Open-Weight AI Model Licenses: Commercial Use Rules Explained

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · open weight models · open source ai license · llama license · apache 2.0 license · mit license · qwen license · ai model licensing · commercial use ai models · enterprise ai compliance

Executive Summary

Open-weight artificial intelligence (AI) models, meaning releases whose trained parameter files can be downloaded and run by anyone, now come from at least eight major vendors actively shipping commercially relevant large language models (LLMs) as of September 2026: Meta (**Llama**), Alibaba Cloud (**Qwen**), Zhipu AI / Z.ai (**GLM**), Moonshot AI (**Kimi**), **MiniMax**, **Mistral AI**, Tencent (**Hunyuan**), and **NVIDIA** (Nemotron). None of these families is licensed identically, and the difference between "open weight" and "open source" is not cosmetic. The Open Source Initiative (OSI) states plainly that "Open Weights differ significantly from Open Source AI because they do not include" (opensource.org) the training code, data, and documentation an Open Source Definition (OSD) license requires, and Stanford's Institute for Human-Centered AI (HAI) likewise cautions that open-weight releases "are not necessarily fully open source, as the underlying code or training data is often withheld" (hai.stanford.edu).

Commercial usability splits along two axes: license family and scale. Some flagship weights, including **Mistral 3**, **Qwen3**'s largest checkpoints, **GLM-4.5** and **GLM-4.6**, **MiniMax-M1** ship under the unmodified **Apache 2.0** or **MIT** license. NVIDIA Nemotron licenses vary by release; for example, NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16 is governed by OpenMDW-1.1 (NVIDIA NGC). Others carry a bespoke, non-OSI community license that permits commercial use up to a defined scale and then requires a separate agreement. For any multimodal Llama 4 model, the incorporated policy withholds the Community License grant from an EU-domiciled individual or a company with its principal place of business in the EU, except for downstream end users of a product or service that incorporates the model (Llama 4 Acceptable Use Policy). Separately, the Llama 4 license applies its 700-million-MAU test on the Llama 4 version release date: it covers the monthly active users in the preceding calendar month of products or services made available by or for the licensee or its affiliates; if that number exceeds 700 million, the licensee must request a license from Meta (Llama 4 Community License). Alibaba's Qwen License Agreement (covering Qwen2.5-72B) sets the same threshold at **100 million** MAU (huggingface.co), a figure Tencent's Hunyuan Community License tests only on the Tencent Hunyuan version release date (detailed below). Moonshot AI's Kimi K2 adds an attribution duty, not a licensing bar, once a licensee exceeds 100 million MAU or **$20 million in monthly revenue** (huggingface.co); MiniMax-M2 has a comparable attribution condition at 100 million MAU or **$30 million in annual recurring revenue** (raw.githubusercontent.com). Kimi K3 instead requires a separate Moonshot agreement before commercial use when the licensee or its affiliates operate a Model as a Service business and their aggregate revenue exceeds $20 million over any consecutive 12 months; that condition does not apply to internal use or use through Moonshot AI's official products or certified inference partners (Kimi K3 License).

Beyond scale thresholds, licenses diverge on redistribution mechanics: naming and "Built with X" attribution clauses appear across Llama, Qwen, Kimi, and MiniMax; a clause barring use of a model's outputs "to improve any other AI model" appeared in Meta's original Llama 2 license before being narrowed in Llama 3.1 and Llama 4 (raw.githubusercontent.com), and a similar restriction persists in Tencent's current Hunyuan license, which additionally excludes the European Union, United Kingdom, and South Korea from its territorial scope (raw.githubusercontent.com). Separately, Mistral AI dual-tracks releases: its Mistral 3 family is Apache 2.0, while other models ship under its Non-Production License (MNPL), which bars any commercial supply of the model outright pending a separate grant (mistral.ai) (mistral.ai). For a managed offering, identify the terms for the selected checkpoint and offering rather than treating hosted-service terms as the license for downloadable weights. Amazon Bedrock says its serverless third-party models may be subject to additional terms or a seller's end-user license agreement, while its Marketplace offerings follow a separate route to the applicable agreement (AWS Bedrock third-party-model terms). Azure AI Foundry says users accept license terms for serverless deployments; its documentation separately describes managed-compute and serverless deployment options (Microsoft Foundry Models overview). Commercial usability instead must be assessed by checkpoint: Mistral models covered by the MNPL cannot be supplied in a commercial activity without a separate Mistral license, regardless of MAU or revenue (mistral.ai).

