WHITE PAPER

Air-gapped and on-premise AI with open-weight models

Some data cannot leave the building. This paper explains how to run governed AI workflows inside your own network, on models you host and control: what open-weight models make possible, what you take on in return, and what public guidance says about doing it safely.

Field	Value
Document type	White paper
Audience	Technology, security, risk and procurement leads in business and government
Version	v1.0, 29 November 2026. First public release.
Classification	Public. This document may be shared freely.
Status	Published. Sources last checked on 29 November 2026.
Owner	Agento (Taskoon Pty Ltd, ABN 51 686 768 776) · contact@agento.com.au
Latest version	www.agento.au/air-gapped
Next review	By 31 May 2027, or sooner if the guidance cited here changes

Version history

Version	Date	What changed
v1.0	29 November 2026	First public release

SUMMARY

The short version

- **Some data cannot leave the building.** For classified, regulated and commercially sensitive work, sending prompts and documents to an outside AI service is not an option.
- **Open-weight models make private AI practical.** A model whose weights are published can be downloaded and run on your own hardware, including on a network with no outside connection.
- **One large model is not the only design.** A core model can run the workflow and call smaller specialist models when a step needs their knowledge.
- **An air gap is not governance.** It stops data leaving. It does not decide what the AI may do inside. Permissions, human approval and an audit trail do that.
- **Owning the model brings control and responsibility.** You gain continuity and independence, and you take on the hardware, the security and the upkeep.

What this paper is, and is not

It explains the approach in concept, sets out the trade-offs honestly and summarises the public guidance that applies. It describes how Agento delivers private deployments with customers. It does not describe the internal design of the Agento platform, and it is general information, not legal, security or procurement advice.

Contents

1 Why AI is moving inside the boundary
2 Three ways to deploy
3 Open-weight models
4 A core model and its specialists
5 Fine-tuning
6 Data knowledge graphs
7 Governance inside the air gap
8 The policy context
9 What you take on
10 How Agento delivers it
11 Questions for any supplier
A Sources

SECTION 1

Why AI is moving inside the boundary

Four pressures push organisations to run AI on their own terms.

Pressure	What it means
Classification	Some information may not leave a controlled network. If the AI cannot come to the data, the data cannot be used with AI at all.
Jurisdiction	Data residency is where data physically sits. Data sovereignty is whose laws control access to it.[S11] The two can differ. Under United States law, for example, a provider must disclose data in its possession, custody or control whether that data is held inside or outside the United States.[S12] A local data centre answers the residency question and leaves the sovereignty question open.
What AI adds	The data at issue is no longer only what you store. It is every prompt, every document given to the model, every answer and every log.
Continuity	A capability that depends on an outside service can be changed, repriced or withdrawn by someone else.

The Australian Signals Directorate advises organisations to check that an AI system meets their residency or sovereignty obligations and to find out whether their inputs are used to retrain a provider's model.[S8] Australia's privacy regulator says privacy obligations apply to personal information put into an AI system and to outputs that contain it.[S14] Where personal information is disclosed to an overseas recipient, the organisation generally stays accountable for how it is handled.[S13]

Governments are acting on the same pressures. Section 8 lists dated examples.

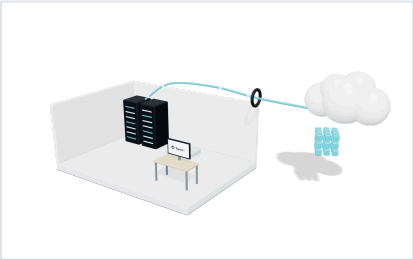
SECTION 2

Three ways to deploy

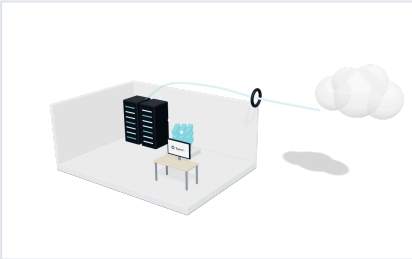
Most organisations that ask for an air-gapped system need one of two private levels. The right one depends on what the data is allowed to touch.

	Hosted cloud	Private deployment	Air-gapped deployment
Where it runs	The supplier's service	Your data centre or your own cloud account	Your own servers, on a network with no outside connection
Outbound connections	Yes	Only to destinations you approve	None
AI models	The supplier's choice of hosted models	Your private models, or providers you approve	Open-weight models on your hardware
Works with	Cloud and on-site tools	The tools you approve	On-site systems only
Suits	Most businesses	Regulated work and data that must stay in-country	Classified, defence and critical infrastructure work

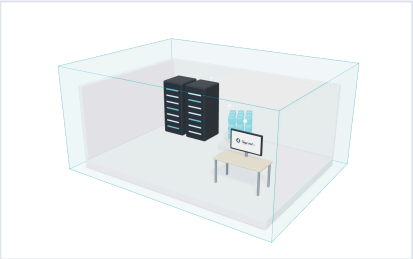
The levels can be reached in stages. Many organisations start with a private deployment, add their own models for the workflows that need them, and disconnect fully only where their rules require it.



