Australian Government
Australian Signals Directorate

cyber.gov.au

Selecting, deploying and operating open-weight **AI models** securely



Contents

Purpose 3

Audience 3

Introduction 3

Governance 4

Principles for selecting, deploying and operating open-weight models 4

Deployment risks assumed by the adopting organisation 6

Deployment and infrastructure security 7

Model maintenance and updates 7

Monitoring, evaluation and incident response 7

Conclusion 8

Purpose

This publication provides guidance on selecting, deploying, operating, monitoring and maintaining open-weight AI models. It is the companion to *Open-weight AI Models: Opportunities and cyber security risks*, which explains what open-weight models are and the risks that arise specifically from the public release of model weights. This publication assumes the organisation has decided, or is close to deciding, to operate an open-weight model itself, and focuses on the responsibilities that decision creates for governance, selection, deployment architecture, ongoing maintenance, and monitoring and incident response.

Organisations can access open-weight models in several ways, by deploying them locally on infrastructure the organisation controls, whether on-premises or in a public cloud environment; by consuming them through an API offered by the model developer or another hosting provider; or by engaging a third party to host and operate the model on the organisation's behalf. This publication focuses primarily on local deployment, where the organisation operates the model itself. Where a provider hosts and operates the model, including through a hosted API for an open-weight model, the organisation is consuming a hosted AI service. When an open-weight model is consumed as a hosted AI service, many of the security, governance and risk considerations are comparable to those associated with cloud computing or hosted commercial AI services. Organisations should therefore apply relevant cloud security risk management practices, including assessing the provider, supply chain risks, service model, and shared responsibility arrangements. Organisations can further refer to ASD's guidance at Cloud computing and Managing cyber supply chains.

Audience

This guidance is intended for executives, technology leaders and cyber security practitioners responsible for the selection, governance, deployment, operation and security of open-weight AI models within their organisation.

Introduction

Operating an open-weight model places the organisation in the role a hosting provider would otherwise fill. Safety and security functions that a provider performs continuously like safeguard maintenance, abuse detection, security updates, monitoring and incident response do not exist in a self-hosted deployment unless the organisation builds and maintains them. These obligations do not disappear when model is self-hosted rather it transfers them, in full, to the adopting organisation. The guidance below is organised accordingly. Governance establishes who is accountable and how models are approved, reassessed and retired. The selection and operation principles support a proportionate decision about which model to adopt and how to run it. The deployment risk areas describe the specific risks the organisation assumes as an operator, with recommended mitigations.

Governance

Organisations should establish governance arrangements for the adoption and use of AI models, including open-weight AI models. This should include an organisational policy defining approved use cases, security requirements, accountability and risk acceptance processes. Organisations should implement a formal onboarding process to ensure models undergo appropriate security, legal, privacy and operational assessment before deployment.

Organisations should maintain a register of approved models and a technical inventory of deployed model versions, artefacts, adapters, dependencies and configurations. The register should record the approved use case, the accountable owner, the risk acceptance decision and the date of next review. The inventory should be sufficient to determine what is running in which environment, and to support patching, reassessment and incident response. Organisations should establish processes to reassess models following significant updates, fine-tuning or other material modifications.