Introduction and Background

Enterprises evaluating an open-weight LLM for a commercial product face a licensing landscape that looks superficially uniform, since almost every vendor now describes its release as "open," but is legally heterogeneous. The Open Source Initiative's Open Source AI Definition (OSAID) requires that an AI system provide, among other things, "sufficiently detailed information about the data used to train the system so that a skilled person can build a substantially equivalent system" (opensource.org), a bar that weights-only releases do not clear because they omit

"the underlying training process" (opensource.org). Researcher Irene Solaiman's peer-reviewed release-gradient framework places weight downloads in a "downloadable access" tier distinct from "fully open" releases that also share code and data, reflecting a documented "tension between concentrating power and mitigating risks" that vendors weigh when choosing how much to publish (arxiv.org) (arxiv.org).

Against that backdrop, this report treats the real question behind "open weight AI model licenses commercial use", whether an open-weight model can be used commercially and under what conditions, as a per-vendor, per-model, and often per-model-size question rather than a single yes-or-no answer. This report assesses license terms at the checkpoint level and does not assign a vendor-wide license-family count. The sections below walk each vendor's current license, then consolidate the recurring restriction mechanisms, dated observations on adoption, and a compliance-review method a legal or procurement team can reproduce against any newly released model. The report cites checkpoint-specific source files, but readers should re-verify the license and model card for the exact checkpoint they plan to deploy. Because model line-ups and license text both change between releases, any reader deploying a specific checkpoint should re-verify against the live source cited.

What "Open Weight" Means, and How It Differs From "Open Source"

An **open-weight model** is one whose trained numerical parameters ("weights") are published for download, typically via Hugging Face or GitHub, without requiring the training code, the training dataset, or a fully open license for the resulting artifact. An **open-source model**, under the OSI Open Source AI Definition, requires all of that: use, study, modification, and sharing "for any purpose without seeking additional permission," plus enough information about training data, code, and parameters that a skilled party could reproduce a substantially equivalent system (opensource.org) (opensource.org). This report does not determine whether any of the eight model families qualifies as Open Source AI: that assessment requires OSAID's Data Information, Code, and Parameters requirements, rather than publication of a full pretraining dataset alone (opensource.org).

The OSD itself, the 25-year-old standard the OSI uses to certify conventional software licenses, is instructive here because Apache 2.0 and MIT inherit its guarantees. OSD criterion 6 holds that a compliant license "may not restrict the program from being used in a business" (opensource.org), and criterion 1 bars charging "a royalty or other fee" on redistribution (opensource.org). The **Apache License 2.0** operationalizes this for code and, by extension, for models distributed under it: it grants "a royalty-free, irrevocable copyright license to reproduce, prepare Derivative Works of, publicly display, publicly perform, sublicense, and distribute" the work (opensource.org), plus a patent license that applies only to claims a contributor can license and that are necessarily infringed by that contributor's contribution alone or in combination with the work; that patent license terminates if the recipient institutes specified patent litigation (apache.org). The **MIT License** is shorter and permits use, copying, modification, publication, distribution, sublicensing, and sale, subject to including both the copyright notice and permission notice in copies or substantial portions of the software (opensource.org). Neither license contains a monthly-active-user threshold, an attribution-display requirement, or a restriction on training competing models, which is precisely what distinguishes them from the bespoke community licenses covered next. Because the Model Openness Framework (MOF), a peer-reviewed classification proposed under the Linux Foundation's Generative AI Commons, was created in part because "some model producers use restrictive licenses whilst claiming that their models are 'open source'" (arxiv.org), buyers should treat a vendor's own "open source" label as a marketing claim to verify, not a legal conclusion, and check the actual license text or identifier attached to the specific checkpoint being deployed.

Table 1 below summarizes how the three release categories differ along the dimensions that matter for a commercial deployment decision.

Dimension	OSI-approved software license (Apache 2.0 / MIT)	Threshold-based Restricted Open Weight (Bespoke Community License)	Closed / API-Only
Training code and data published	Not required for the license itself, but the license permits full reuse of what is published	Rarely full data; weights published	No
Field-of-use restrictions	None (opensource.org)	Common: MAU/revenue thresholds, no-compete-training clauses, territorial carve-outs	Governed entirely by API terms of service
Commercial use below scale threshold	Permitted, subject to applicable license conditions	May be permitted until a defined threshold; non-production licenses can prohibit commercial supply (Mistral AI Non-Production License).	Subject to provider's commercial terms and pricing
Redistribution rights	Permitted, including sublicensing, subject to applicable license conditions (opensource.org)	Conditioned on naming/attribution clauses (e.g., "Built with Llama")	Not applicable; no weights to redistribute
Example (September 2026)	Mistral 3, GLM-4.5/4.6, MiniMax-M1, Qwen3	Llama 4, Qwen2.5-72B, Kimi K2, MiniMax-M2, Tencent Hunyuan	Hosted-only frontier models accessed via API

The practical takeaway from *Table 1* is that "open weight" is a spectrum, not a binary: several vendors ship both a permissive-licensed checkpoint and a restricted-licensed checkpoint within the same model family and release date, so the license question must be asked per checkpoint rather than per brand name.

