Australian Government
Australian Signals Directorate

cyber.gov.au

Open-weight **AI models**

Opportunities and cyber security risks



Table of contents

Purpose 3

Audience 3

Introduction 3

Closed, open-weight and open-source models 4

Opportunities and benefits 4

Risks arising from the public release of model weights 5

Model, data and software supply chain 5

Fine-tuned model security 6

Safeguards, integrations and agentic systems 6

Conclusion 7

Further information 7

Purpose

This publication explains what open-weight AI models are, the benefits they can provide, and the cyber security and supply chain risks that arise specifically because model weights have been publicly released. Open-weight models are AI models whose underlying parameters, or 'weights', are publicly released which allows organisations to download, modify and operate them on their own infrastructure. Adopting open-weight models gives organisations greater control over model deployment, operation and lifecycle. It also places greater responsibility on the organisation for security, safeguards, monitoring, updates and potential incident response. The selection principles, deployment, operational and governance risks that organisations assume when self-hosting any AI model are addressed in the companion publication *Selecting, deploying and operating open-weight AI Models securely*.

Audience

This guidance is intended for executives, technology leaders and cyber security practitioners considering whether to adopt open-weight AI models within their organisation.

Introduction

A series of open-weight models have been released recently, with performance on selected benchmarks approaching that of leading closed models. Current and emerging open-weight models may be capable of supporting activities such as vulnerability discovery, autonomous multi-step operations, software supply chain compromise and sophisticated social engineering. Their increasing capability, availability and customisability make them of growing interest to both defenders and malicious actors. Not all vendor-reported benchmark results may have been independently verified but these releases indicate that highly capable AI systems are increasingly available outside provider-controlled environments.

Self-hosting open-weight models enables greater customisation and control, supporting research, innovation and operational resilience. However, it also transfers greater responsibility for security, safeguards, monitoring and incident response to the organisations that deploy or integrate them. While self-hosting may reduce usage costs at scale, organisations should assess the total cost of ownership associated with securely operating the model, including infrastructure, energy consumption, cyber security controls, monitoring, maintenance and access to appropriately skilled personnel.

Closed, open-weight and open-source models

AI models can generally be categorised as either closed, open-weight or open-source based on which model components (weights, source code and training data) are released, and under what licence conditions. Each category brings its own benefits and constraints. The key characteristics of each category can be summarised below. This publication focuses on open-weight models because they are increasingly available, capable and widely adopted, whereas fully open-source models remain uncommon. Their capability, accessibility and customisability create distinct cyber security considerations that warrant specific guidance.

AI models are released along a scale of openness

Greater openness enables wider use, customisation and auditability, but also transfers more responsibility for security, safety and compliance to the user or operator.



Closed models

Operated by the developer or a partner platform

- Customers access them as a service; the weights never leave the provider.
- The provider operates safeguards, monitors misuse and can update or withdraw the model.
- Updates, controls and visibility are managed by the provider.
- Models include Olmo 3 family.

Key characteristics

- Access via provider service.
- Weights stay with provider.
- Provider runs safeguards and monitoring.
- Provider can update or withdraw model.

Open-weight models

Weights (trained parameters) are published for download.

- Typically without the training data, training code or full documentation.
- Anyone can run, fine-tune and redistribute them, subject to licence terms that vary by model and must be assessed individually.
- Safeguards included at release can be modified or removed by users, and are not updated or enforced by the provider.
- Models include Kimi K3, GLM5.3, Nemotron 3 family.

Key characteristics

- Weights published for download.
- Training data, code and docs usually not included.
- Free to run, fine-tune and redistribute (licence applied).

Open-source models

Weights, training data, code and recipes are published.

- Includes the data, code and documentation needed to reproduce or fully audit the model.
- Enables transparency, reproducibility and independent verification.
- Genuinely open-source released remain rate; mode ‘open’ models are open-weight only.
- Models include Mythos/Opus.

Key characteristics

- Weights, training data, code and recipes published.
- Can be reproduced and fully audited.
- Highest transparency and customisation.
- Genuinely open-source released are rare.

Regardless of the model selected, organisations remain responsible for implementing appropriate controls, oversight and monitoring to prevent, detect and respond to harmful outputs, unintended actions and misuse.