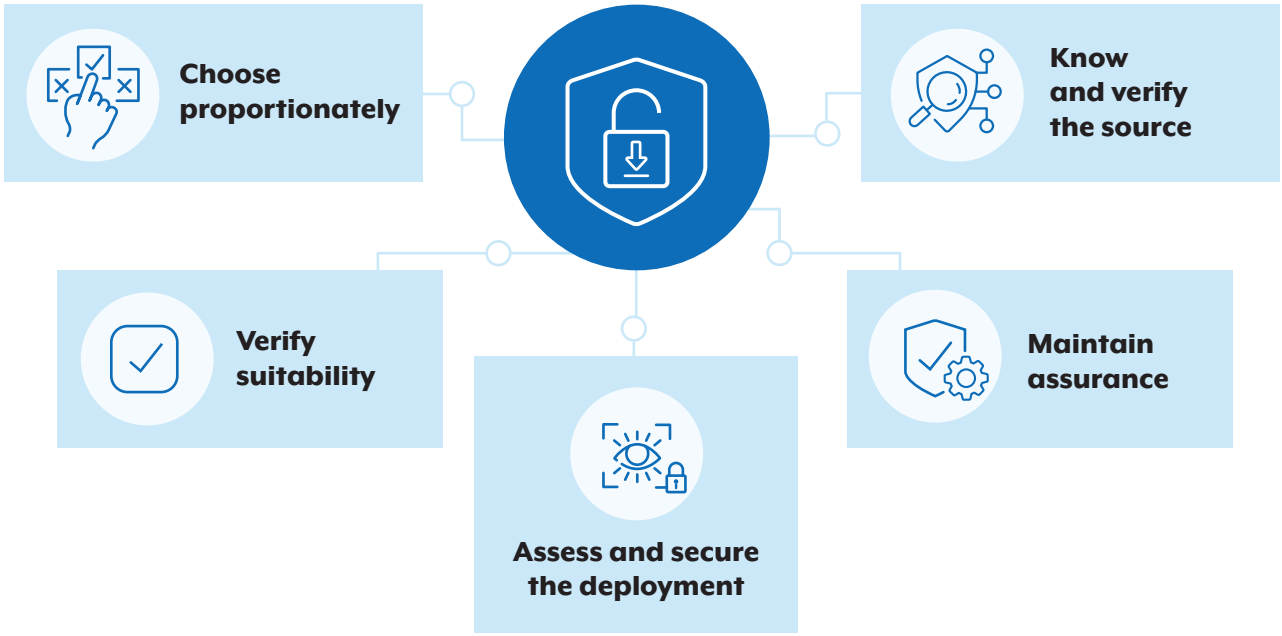
Organisations should establish processes for updating, replacing or retiring open-weight models throughout their operational lifecycle, including evaluating replacement models and revoking unnecessary access when models are decommissioned. Model artefacts, credentials, integrations and related records should be managed securely to support continuity, compliance and incident response requirements.

Principles for selecting, deploying and operating open-weight models

Selecting an open-weight model requires an organisation to make its own assessment of the model's suitability, trustworthiness and security. The following principles support a proportionate selection decision by considering the model's source, capabilities, safeguards, assurance arrangements and associated risks. While these principles reflect good practice for adopting any AI model, open-weight deployment places responsibility for applying them entirely on the adopting organisation.

Choose proportionately. Select a model whose capabilities meet the operational need, provide a cyber security or operational benefit, and align with the organisation's risk tolerance. Consider whether a smaller or specialised model could perform the task effectively with fewer permissions and lower resource requirements. Where a higher-capability model is required, balance its expected benefits against the potential consequences of misuse, or unsafe or biased outputs. Determine

Principles for selecting trusted open-weight models



whether controls available to the organisation can reduce the residual risk to an acceptable level. Where they cannot, the model should not be adopted for that use.

Know and verify the source. Establish who developed, modified and distributed the specific model version. Assess the reputation, transparency and security track record of the developer and distributor, including their country of origin and the jurisdictions in which they operate. Review the model's provenance, lineage, training and alignment information, licence conditions, known limitations and history of security or safety disclosures. Determine whether the model has been fine-tuned, distilled, merged, quantised, converted or otherwise modified. Consider whether applicable laws, government authorities, export controls, licensing requirements or vulnerability disclosure practices could affect the model's availability, maintenance, distribution or use. Country of origin should be considered as one element of this assessment and not, by itself, as evidence that a model is trustworthy or untrustworthy. Where provenance, documentation or assurance information cannot be established to a level appropriate to the intended use, the model should not be adopted.

Verify suitability. Independently test the specific model version to confirm that it performs the intended task reliably and operates within defined security, safety and operational requirements. Test it with inputs that reflect real operating conditions, including the natural languages, data formats and task scenarios the model will encounter in use. Assess whether it produces biased, systematically inaccurate or insecure outputs, follows unauthorised instructions, discloses sensitive information or misuses connected tools. Review available security, safety and red-team evaluations, but do not rely on general benchmarks or malware scanning alone. . Where the model cannot be demonstrated to meet the organisation's security, safety and operational requirements for the intended use, it should not be adopted.

For largely technical or objective tasks, such as coding, data processing or cyber security analysis, ideological bias may have limited impact. However, for tasks involving judgement, advice, decision support, content generation or narrative framing, organisations should carefully consider how known model biases could influence outputs and whether additional controls, oversight or alternative models are warranted.