The License Landscape by Model Family

Meta Llama

Meta licenses Llama 4 (April 2025) and Llama 3.1 (July 2024) under near-identical bespoke community licenses, but Llama 4's incorporated policy withholds the grant for any multimodal Llama 4 model from an EU-domiciled individual or a company with its principal place of business in the EU, other than downstream end users of a product or service that incorporates the model (Llama 4 Acceptable Use Policy). For licensees to whom the grant applies, the Llama 4 license tests, on the Llama 4 version release date, whether the preceding calendar month's MAU of products or services made available by or for the licensee or its affiliates exceeded 700 million; only if it did must the licensee request a license from Meta (Llama 4 Community License). Any distributed use requires prominently displaying "'Built with Llama' on a related website, user interface, blogpost, about page, or product documentation" (raw.githubusercontent.com), and any derivative model trained using Llama outputs must "include 'Llama' at the beginning of any such AI model name" (raw.githubusercontent.com). This is a meaningful loosening from the original Llama 2 license (July 2023), which flatly stated licensees "will not use the Llama Materials or any output or results of the Llama Materials to improve any other large language model" (raw.githubusercontent.com);

Llama 3.1 and Llama 4 replaced that outright ban with the narrower naming requirement. Use must also comply with Meta's separate Acceptable Use Policy incorporated by reference (raw.githubusercontent.com) (developer.meta.com).

Alibaba Qwen

Alibaba Cloud splits its Qwen line by size and generation. Qwen2.5-72B-Instruct ships under a custom "qwen" license identifier (huggingface.co) whose text requires that "your product or service has more than 100 million monthly active users, you shall request a license from us" (huggingface.co) The license requires a prominent "Built with Qwen" or "Improved using Qwen" display in related product documentation only when the Materials or their outputs or results are used to create, train, fine-tune, or improve an AI model that is distributed or made available (huggingface.co). By contrast, the entire Qwen3 generation, including its largest mixture-of-experts flagship, carries a plain `license: apache-2.0` identifier with no added conditions (huggingface.co), meaning the MAU-threshold clause that applied to the largest Qwen2.5 model does not carry forward to the current Qwen3 flagship.

Zhipu AI / Z.ai GLM

Zhipu AI's (operating its open releases as Z.ai) GLM-4.5 and GLM-4.6 model weights are both declared under a plain MIT license on their Hugging Face model cards (huggingface.co) (huggingface.co). Neither text carries a usage-scale threshold, a marked change from the earlier GLM-4-9B-Chat release from Zhipu's predecessor project THUDM, whose license required users who wished to use the model for commercial purposes to complete registration through a designated web form before commercial deployment (huggingface.co). The move to MIT for the current GLM-4.5/4.6 generation represents the most permissive licensing stance among the vendors that previously used a registration-gated model.

Moonshot AI Kimi

Moonshot AI licenses its Kimi K2 family, including Kimi-K2-Instruct, its September 2025 point release Kimi-K2-Instruct-0905, and the newer Kimi-K2-Thinking reasoning model, under a "modified-mit" identifier (huggingface.co). The added clause requires that once a licensee's product or service has "more than 100 million monthly active users, or more than 20 million US dollars" in monthly revenue, the licensee "shall prominently display 'Kimi K2' on the user interface of such product or service" (huggingface.co) (huggingface.co). Unlike Meta's, Alibaba's, and Tencent's thresholds, Kimi K2's clause only adds a branding obligation; the underlying MIT grant to use, modify, and redistribute continues to apply regardless of scale. Kimi K3 is different: if the licensee or its affiliates operate a Model as a Service business and exceed $20 million in aggregate revenue over any consecutive 12 months, they must enter a separate agreement with Moonshot AI before commercial use. That requirement does not apply to internal use or use through Moonshot AI's official products or certified inference partners (Kimi K3 License).

MiniMax

MiniMax uses at least three distinct license structures across its recent releases, so the specific checkpoint matters. Its MiniMax-M1 reasoning model carries a plain, unmodified Apache License 2.0 (raw.githubusercontent.com), while its newer MiniMax-M2 uses a modified MIT license that requires licensees exceeding "100 million monthly active users, or more than 30 million US dollars" in annual recurring revenue to "prominently display 'MiniMax M2' on the user interface" (raw.githubusercontent.com) (raw.githubusercontent.com), the same attribution-only pattern

Moonshot uses for Kimi K2. MiniMax-Text-01 uses yet a third structure, a split license with a distinct "MINIMAX MODEL LICENSE" (huggingface.co) governing the weights alongside a separately published "MIT License" (huggingface.co) governing the code repository, underscoring that even within one vendor, code and weights can carry different terms.