1. On-premise
Agento runs on your servers. The AI model can still be an outside service, reached through a connection you approve.



2. Open-weight model
The weights are published, so the model can be downloaded and run inside with everything else.



3. Air-gapped
With the model inside, the outside connection is removed. Nothing leaves and nothing arrives.

An illustration of the idea, not a picture of how the product is built.

SECTION 3

Open-weight models

An AI model is a very large file of numbers called weights. Whoever holds the weights can run the model.

Kind of model	What it means
Closed	The developer keeps the weights. You use the model through their service.
Open weight	The developer publishes the weights. You can download the model and run it on hardware you control.
Open source AI	A stricter standard. The Open Source AI Definition asks for the weights, the code and information about the training data to be available.[S1]

For private deployment, open weight is the line that matters. A model you can only reach through someone else's service cannot run on a network with no outside connection. The licence still matters: it decides whether you may use the model commercially, modify it and pass it on.

Small and large

Open-weight models range from small language models that run on a single server to large general models that need several graphics processors. Small models are fast and cheap for narrow, well-defined tasks. Large models bring broader knowledge and stronger reasoning. A private deployment usually mixes the two.

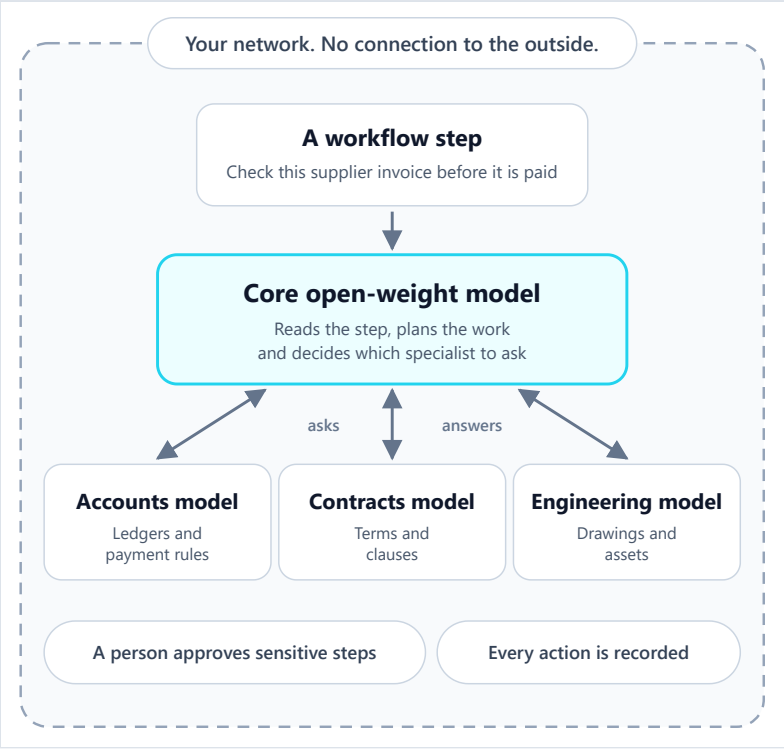
> **Both sides of the ledger**
>
> In November 2026 the Australian Signals Directorate published guidance on open-weight models. It says they can give an organisation greater control and flexibility and support the handling of sensitive data, and that hosting one places responsibility for security, safeguards, monitoring, updates and incident response on the organisation.[S2, S3] A sound plan accounts for both.

SECTION 4

A core model and its specialists

Inside a business, a workflow touches accounts, contracts, engineering records and customer history. Each has its own language and rules. One model does not have to carry all of it.

In the pattern we call specialist model calling, a core model runs the workflow. When a step needs knowledge from one part of the business, the core model asks a model that specialises in that part, over the internal network. A question about payment terms goes to the accounts model. A question about what a contract allows goes to the contracts model. The answer returns to the core model, which moves the workflow on or passes it to a person for approval.



An illustration of the approach. The models and their names are examples.

Why specialists help

- **Better answers on your own material.** A small model set up for one area can outperform a much larger general model on that area.
- **Less hardware.** Researchers have argued that small language models are capable enough, and more economical, for much of the routine work AI agents do.[S4]
- **Clear boundaries.** Access can follow the structure of the business, so the accounts model is the only one that sees the ledger.
- **Easier to improve.** One specialist can be retrained or replaced without touching the rest.

Not every organisation needs a team of models at the start. Where the first workflows sit in one part of the business, a single good open-weight model with the right documents may be enough.

