AI Explained

Open Weight v Closed Weight Models

Part 1 – What an open model gives you

By Dr Peter Fields

THE 2-MINUTE SUMMARY

Weights are the numbers a model learns during training. An open-weight release makes those numbers available to download under a licence. A closed-weight model is usually available through a service while the provider keeps the weights private.

Open source goes further. The Open Source Initiative looks at the rights people are given and whether they can get the code and training-data information needed to study and modify the system. Two downloadable models can have very different licences.

Hosting is a separate choice. A supplier can host an open-weight model. A firm can also run a closed-weight model in its own environment if the supplier offers that arrangement. Where your documents go depends on how the whole tool is set up.

For most users, the answer, price and handling of their files matter more than the release label. For the company building the tool, weight access can change who runs the model, what can be adapted and who pays to keep it working.

The same screen, a different arrangement

Suppose your firm has a tool that pulls out the renewal dates and notice periods from contracts. You upload an agreement and get a table. The people running the tool have questions you might never think to ask. Can they get a copy of the model? Can they keep using a version they have tested if the supplier retires it?

The weights are the numbers adjusted during training as the model learns patterns from examples. To answer a prompt, the model uses those numbers according to its architecture, the design that determines how they are put to work.

Open weights and open source

An open-weight release makes the trained weights available to download under stated terms. With suitable software and hardware, you can run a copy. With closed weights, the developer keeps those numbers private. A customer might use the model in a chat app or through an API, a service that lets one program send requests to another. Access to the service depends on the provider's terms.

The Open Source Initiative's definition asks more of a release. People must have the freedom to use, study, modify and share the system for any purpose. They also need the weights, the code used to train and run the model, and sufficiently detailed information about the training data for a skilled person to

build a substantially equivalent system. Some of the original training data may be unavailable for redistribution. The definition calls for detailed information about such data as well. [1]

The terms can vary even within one developer's range. Mistral released its Mistral 3 family under Apache 2.0, which gives broad rights to use, modify and distribute the release. Its later Large 4 also has downloadable weights, under terms that require a separate commercial licence to offer the model itself as a service. Meta's Llama 3.3 has its own conditions, including a separate licence for companies above the agreement's 700 million monthly-user threshold. For an example of a fuller release, Ai2 published Olmo 3 with its weights, data, training code and copies from intermediate stages of training. [2–4]

Where the model runs

Open weights describe what the developer releases. Hosting describes where an installed copy runs and who operates it. Amazon lists Meta's open-weight Llama 3.3 on its Bedrock service, while LM Studio explains how to run a downloaded model offline. A supplier can also arrange a private installation of a closed-weight model. Mistral offered that option for Medium 3, on the customer's premises or in a private cloud environment. [5–6]

The screen may look the same whichever arrangement the firm chooses. To find out where the contract goes, check how the app handles files and which outside services it calls. The Confidentiality and Privacy articles take that question further.

Geek Box – Will it fit on my computer?

For the weights alone, allow about 2 GB per billion parameters at 16-bit precision or 0.5 GB at 4-bit. So a 7-billion-parameter model at 4-bit starts at 3.5 GB and a 70-billion-parameter model at 35 GB. Both need extra working memory, especially for long inputs. Storing weights with fewer bits is called quantisation. It saves memory and can affect the quality of the answers. [7]

What changes when you can get the weights?

Access to the weights may let the firm keep a version it has tested, adapt it under the licence or move it to other equipment. Operating its own copy means taking on computing costs, updates and support. The AI Agents and Tool Use article covers what else a firm may have to change when it switches providers.

Stanford's 2026 AI Index gives a snapshot. On Arena's text leaderboard, where people compare pairs of answers, the top closed-weight model led the top open-weight model by 49 points in March 2026, up from seven in August 2024. The scale works roughly like chess ratings. For the contract tool, what matters is which model picks up a notice period qualified by a schedule. A general leaderboard gives you little basis for judging that. [8]

There is a safety trade-off too. A provider can change or withdraw access to a model on its own service. Copies already downloaded remain outside its control. Access to those copies gives independent researchers more scope to examine the model. It also makes sense to check safeguards after any fine-tuning. Qi and colleagues found that further training could weaken them, including on models tuned through a hosted service. [9]

How to use it better

If you're choosing an app, try it on work you recognise and check what happens to the files you give it. If you're choosing a model to build into a tool, check the licence for that exact release and decide who will run it.

Finding the notice period in a contract is one job. Advising on whether it is a term your firm would usually accept takes a different kind of evidence. Part 2 asks how to bring the firm's own precedents and past decisions into the answer, whoever hosts the model.

Sources and notes

1. Open Source Initiative, The Open Source AI Definition, v1.0, and accompanying FAQ on data information. Definition • FAQ

2. Meta (2024), Llama 3.3 Community License Agreement, §§1–2. The threshold in §2 is more than 700 million monthly active users in the preceding calendar month, measured at the release date and subject to the terms of that section. Licence

3. Mistral, Introducing Mistral 3 (2 December 2025), and Mistral Large 4 licence. See also Apache License 2.0, §§2–4. Mistral 3 • Large 4 licence • Apache 2.0

4. Ai2 (2025), Olmo 3 release announcement, on data, code, weights and intermediate checkpoints. Release notes

5. Amazon Web Services, Llama 3.3 70B Instruct on Bedrock. Mistral, Medium is the new large (7 May 2025), on private deployment of Medium 3. AWS • Mistral

6. LM Studio, Offline Operation, on using downloaded models locally. Documentation

7. Elias Frantar, Saleh Ashkboos, Torsten Hoefler and Dan Alistarh (2023), GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers, ICLR, §§2, 4–5. Paper

8. Stanford Institute for Human-Centered Artificial Intelligence (2026), AI Index Report, ch. 2, §2.1, fig. 2.1.2. The figures are Arena rating points from the March 2026 export. Chapter PDF

9. Xiangyu Qi, Yi Zeng, Tinghao Xie and colleagues (2024), Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!, ICLR. Paper abstract