Mistral AI

Mistral AI operates the most explicitly dual-tracked licensing program of any vendor in this survey. Its newest flagship generation, Mistral 3 (including Mistral Large 3 and Ministral 3, announced December 2025), states "All models are released under the Apache 2.0 license" (mistral.ai). Separately, Mistral introduced the Mistral AI Non-Production License (MNPL) in May 2024 for models it does not want used commercially without a paid arrangement, stating the license "allows developers to use our technology for non-commercial purposes and to support research work" and confirming the company would keep "releasing models and code under Apache 2.0 as we progressively consolidate two families of products released under Apache 2.0 and the MNPL" (mistral.ai) (mistral.ai); Codestral was the first model shipped under the MNPL (mistral.ai). The MNPL prohibits supplying the defined Mistral Models or Derivatives in the course of a commercial activity: "You shall not supply the Mistral Models or Derivatives in the course of a commercial activity" (mistral.ai), and a licensee wanting commercial use "must request a license from Mistral AI, which Mistral AI may grant to You in Mistral AI's sole discretion" (mistral.ai), though the license clarifies that "Outputs are not considered as Derivatives under this Agreement" (mistral.ai). A still-stricter Mistral AI Research License (MRL) applies to some models and limits use to "Research Purposes" only, explicitly excluding "any usage of the Mistral Model, Derivative or Output by individuals or contractors employed" by a company in the course of business (mistral.ai) (mistral.ai). The practical rule for Mistral is: check the specific model's license identifier before assuming Apache 2.0 applies.

Tencent Hunyuan

Tencent releases its open-weight Hunyuan models, including Hunyuan-Large and the smaller Hunyuan-A13B (June 2025), under the "Tencent Hunyuan Community License Agreement" (raw.githubusercontent.com). For Hunyuan-A13B, the 100-million-MAU condition is tested on the Tencent Hunyuan version release date and counts the preceding calendar month's MAU of all products or services made available by or for the licensee; only if that number exceeds 100 million must the licensee request a license from Tencent (raw.githubusercontent.com). Tencent's license also bars using Hunyuan or its outputs "to improve any other AI model," a no-compete-training restriction similar in spirit to the one Meta dropped after Llama 2 (raw.githubusercontent.com). Distinctively among the licenses surveyed, Tencent's license carries an express territorial carve-out stating it "DOES NOT APPLY IN THE EUROPEAN UNION, UNITED KINGDOM AND SOUTH KOREA" (raw.githubusercontent.com), meaning deployers in those jurisdictions cannot rely on this license as written and should seek separate confirmation from Tencent before commercial use there. The incorporated acceptable-use policy also prohibits high-stakes automated decisions affecting an individual’s safety, rights, or wellbeing, including in medicine/health, credit, employment, housing, education, insurance, law enforcement, migration, and critical infrastructure, as well as unauthorized or unlicensed financial, legal, or medical/health practice.

NVIDIA Nemotron

Nemotron governing terms vary by checkpoint. For example, NVIDIA's governing-terms page for NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16 states that the model is governed by the OpenMDW License Agreement, version 1.1 (NVIDIA NGC). Deployers should identify the exact checkpoint and read its governing terms rather than apply a Nemotron family-wide license label.

Table 2 compares selected model licensing positions across the eight vendors.

Vendor	License position (last verified September 5, 2026)	License type	Commercial use	Scale trigger requiring new license or attribution
Meta	Llama 4	Bespoke community license	For multimodal Llama 4, the grant is unavailable to EU-domiciled individuals and EU-principal-place-of-business companies, except downstream end users; otherwise subject to license conditions	On the Llama 4 version release date, 700M preceding-month MAU across products/services made available by or for the licensee or its affiliates requires a separate license
Alibaba Cloud	Qwen3 (flagship); Qwen2.5-72B (prior gen)	Apache 2.0 (Qwen3); custom "qwen" license (Qwen2.5-72B)	Yes	Qwen2.5-72B only: 100M MAU requires separate license; Qwen3 has none
Zhipu AI / Z.ai	GLM-4.6	MIT (weights); Apache 2.0 (code)	Yes, subject to applicable license conditions	None found in current text
Moonshot AI	Kimi K3	Kimi K3 License	Subject to the license conditions	A Model as a Service business with more than $20M aggregate revenue over 12 months must enter a separate agreement before commercial use; 100M MAU or $20M monthly revenue triggers a Kimi K3 UI-display condition, subject to stated exceptions
MiniMax	MiniMax-M2 (modified MIT); MiniMax-M1 (Apache 2.0)	Mixed by model	Yes	M2: 100M MAU or $30M in annual recurring revenue triggers attribution display
Mistral AI	Mistral 3 (Apache 2.0); other lines under MNPL/MRL	Mixed by model	Apache 2.0 models: yes, subject to license conditions. For MNPL models, assess the proposed workflow under the license's defined terms	MNPL prohibits commercial supply of defined Models or Derivatives; it treats Outputs separately from Derivatives