SECTION 5

Fine-tuning

Fine-tuning is further training on your own examples, so a model gets better at one specific job. It is the last tool to reach for. Most gaps close with clearer instructions, the right documents at the right moment, answers in a fixed format, or a better-suited model.

It is worth considering when a task is narrow and repeated, good examples exist, a general model keeps making the same mistake, or a small model could take over a job that currently needs a large one. Modern methods keep the cost down. One widely used approach leaves the original model frozen and trains a small add-on beside it.[S5]

> **The risks, stated plainly**
>
> - A tuned model knows only what it was trained on. Facts that change belong in documents, not in training.
> - A model trained on confidential material should be handled as confidential.
> - When the base model is replaced, the tuning is repeated and tested again.

SECTION 6

Data knowledge graphs

A knowledge graph stores things and the links between them: this customer holds this contract, this contract covers these assets, this asset was inspected on this date. Together the facts form a map of the organisation.

A language model is good with words and knows nothing about your business. Documents help when the answer sits in one place. A graph helps when the answer is spread across many, because the model can follow the links instead of guessing. Research from Microsoft reported that building a graph from a body of documents gave more complete and more varied answers to questions about the collection as a whole than retrieving passages alone.[S6]

Graphs and fine-tuning do different jobs. Facts that change belong in the graph. Skills that stay stable are what fine-tuning is for. In a deployment with specialist models, the graph is what connects their areas.

SECTION 7

Governance inside the air gap

Taking a network offline answers where data can go. It does not answer what the AI may do inside. An agent on a disconnected network can still change the wrong record or act on a document written to mislead it.

National cybersecurity agencies give consistent advice on this. Joint guidance from Australia, the United States, the United Kingdom, Canada and New Zealand on agentic AI recommends limiting first use to low-risk, well-defined tasks and applying strong oversight, continuous monitoring and human control.[S7] Earlier joint guidance on AI systems that an organisation runs itself recommends protecting model weights, restricting network access and keeping human oversight in place.[S9] The risks named in the OWASP list for language-model applications, including prompt injection and excessive agency, are not removed by where the model runs.[S10]

Control	What it does
Permissions	Each agent can reach only the systems and data its job needs
Human approval	Sensitive steps stop and wait for a named person
Audit trail	Every action is recorded: what was done, on what, and on whose approval
Limits and a stop	Agents work within set limits, and a person can halt them

> **The rule for a disconnected network**
>
> Your own team is the only line of oversight. The record of what the AI did has to be complete, and it has to be yours.

SECTION 8

The policy context

Australia

Australian Government agencies are encouraged to use AI under a mandatory policy. Version 2 of the Policy for the responsible use of AI in government took effect on 15 December 2025 and requires, among other things, accountable officials, transparency statements, internal registers of AI use cases and impact assessment.[S15]

What an AI tool may handle depends on the classification of the information. Staff guidance allows public generative AI tools for OFFICIAL information, subject to agency policy.[S16] The Protective Security Policy Framework sets what agencies must do to protect their information.[S17] There is no general rule that government AI must be on-premise. As classification rises, the choice of where a system may run narrows.

> **Two terms that are often misused**
>
> The Hosting Certification Framework certifies hosting providers, not software products.[S19] IRAP is a program under which the Australian Signals Directorate endorses individual assessors. Those assessors do not accredit or endorse systems on its behalf.[S18]

What other governments are doing

When	Where	What was announced
April 2024	United States	Google said its air-gapped cloud product was authorised for Top Secret and Secret workloads[S24]
July 2024	Australia	A partnership worth at least 2 billion Australian dollars over a decade for a sovereign Top Secret cloud[S21]
January 2025	United States	Microsoft said Azure OpenAI Service was authorised for its Top Secret government cloud[S23]
June 2025	Germany	The armed forces' IT provider announced plans for an air-gapped cloud it will operate in its own data centres[S25]
September 2025	United Kingdom	A 400 million pound Ministry of Defence contract for a sovereign cloud on an air-gapped product[S22]
September 2025	United States	Meta's open Llama models made available to all federal departments and agencies[S26]
September 2025	Oman	An Omani language model, Mu'een, launched for use within government first[S27]

The clearest statement of why open weight matters to government is in America's AI Action Plan, which backs open-source and open-weight models and notes that many businesses and governments hold sensitive data they cannot send to the vendors of closed models.[S20]

Several of these sources are a provider's own announcement. The Australian announcement describes a sovereign cloud and does not say it is air-gapped. Agento is not involved in any of these programmes, and nothing here implies endorsement by any government body.

SECTION 9

What you take on

Owning your models is not the right answer for everyone.