Opportunities and benefits

Open-weight models may offer genuine benefits to Australian organisations, including local control of sensitive data, reduced dependence on offshore providers, support for sovereign research capability, and availability for defensive cyber security operations.

Open-weight models can provide organisations with greater control over deployment, availability, operation and lifecycle management. Organisations may be able to host models within their own environments, including disconnected environments, to process sensitive data. They may also determine when updates are adopted, maintain access to specific model versions, apply custom security controls and reduce dependence on a single provider's availability, policies or commercial decisions.

Provider-managed safeguards, safety policies and content restrictions may limit some legitimate operational, research or defensive cyber security activities, and organisations may have limited ability to adjust these controls. Open-weight models allow an organisation to configure safeguards for its approved use case under its own governance arrangements. This is distinct from weakening or removing safeguards, which is addressed below. However, the benefits of open-weight models should be balanced against the additional security, governance, monitoring and operational responsibilities assumed by the operating organisation.

Risks associated with open-weights AI models

The following risk areas arise because released model weights can be copied, modified and redistributed by any party, without the original developer's oversight. In provider managed AI services ,these risks either do not arise, or are managed by the provider. The risks associated with open-weight AI models are not uniform and will vary between organisations and use cases. Organisations should assess how these risks apply to their environment and implement security controls proportionate to the potential impact. Although focused on open-weight AI models, these risks may also be relevant to varying degrees to other AI deployment models and architectures.

Model selection, provenance and developer jurisdiction

Risk. Organisations can obtain open-weight models from original developers or from third parties that may have converted, quantised, fine-tuned or otherwise modified them. Available documentation and assurance information may be incomplete, contain errors or lack sufficient detail. It may also be difficult to establish the provenance of modifications, and their impact on model behaviour, security and operational outcomes may not be well understood.

Consequently, organisations may deploy a model without adequately understanding its capabilities, limitations, known weaknesses, licence conditions or exposure to the legal and regulatory environment of its developer or distributor. The model may produce biased, unreliable or otherwise unsuitable outputs for the intended use.

Recommended actions. Organisations should define the model's intended and prohibited uses and should not adopt a model whose provenance, documentation or assurance information cannot be established to a level appropriate to that use. Material limitations identified during assessment should be documented, and appropriate human oversight applied to consequential outputs.

Organisations deploying open-weight models should also refer to ASD's publication *Choosing secure and verifiable technologies: Executive guidance*.

Model, data and software supply chain

Risk. As with any software, open-weight models carry supply chain risks. Model weights, model serving frameworks, libraries, AI harnesses and plugins, container images and third-party conversions may contain vulnerabilities or malicious modifications. Some model formats and loading processes can execute code. Tampered weights, or manipulated training, fine-tuning and reinforcement learning datasets, may introduce hidden behaviours, vulnerabilities or backdoors that routine testing is unlikely to detect. Open-weight models may exhibit behaviours that are not apparent during routine evaluation and may only emerge when responding to particular identities, organisations, jurisdictions, topics or operational contexts.

Recommended actions. Organisations should obtain models and related components only from approved sources that have been assessed in accordance with the organisation's procurement, governance, security and risk management processes. They should also verify available hashes, signatures, provenance information, and scan artefacts and dependencies for known vulnerabilities, malware and signs of tampering before use. New models should initially be loaded in a sandboxed environment until the organisation has gained sufficient assurance that the model and associated components do not introduce unacceptable risk. The duration of testing will vary depending on the model source, deployment method, intended use and the organisation's risk tolerance. Organisations should prefer model file formats that cannot contain executable code and avoid formats that serialise arbitrary objects. Where custom code is required, organisations should review it before execution and run it in an appropriately isolated environment.

Organisations should also assess any available information about the provenance of training, fine-tuning and reinforcement learning datasets, and perform adversarial testing to identify

hidden behaviours or backdoors that routine testing may not reveal.

Organisations deploying open-weight models should also refer to ASD's joint publication *Artificial intelligence and machine learning: Supply chain risks and mitigations*.

Fine-tuned model security

Risk. Organisations can fine-tune open-weight models using their own data and retain the resulting model files. These models may memorise or reveal sensitive training information. If they are stolen or disclosed, a malicious actor could copy, operate and examine them to extract that information.