Vendor	License position (last verified September 5, 2026)	License type	Commercial use	Scale trigger requiring new license or attribution
Tencent	Hunyuan-A13B	Tencent Hunyuan Community License	Subject to license and acceptable-use restrictions; not licensed in EU/UK/South Korea	On the Tencent Hunyuan version release date, 100M preceding-month MAU across all products/services made available by or for the licensee requires a separate license; high-stakes automated decisions and unauthorized or unlicensed professional practice are prohibited
NVIDIA	NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16	OpenMDW-1.1	Check the governing terms for the selected checkpoint	Check the selected checkpoint's governing terms

Table 2 shows that scale-based conditions include release-date MAU tests for Llama 4 (700 million) and Hunyuan-A13B (100 million), an ongoing 100-million-MAU requirement for Alibaba's older Qwen2.5-72B, and Kimi K3's Model as a Service and attribution conditions, alongside MiniMax's attribution condition. The cited Zhipu/Z.ai and current-generation Qwen3 checkpoints impose no scale condition in their license text; Nemotron terms must be assessed by checkpoint.

Commercial-Use Restriction Patterns Across Open-Weight Licenses

Reviewing the eight license texts above together surfaces a small, recurring set of restriction mechanisms that any enterprise legal or procurement review should check for by name, rather than relying on a vendor's "open" label. The most consequential is the **usage-scale condition**: Llama 4 and Hunyuan-A13B require a license request only if their release-date MAU tests are met, Qwen2.5-72B has an ongoing 100-million-MAU license-request requirement, and Kimi K2 and MiniMax-M2 impose attribution-display duties. Kimi K3 instead requires a separate Moonshot agreement before commercial use when a qualifying Model as a Service business exceeds $20 million in aggregate revenue over any consecutive 12 months, subject to its internal-use and official-product/certified-partner exceptions (Kimi K3 License). A second mechanism is the **branding or naming requirement** ("Built with Llama," "Built with Qwen," displaying "Kimi K2" or "MiniMax M2" in a product's interface), which is a compliance obligation rather than a bar on use but is easy to miss in a rapid deployment. A third is the **no-compete-training restriction**, barring use of a model or its outputs to train or improve a different, competing AI model; this appears in Tencent's current Hunyuan license and appeared in Meta's original Llama 2 license, but Meta narrowed it for Llama 3.1 and Llama 4 to the lighter naming rule described above. A fourth is a **territorial carve-out**: Tencent's Hunyuan license excludes the EU, UK, and South Korea, while Llama 4's incorporated policy withholds the Community License grant for multimodal Llama 4 models from EU-domiciled individuals and EU-principal-place-of-business companies, except downstream end users (Llama 4 Acceptable Use Policy). A fifth is a **litigation-termination clause**, seen in Alibaba's Qwen License Agreement, which automatically terminates the license if the licensee brings a patent or IP

claim against the licensor. Finally, several licenses include an **outputs disclaimer**, in which the vendor states it claims no ownership over content generated using the model, shifting responsibility for output use to the licensee, as both Tencent and NVIDIA do explicitly in their license texts.

Table 3 isolates each mechanism with a representative example and its practical consequence for a deployment team.

Restriction mechanism	Representative license	Trigger or detail	Practical implication for deployers
Release-date MAU license condition	Meta Llama 4	On the Llama 4 version release date: 700M preceding-month MAU of products/services made available by or for the licensee or its affiliates	A separate license is required only if that release-date condition is met
License-request condition	Alibaba Qwen2.5-72B; Tencent Hunyuan-A13B	Qwen2.5-72B: 100M MAU; Hunyuan-A13B: on its version release date, 100M preceding-month MAU of all products/services made available by or for the licensee	Qwen2.5-72B requires a license when its threshold is exceeded; Hunyuan-A13B requires one only if its release-date condition is met
Attribution-display duty (not a use bar)	Moonshot Kimi K2; MiniMax-M2	100M MAU, or $20M/month revenue for Kimi K2 or $30M in annual recurring revenue for MiniMax-M2	Requires a visible UI credit once scale is reached; does not block use itself
Branding-display requirement	Meta Llama 4	Distributing Llama Materials, derivatives, or a containing product or service requires a "Built with Llama" display (Llama 4 Community License)	Do not treat this display requirement as a model-naming condition for every distribution
No-compete-training restriction	Tencent Hunyuan (current); Meta Llama 2 (historical, since narrowed)	Using outputs to improve a different AI model	Assess this condition against the proposed use
Territorial carve-out	Tencent Hunyuan; Llama 4 multimodal models	Tencent excludes EU, UK, South Korea; Llama 4's incorporated policy withholds the grant from EU-domiciled individuals and EU-principal-place-of-business companies, except downstream end users (Llama 4 Acceptable Use Policy)	Check the exact model and incorporated policy before deployment
Acceptable-use restrictions	Tencent Hunyuan	Prohibits high-stakes automated decisions affecting safety, rights, or wellbeing and unauthorized or unlicensed professional practice	Screen proposed use cases, including regulated workflows, against the incorporated policy before deployment
Commercial-supply restriction	Mistral MNPL	Supplying the defined Mistral Models or Derivatives in the course of a commercial activity	Assess the proposed workflow under the defined terms; Outputs are not Derivatives under the MNPL

