Xinference

Open models, explained.

What they are, what people get wrong and how to get started

xinference.co

Xinference

Can we get results like ChatGPT without sending everything to one AI provider?

Open models are one answer. Here’s what they are, how good they’ve become and how to try them.

Xinference

Who we are

We run open models through one API, hosted in Australia.

A model API means nothing to install or run. Send a request, get an answer, pay per token.

OpenAI	Per token	Your data
Compatible API	Pay only for what you use	Not used for training

Sources xinference.co/model-api, read 30 Oct 2026

Xinference

What we'll cover

01 **What open models are**
And what the licences let you do

02 **How good they are**
Compared with the best closed models

03 **Common misconceptions**
Six we hear all the time

04 **Ways to run them**
From an API to your own servers

05 **Why it matters here**
For Australian teams

06 **How to get started**
Then your questions

Xinference

What is an open model?

Open weights

You can download the trained model and run it yourself. Most models people call open are this.

Open source

Also shares enough about the training data and code for a skilled person to build an equivalent model.

For running a model yourself, open weights is what matters.

Source Open Source Initiative, Open Source AI Definition 1.0, October 2024

Xinference

Licences in plain English

APACHE 2.0 AND MIT

Permissive

Use it commercially, change it and keep the licence notice.

FOR EXAMPLE

Gemma 4, gpt-oss, DeepSeek V4.1, Mistral Large 3

THE MODEL MAKER'S OWN TERMS

Custom

Free to use, with extra conditions once a business or product gets very big.

FOR EXAMPLE

Llama 4, Kimi K3, GLM-5.3

BEFORE YOU BUILD ON IT

Our tip

Read the licence for the exact model you pick, and check it again when you switch models.

Sources Apache 2.0 and MIT licence texts, and model cards and licence files on Hugging Face from Google, OpenAI, DeepSeek, Mistral AI, Meta, Moonshot AI and Z.ai, read 8 to 12 Oct 2026

Xinference

Three words you'll hear

01	**Tokens**	The chunks of text a model reads and writes, roughly three quarters of a word each. Usage is priced per million tokens.
02	**Parameters**	The learned numbers inside a model. More usually means more capable, and more costly to run.
03	**Context window**	How much text the model can work with in one go, including its answer.

Source OpenAI Help Center, What are tokens and how to count them (a rule of thumb for English text)

Xinference

How good are open models now?

4 months

Roughly how far the best open models trail the best closed ones



Under 1/10

DeepSeek V4.1 Flash's price per token, compared with Claude Opus 5.5

DeepSeek V4.1 Flash: $0.30 in, $1.20 out
Claude Opus 5.5: $4 in, $20 out

Close to the best, for a fraction of the price.

Sources Epoch AI, open and closed model gap on its Capabilities Index, 29 May 2026, and list prices in USD per million tokens on xinference.co/model-api and Anthropic's pricing page, read 29 Oct 2026

Xinference

Open models you'll hear about

MODEL	MADE BY	BASED IN	LICENCE
Kimi K3.1 pro	Moonshot AI	China	Its own licence
GLM-5.3	Z.ai	China	Its own licence
DeepSeek V4.1 Flash	DeepSeek	China	MIT
Qwen3.8-27B	Alibaba (Qwen)	China	Apache 2.0
Gemma 4	Google	United States	Apache 2.0
Mistral Large 3	Mistral AI	France	Apache 2.0
Muse	Meta AI	United States	Its own licence

As of October 2026. New ones arrive every few weeks.

Sources Model cards and licence files on Hugging Face from Moonshot AI, Z.ai, DeepSeek, Qwen, Google and Mistral AI, read 12 Oct 2026

Xinference

Six common misconceptions

“Open models are way behind.”

Months behind the very best, not years. Many everyday tasks won’t show a difference.

“Open means free.”

Free to download, not to run. You still pay for the computing power.

“Open means anyone can see my data.”

The model is open. Your data isn’t. Who sees it depends on who runs the model.

“A Chinese model sends my data to China.”

Your data goes to whoever hosts the model. Use the lab’s own app or API and that’s the lab.

“Cheapest per token is cheapest overall.”

What counts is the cost of a good answer on your task.

“We’d have to run it ourselves.”

You can use an API, dedicated GPUs or your own servers.

Xinference

Who made a model vs where it runs

WHO MADE IT

Shapes how it behaves

Training choices shape its answers, including on political topics, and the maker sets the licence. Test it on your own work.

WHERE IT RUNS

Decides who sees your data

Prompts go to whoever hosts the model. Use the lab's own app or API and that's the lab. Run it elsewhere and it isn't.

OPEN MODELS COME FROM MANY PLACES

UNITED STATES	FRANCE	CHINA
Gemma, gpt-oss, Nemotron, Llama	Mistral	Kimi, GLM, DeepSeek, Qwen

Sources Model cards on Hugging Face from Google, OpenAI, NVIDIA, Meta, Mistral AI, Moonshot AI, Z.ai, DeepSeek and Qwen, read 8 Oct 2026

Xinference

Three ways to run them

Managed API

Send requests, pay per token. Nothing to run.

GOOD FOR

Getting started, and spiky or low-volume use

Dedicated GPUs

A model running on GPUs reserved for you.

GOOD FOR

Steady, heavy use and predictable speed

Self-hosted

The model runs on your own servers or cloud account.

GOOD FOR

Strict data control and customisation

Xinference

Why Australian teams are looking at open models

Data control

Under the Privacy Act (APP 8), you generally stay accountable for personal information you send overseas.

Transparency

From 10 December 2026, privacy policies must explain when a computer program makes or informs decisions that significantly affect people.

Cost

Many everyday tasks don’t need the most expensive model.

Choice

Switch models as better ones arrive, without rewriting your app.

Sources OAIC on Australian Privacy Principle 8, and the OAIC release of 30 Sep 2026 on transparency for automated decisions

Xinference

How to get started

01

Pick one real task

Something your team does often, with a clear idea of a good answer.

02

Gather examples

Ten to twenty real examples, with the answers you'd expect.

03

Test a few models

Compare quality, speed and cost on those examples.

04

Choose how to run

An API, dedicated GPUs or your own servers.

Xinference

Where Xinference fits

DEDICATED INFERENCE

MANAGED API

Model API

Open models through an OpenAI-compatible API. Pay per token.

DEDICATED GPUS

Cloud deployment

A model on GPUs reserved for your workload, fully managed.

SELF-HOSTED

Private deployment

Run models on your own infrastructure, with fine-tuning, monitoring and access controls.

Don’t see the model you need? Talk to us about adding it.

Xinference

Questions?

Want to try an open model on your own work? Let’s talk.

xinference.co