Recommended actions. Before training or fine-tuning an open-weight model, organisations should assess the sensitivity of the data used. Organisations should assess the potential impact of data disclosure, extraction or inference from the resulting model. They should treat fine-tuned model files as potentially containing information derived from the data used to create them. Organisations should also restrict and monitor access, prevent unauthorised copying or export, encrypt the files at rest and in transit, and test models for disclosure of sensitive training data before deployment. Where an open-weight model has been fine-tuned using classified or commercially sensitive information, organisations should assess whether the resulting checkpoints and adapters contain, or could be used to reveal, that information. Any classification and protective controls determined by the relevant classification authority should be applied to those artefacts. A model fine-tuned on large volumes of unclassified or lower-classified data may, in aggregate, warrant a higher classification or additional protections than any individual source.

Safeguards, integrations and agentic systems

Risk. Once an open-weight model is released, the original developer typically no longer controls how model checkpoints are hosted, configured, fine-tuned or safeguarded. As a result, safeguards built into the model may become outdated, operate inconsistently, or be weakened or removed through fine-tuning or modification. This can increase the likelihood that prompt injection attacks succeed, that connected tools are misused, that unauthorised actions are performed, and that sensitive information is disclosed. These risks are amplified in agentic systems, where the model can invoke tools, execute code or take actions autonomously, because a weakened safeguard or successful prompt injection can result in unauthorised actions rather than only unsafe outputs.

Models that have been fine-tuned, quantised or converted, or that have been created through distillation or merging, may not retain the safeguards, alignment properties or behavioural characteristics of the original model.

Recommended actions. Organisations should place external controls between the model and sensitive systems rather than relying solely on safeguards built into the model. Tools, connectors and agent permissions should be restricted to the approved use case, and high-risk or irreversible actions should require human approval. Organisations should apply risk-based controls to inputs, outputs and tool use, including content filtering, policy enforcement and parameter constraints where appropriate. Untrusted content should be prevented from modifying system instructions, tool definitions and other trusted context. Deployments should be tested for prompt injection, unsafe tool use and data exfiltration. Organisations should treat models with deliberately removed safeguards as materially different from the original release and reassess associated risks before use.

Organisations deploying open-weight models in agentic systems should also refer to ASD's publications *Careful adoption of agentic AI services* and *Agentic AI harnesses*.

Conclusion

Open-weight AI models can offer greater control, flexibility and resilience, but they also transfer responsibility for security, monitoring and lifecycle management to the adopting organisation. Organisations should conduct a risk assessment before deciding whether to use an open-weight AI model, apply controls proportionate to the intended use and sensitivity of the data involved, and comply with applicable frameworks. Organisations that decide to adopt an open-weight model also assume the deployment and operational risks common to AI systems more broadly.

Further information

Artificial intelligence

Artificial intelligence and machine learning: Supply chain risks and mitigations

Careful adoption of agentic AI services

Agentic AI harnesses

Deploying AI systems securely

Managing cyber supply chains

Choosing secure and verifiable technologies

Secure by Design

Selecting, deploying and operating Open-Weight AI Models securely

For more information, or to report a cyber security incident, contact us:

cyber.gov.au

1300 CYBER1 (1300 292 371)

Disclaimer

The material in this guide is of a general nature and should not be regarded as legal advice or relied on for assistance in any particular circumstance or emergency situation. In any important matter, you should seek appropriate independent professional advice in relation to your own circumstances.

The Commonwealth accepts no responsibility or liability for any damage, loss or expense incurred as a result of the reliance on information contained in this guide.

Copyright

© Commonwealth of Australia 2026

With the exception of the Coat of Arms, the entity's logo, third party content, and where otherwise stated, all material presented in this publication is provided under a Creative Commons Attribution 4.0 International license
https://creativecommons.org/licenses/by/4.0/

For the avoidance of doubt, this means this license only applies to material as set out in this document.

The details of the relevant license conditions are available on the Creative Commons website as is the full legal code for the CC BY 4.0 license
https://creativecommons.org/licenses/by/4.0/legalcode.en

Use of the Coat of Arms

The terms under which the Coat of Arms can be used are detailed on the Department of the Prime Minister and Cabinet website
https://www.pmc.gov.au/resources/commonwealth-coat-arms-information-and-guidelines