The license texts in *Table 3* use different mechanisms and must be evaluated for the specific checkpoint and proposed deployment; Mistral's MNPL applies a commercial-supply restriction regardless of scale.

Implementation Guidance: Conducting a License Compliance Review Before Deployment

Because license terms vary by vendor, by model size within a vendor's own line-up, and by release date, a repeatable review method matters more than memorizing any single vendor's current terms. First, **identify the exact license attached to the specific checkpoint**, not the model family name, since Hugging Face model cards declare a machine-readable `license` field in their YAML metadata, which Hugging Face's own documentation confirms is used to "filter models by license" on the Hub (huggingface.co) and that any repository owner is able to add a declared license to (huggingface.co). A declared identifier of `apache-2.0` or `mit` signals an OSI-approved license whose full text should still be read before deployment; an identifier of `other` with a custom `license_name` (as Qwen, Kimi, and MiniMax's M2 all use) signals a bespoke text that must be read in full before deployment.

Second, **check the timing and scope of every usage-scale condition**: Llama 4 and Hunyuan-A13B use release-date MAU tests, Qwen2.5-72B has an ongoing license-request requirement, Kimi K2 and MiniMax-M2 impose attribution-display duties, and Kimi K3 can require a separate Moonshot agreement before commercial use by a qualifying Model as a Service business that exceeds $20 million in aggregate revenue over any consecutive 12 months, subject to its stated exceptions (Kimi K3 License). Third, **check for a no-compete-training clause and a branding/naming requirement**, both of which are easy to overlook because they do not prevent initial deployment but create a compliance gap that surfaces later, for example, when a company later wants to fine-tune a competing model on the same infrastructure. Fourth, **check for territorial carve-outs and incorporated acceptable-use restrictions**: Tencent’s license can exclude entire jurisdictions from coverage regardless of scale, and Llama 4's incorporated policy withholds the grant for multimodal Llama 4 models from EU-domiciled individuals and EU-principal-place-of-business companies, except downstream end users (Llama 4 Acceptable Use Policy). Fifth, where a license is genuinely ambiguous or the deployment approaches a stated threshold, treat the Model Openness Framework's underlying rationale, that "some model producers use restrictive licenses whilst claiming that their models are 'open source'" (arxiv.org), as a reason to have counsel review the primary license text rather than a vendor blog post or third-party summary; the framework was designed explicitly to help "model consumers identify open models and their constituent components that can be permissively used, studied, modified, and redistributed" (arxiv.org) and defines 17 separate components spanning code, data, and documentation that a fully open release should provide (arxiv.org).

Sixth, if deploying via a managed cloud platform rather than self-hosting, remember that the platform's terms of service do not replace the model vendor's license: Amazon Web Services (AWS) states that "Serverless Third-Party Models offered on Amazon Bedrock may be subject to additional terms or a seller's end user license agreement" (aws.amazon.com), and its documentation instructs teams to "verify that your grant is in an Active state" before invoking a third-party model (docs.aws.amazon.com). Microsoft's Azure AI Foundry documentation similarly states that for partner and community models, "Model providers define the license terms and set the price for use of their models using Azure Marketplace" (learn.microsoft.com), and that "Users accept license terms for use of the models" as a discrete step (learn.microsoft.com). A cloud contract's standard commercial terms do not override or satisfy a vendor's separate MAU threshold, attribution clause, or (in Mistral's MNPL case) commercial-supply restriction.

For organizations in regulated industries such as life sciences, this review is not purely a legal exercise: it intersects with existing vendor and software risk assessment processes, since a model's license terms, redistribution obligations, and any output-use restrictions need to be tracked alongside the validation and audit-trail requirements that already apply to any AI system touching regulated data or workflows. Where electronic records are subject to FDA requirements, 21 CFR Part 11 applies to electronic records created, modified, maintained, archived, retrieved, or transmitted under those requirements (21 CFR Part 11).

