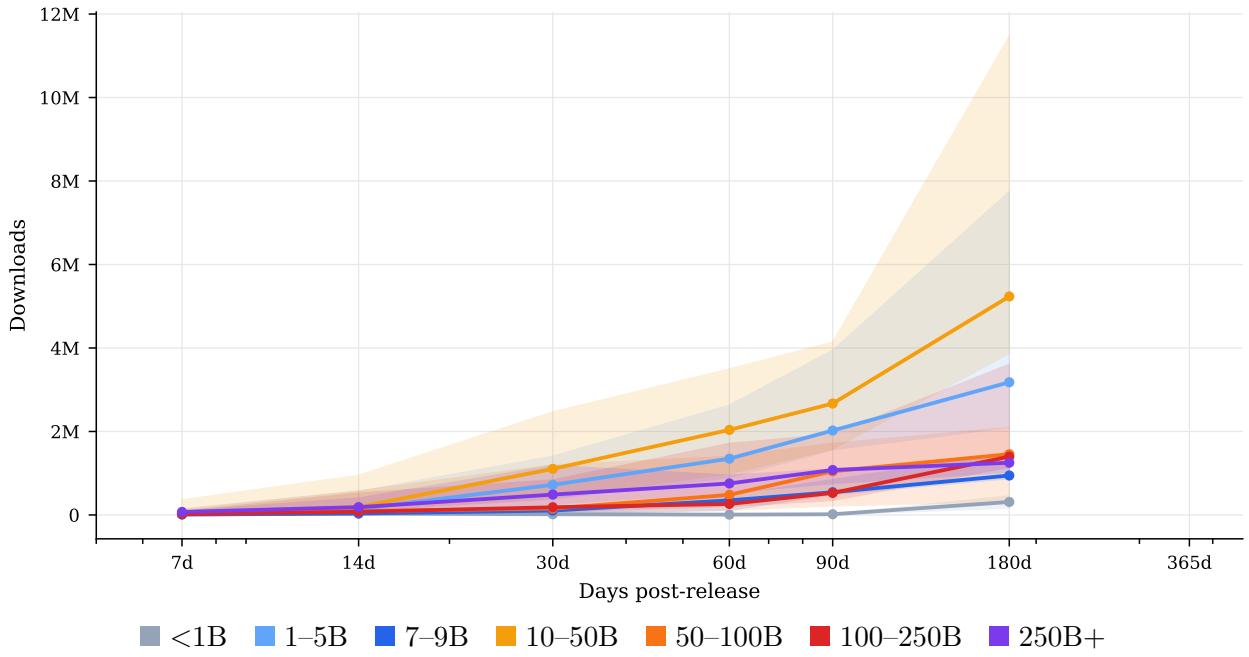

Figure 17: RAM reference curves showing median cumulative downloads (with IQR) for the top 10 models in each size category over time. The 1–5B and 7–9B ranges show the highest adoption baselines.

Table 1: RAM scores (top row, × median) and cumulative downloads (bottom row) for a set of case-study models at each post-release milestone. Empty cells indicate the model has not yet reached that milestone. Reference date: 2026-04-02.

Bucket	Model	7d	14d	30d	60d	90d	180d
1–5B	DeepSeek OCR (3B)	6.30× 302K	9.42× 1.3M	5.92× 4.3M	6.97× 9.4M	6.31× 12.8M	
	Qwen 3.5 4B	3.45× 166K	5.29× 751K	3.27× 2.4M			
10–50B	Qwen 3.5 35B	16.13× 822K	11.10× 2.1M	4.54× 5.0M			
	Nemotron Nano 3 (30B)	3.25× 166K	3.10× 585K	0.98× 1.1M	1.45× 3.0M	2.22× 5.9M	
100–250B	GPT-OSS 120B	20.45× 429K	10.23× 788K	14.46× 2.7M	24.53× 6.5M	19.47× 10.3M	15.35× 21.5M
	Nemotron Super 120B	18.03× 379K	19.93× 1.5M				
	Qwen 3.5 122B	7.86× 165K	4.94× 381K	8.01× 1.5M			
	MiniMax M2.1 (229B)	4.45× 93K	2.53× 195K	1.33× 246K	1.22× 323K	0.70× 372K	
250B+	GLM-5	4.96× 362K	4.06× 759K	8.99× 4.4M			
	Qwen 3.5 397B	3.54× 258K	6.60× 1.2M	4.57× 2.2M			
	Kimi K2.5	1.38× 100K	2.48× 463K	2.86× 1.4M	7.42× 5.6M		
	Kimi K2 Thinking (1T)	1.23× 90K	0.85× 159K	0.79× 385K	0.97× 732K	1.02× 1.1M	
	GLM 4.7 (358B)	0.53× 39K	0.27× 51K	0.29× 144K	0.65× 491K	0.64× 685K	
	DeepSeek V3.2 (685B)	0.35× 25K	0.33× 61K	0.23× 114K	0.45× 339K	0.60× 648K	