Assess and secure the deployment. Determine how the model will be hosted, isolated and connected to data, tools and other systems before committing to a deployment design. Identify the network connections and data flows created by the model-serving framework, AI harness, plugins, telemetry, update mechanisms and other supporting components, and establish whether each is necessary and authorised. Establish whether the intended environment can enforce the separation, access control and egress restrictions the use case requires, and confirm that the deployment cannot communicate with unauthorised systems or transfer information outside the approved environment. Do not rely solely on safeguards built into the model: determine which controls must sit outside it, covering inputs, outputs, tool use, and approval of high-risk or irreversible actions. Where the environment or the organisation's control capability cannot meet these requirements, reconsider the deployment model or the model itself.

Maintain assurance. Determine whether the developer, distributor or maintainer provides timely security updates, vulnerability information, updated model versions and clear maintenance or end-of-support information. Identify where the organisation must provide its own monitoring, evaluation, patching and incident response because vendor support, warranties or indemnities are limited or unavailable. Establish what will trigger reassessment or retirement, including changes to the model's weights, adapters, software, safeguards, integrations, licence or distribution arrangements, before the model is approved. Where assurance cannot be obtained from the supplier and cannot be provided by the organisation, the model should not be adopted.

Deployment risks assumed by the adopting organisation

The following risk areas apply to most AI deployments. However, organisations operating open-weight models are solely responsible for managing them, without the monitoring, safeguards and incident response capabilities that a provider would otherwise supply. These practices are described in more detail in ASD's joint publication *Deploying AI systems securely*.

Deployment and infrastructure security

Risk. Operating an open-weight model makes the organisation responsible for securing the inference service and its supporting infrastructure. Exposed interfaces, weak authentication, vulnerable dependencies, excessive network access or poor secrets management may enable unauthorised use, data exfiltration, model theft, denial of service or compromise of connected systems.

Recommended actions. Organisations should place inference services in a separate network segment or security boundary from sensitive systems and data, separate development and production environments, authenticate administrative and inference interfaces and enforce multi-factor authentication for administrative access and apply least privilege. Outbound network access should be controlled, credentials protected and dependencies kept up to date. Rate limits and resource quotas should be applied to inference interfaces to limit the impact of excessive or abusive use.

Organisations should monitor the inference service for compromise, misuse and unexpected behaviour. They should regularly test procedures for rapidly containing an incident, including stopping the service, revoking access and credentials, and isolating affected infrastructure from networks, sensitive data and connected systems.

Model maintenance and updates

Risk. As with any software, organisations operating open-weight models are responsible for tracking new model releases, security updates, vulnerability disclosures and safety improvements. Failure to update models, supporting software or safeguards may expose deployments to known vulnerabilities, degraded performance or emerging threats.

Recommended actions. Organisations should establish processes to monitor developer releases, security advisories and vulnerability disclosures. They should assess, test and approve updates before deployment, and maintain rollback procedures to recover from unsuccessful upgrades. Where a model is no longer maintained by its developer or distributor, organisations should reassess whether continued use remains acceptable and plan for replacement or retirement.

Monitoring, evaluation and incident response

Risk. Open-weight deployments place primary responsibility for monitoring, abuse detection, safety updates and incident response on the operating organisation. Misuse, attempts to steal or replicate the model, data leakage and safeguard bypass attempts may remain undetected unless the operating organisation collects appropriate telemetry and maintains its own detection and response capabilities. The same applies to unexpected behavioural changes, such as topic avoidance, selective refusals or systematically skewed outputs.

Recommended actions. Organisations should log administrative access, security events, model interactions, prompts, responses and tool activity commensurate with risk, while protecting logs that may contain sensitive information and complying with privacy, legal and data handling requirements. Organisations should establish behavioural baselines, monitor for abnormal use, data exfiltration attempts and safeguard bypass attempts, and repeat evaluations after material changes to models, configurations, prompts, tools or connected systems. Tested procedures should be maintained for rollback, model replacement, credential revocation, evidence preservation and incident reporting. Security testing should include adversarial and red-team evaluation techniques where proportionate.

Conclusion

Open-weight AI models can offer greater control, flexibility and resilience, but an organisation that operates a model itself assumes the safety and security responsibilities a provider would otherwise hold. Governance arrangements should establish who is accountable for those responsibilities and how models are approved, reassessed and retired. Selection, deployment and operation should be proportionate to the intended use and the sensitivity of the data involved, and supported by the monitoring, maintenance and incident response capability needed to sustain them.

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

Open-weight AI Models: Opportunities and cyber security risks

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