Data Analysis and Evidence

Method note: the figures in this section are drawn from Stanford HAI's 2025 AI Index Report and the Model Openness Framework paper, both independent third-party sources rather than vendor claims; percentages describe the composition of publicly tracked "notable" AI model releases as classified by Epoch AI and reproduced in the HAI report, not a survey of enterprise deployment share, and should not be conflated with the license-specific facts in the sections above, which come from each vendor's own license text.

Stanford HAI's 2025 AI Index Report found that the capability gap between the best closed-weight and best open-weight models narrowed sharply within about one year: on the Chatbot Arena leaderboard, the top closed model "outperformed the top open-weight model by 8.0%. By February 2025, this gap had narrowed to 1.7%" (hai.stanford.edu). On the widely cited MMLU (Massive Multitask Language Understanding) benchmark, the same report found that at the start of 2024 "closed-weight models led open models on MMLU by 15.9 points, but by the end of 2024, that difference had shrunk to just 0.1 percentage point" (hai.stanford.edu). This closing gap is the underlying reason enterprise interest in open-weight licensing terms has grown: a model that is nearly competitive with closed frontier systems, but downloadable and self-hostable, changes procurement and compliance calculations in a way a large capability gap would not.

On release composition, the same 2025 AI Index Report's breakdown of notable AI models released in 2024 by access type found that "18.03%" of tracked releases were classified as "Open weights (restricted use)" and a further "11.48%" as "Open weights (unrestricted)" (hai.stanford.edu) (hai.stanford.edu), together accounting for roughly three in ten notable model releases tracked that year, alongside API-only and fully closed releases. The same report explicitly warns against treating that access as equivalent to open-source status, noting that open-weight models "are not necessarily fully open source, as the underlying code or training data is often withheld" (hai.stanford.edu), corroborating the OSI's own position cited earlier in this report.

The Model Openness Framework paper adds a structural explanation for why license diligence matters at the component level rather than the release level: it defines "17 components" spanning "code, data, and documentation" (arxiv.org) against which any given release can be scored, and the eight vendors surveyed in this report each publish a different subset of those components under a different mix of licenses, which is why no single "open" or "restricted" label accurately describes an entire vendor's output.

Implications and Future Directions

Two trends observed in this survey are likely to continue shaping how open-weight licensing evolves. First, several vendors, most visibly Alibaba with its shift from the custom Qwen2.5-72B license to Apache 2.0 across the entire Qwen3 generation, and Zhipu AI/Z.ai with its move from THUDM's registration-gated GLM-4 license to a plain MIT

license for GLM-4.5 and GLM-4.6, have moved toward more permissive terms release over release rather than less permissive ones. Whether this trend holds as models approach or exceed the capability of the best closed systems, per the narrowing Chatbot Arena and MMLU gaps described above, or reverses as vendors seek more leverage over how their weights are used at scale, is not yet resolvable from public license texts alone and should be tracked release by release rather than assumed to continue.

Second, the attribution-only threshold structure used by Moonshot's Kimi K2 and MiniMax's MiniMax-M2, which requires a UI credit at scale rather than a new license, may represent a middle path between Meta's, Alibaba's, and Tencent's outright license-request requirements and full permissive licensing; if other vendors converge on this lighter-touch model, commercial-use analysis would increasingly focus on the applicable attribution condition rather than an additional license requirement. Irene Solaiman's release-gradient framework frames this kind of graduated restriction as a rational response to a documented "tension between concentrating power and mitigating risks" (arxiv.org) rather than a purely commercial calculation, and that framing is likely to remain the standard lens for explaining why any two vendors' otherwise similar-looking open-weight releases carry different legal terms.

For organizations whose electronic records are subject to FDA requirements, the applicable controls include procedures and controls designed to ensure the authenticity, integrity, and, when appropriate, confidentiality of electronic records (21 CFR Part 11).

Frequently Asked Questions (FAQs)

Can Llama models be used commercially? It depends on the exact model and license. For any multimodal Llama 4 model, the incorporated policy withholds the Community License grant from an EU-domiciled individual or a company with its principal place of business in the EU, except downstream end users of a product or service that incorporates the model (Llama 4 Acceptable Use Policy). For licensees to whom the Llama 4 grant applies, a separate license is required only if, on the Llama 4 version release date, the preceding calendar month's MAU of products or services made available by or for the licensee or its affiliates exceeded 700 million (Llama 4 Community License).

What's the difference between open weight and open source AI models? Open weight means the trained parameters are downloadable; open source, per the OSI's Open Source AI Definition, additionally requires the training code, sufficient data information to reproduce the system, and a license granting the freedom to use it "for any purpose without seeking additional permission" (opensource.org) (opensource.org). This report does not determine whether any of the eight vendors qualifies as Open Source AI, because that assessment requires OSAID's Data Information, Code, and Parameters requirements rather than publication of a full pretraining dataset alone (opensource.org).

Which open-weight models use Apache 2.0 or MIT versus a custom license? Last verified September 5, 2026: Mistral 3, GLM-4.5/4.6, Qwen3, and MiniMax-M1 use plain Apache 2.0 or MIT. Nemotron licenses vary by release: NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16 is governed by OpenMDW-1.1 (NVIDIA NGC). Llama 4, Qwen2.5-72B, Kimi K2, Kimi K3, MiniMax-M2, and Tencent's Hunyuan line use bespoke community licenses with added conditions; K3 includes a separate-agreement condition for qualifying Model as a Service businesses above its aggregate-revenue threshold, subject to stated exceptions (Kimi K3 License). See *Table 2* above for the full comparison.

What are the redistribution rules for open-weight models? Apache 2.0 and MIT permit redistribution, including sublicensing, subject to their notice and license-preservation conditions (opensource.org). Bespoke licenses typically permit redistribution but attach a naming or attribution condition, for example Meta's "Built with Llama" requirement (raw.githubusercontent.com) or Kimi K2's UI-display requirement at scale (huggingface.co).

Are there deployment restrictions tied to where a company operates? Yes. Tencent's Hunyuan Community License states it "DOES NOT APPLY IN THE EUROPEAN UNION, UNITED KINGDOM AND SOUTH KOREA" (Tencent Hunyuan Community License). In addition, for any multimodal Llama 4 model, the incorporated policy withholds the Community License grant from an EU-domiciled individual or a company with its principal place of business in the EU, except downstream end users of a product or service that incorporates the model (Llama 4 Acceptable Use Policy).

Does deploying an open-weight model through a cloud provider change the license? Review the exact checkpoint and offering. Amazon Bedrock says serverless third-party models may be subject to additional terms or a seller's end-user license agreement, while Bedrock Marketplace models follow a separate route to the applicable agreement (AWS Bedrock third-party-model terms). Azure AI Foundry says users accept license terms for serverless deployments (Microsoft Foundry Models overview). Those hosted-service terms should be reviewed alongside, rather than assumed to replace, the license governing downloadable weights.

Is Mistral's Codestral usable commercially? Codestral was the first Mistral model released under the Mistral AI Non-Production License (mistral.ai). The MNPL prohibits supplying the defined Mistral Models or Derivatives in the course of a commercial activity, while stating that Outputs are not Derivatives under the agreement (mistral.ai). A proposed commercial workflow should be assessed under those defined terms and the full license. Mistral's Apache 2.0-licensed Mistral 3 family has different terms.

Conclusion

Open-weight AI model licenses commercial use is not answerable with a single rule, because the eight vendors surveyed here, Meta, Alibaba, Zhipu AI/Z.ai, Moonshot AI, MiniMax, Mistral AI, Tencent, and NVIDIA, each apply different terms, and several apply more than one license across their own current line-up. The dividing line that matters most in practice is not "open" versus "closed" but whether a given checkpoint carries an OSI-approved license (Apache 2.0 or MIT, as with Mistral 3, GLM-4.5/4.6, Qwen3, and MiniMax-M1) subject to its terms, NVIDIA Nemotron checkpoints, whose governing terms must be checked individually (NVIDIA NGC), or a bespoke community license with a usage-scale threshold, an attribution duty, a no-compete-training clause, or a territorial carve-out (as with Llama, Qwen2.5-72B, Kimi K2, Kimi K3, MiniMax-M2, and Tencent Hunyuan). For K3, a qualifying Model as a Service business exceeding $20 million in aggregate revenue over any consecutive 12 months must enter a separate Moonshot agreement before commercial use, subject to its stated exceptions (Kimi K3 License).

For many enterprise deployments below the stated scale thresholds, commercial use may be available, but it remains subject to model-specific conditions. For any multimodal Llama 4 model, the incorporated policy withholds the Community License grant from EU-domiciled individuals and EU-principal-place-of-business companies, except downstream end users of a product or service that incorporates the model (Llama 4 Acceptable Use Policy). Tencent Hunyuan, for example, prohibits high-stakes automated decisions affecting an individual’s safety, rights, or wellbeing and unauthorized or unlicensed professional practice, and is not licensed in the EU, UK, or South Korea.

Because license text changes between point releases, and because a single vendor can and does apply different licenses to different model sizes within the same generation, the specific checkpoint's license file, not the model family's public reputation, should be the last thing checked before a commercial deployment ships, and the first thing re-checked when that deployment is updated to a newer release.