Cost	What it involves
Hardware	Servers with graphics processors, and somewhere to run them
People	Someone has to keep the system patched, monitored and backed up
Security	The model files and the servers they run on become assets you must protect
Keeping up	New models arrive often. You decide which upgrades are worth the testing
A gap at the top end	The largest hosted models are still stronger on the hardest open-ended reasoning
Fewer connections	On a disconnected network, tools that live in someone else's cloud cannot be reached

For many workflow steps none of this is a barrier. For some it is, and it is better to find that out in testing than in production.

SECTION 10

How Agento delivers it

Agento runs governed AI workflows. A process is described in plain language, AI agents carry it out across the tools an organisation already uses, a person approves anything sensitive and every action is recorded.

Available now on Enterprise

- Deployment on-premise or in your own cloud account
- Your own private models connected to your workflows
- Human approval on sensitive steps
- A tamper-evident audit trail you can export

Scoped and delivered with you

- Fully disconnected (air-gapped) operation
- Open-weight models on your own hardware
- Specialist models and fine-tuning
- A knowledge graph built from your data

The second list is delivered as an Enterprise project, not switched on from a settings page. It runs in four steps.

Step	What happens
1. Scope	We agree which workflows come first, what data they touch, which deployment level you need and the tests that will prove it works.
2. Choose and size the models	We test candidate open-weight models against your real workflows and size the hardware around the ones that pass.
3. Install and prove	Agento is installed inside your network. Your security team runs the acceptance tests, including a check that nothing connects out.
4. Operate and update	New versions and new models arrive as packages you verify and bring in through your own change process.

Our current compliance position is published in the Trust Center at www.agento.au/trust. We do not claim certifications or assessments we do not hold.

SECTION 11

Questions for any supplier

1. Where is our data processed, including prompts and outputs, and not only where it is stored?
2. Which company operates the AI model, and which country's laws apply to it?
3. Are our inputs used to train or improve any model?
4. What is logged, for how long, and who can read it?
5. Can the system run entirely in our own environment, with a model we host?
6. What can the AI do without a person's approval, and how is that limited?
7. Can we see and export a complete record of what the AI did?
8. Where do model files come from, and how are they verified?
9. How do updates reach a disconnected network, and who tests them?
10. What happens to our data and our models when the contract ends?

> **Talk to us**
>
> Send us the workflows you have in mind and the rules your data has to follow. We will tell you which deployment level fits and what it would take. Email contact@agento.com.au or see www.agento.au/air-gapped.

APPENDIX

Sources

Every link was opened and checked on 29 November 2026.

S1 Open Source Initiative, The Open Source AI Definition 1.0 https://opensource.org/ai/open-source-ai-definition

S2 Australian Signals Directorate, Open-weight AI models: Opportunities and cyber security risks (5 November 2026) https://www.cyber.gov.au/publication/open-weight-ai-models-opportunities-and-cyber-security-risks

S3 Australian Signals Directorate, Open-weight AI models: Selecting, deploying and operating open-weight AI models securely (5 November 2026) https://www.cyber.gov.au/publication/open-weight-ai-models-selecting-deploying-and-operating-open-weight-ai-models-securely

S4 Belcak and others, Small Language Models are the Future of Agentic AI (2025) https://arxiv.org/abs/2506.02153

S5 Hu and others, LoRA: Low-Rank Adaptation of Large Language Models (2021) https://arxiv.org/abs/2106.09685

S6 Edge and others, From Local to Global: A Graph RAG Approach to Query-Focused Summarization (2024) https://arxiv.org/abs/2404.16130

S7 Australian Signals Directorate and partner agencies, Careful adoption of agentic AI services (1 May 2026) https://www.cyber.gov.au/publication/careful-adoption-of-agentic-ai-services

S8 Australian Signals Directorate and partner agencies, Engaging with artificial intelligence (January 2024) https://www.cyber.gov.au/publication/engaging-with-artificial-intelligence

S9 Australian Signals Directorate and partner agencies, Deploying AI systems securely (April 2024) https://www.cyber.gov.au/publication/deploying-ai-systems-securely

S10 OWASP GenAI Security Project, Top 10 for LLM Applications 2026 (August 2026) https://genai.owasp.org/resource/owasp-genai-llm-top-10-2026/

S11 Government of Canada, White Paper: Data Sovereignty and Public Cloud https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/cloud-services/digital-sovereignty/gc-white-paper-data-sovereignty-public-cloud.html

S12 United States Code, title 18, section 2713 https://www.law.cornell.edu/uscode/text/18/2713

S13 Office of the Australian Information Commissioner, Australian Privacy Principles guidelines, Chapter 8: Cross-border disclosure of personal information https://www.oaic.gov.au/privacy/australian-privacy-principles/australian-privacy-principles-guidelines/chapter-8-app-8-cross-border-disclosure-of-personal-information

S14 Office of the Australian Information Commissioner, Guidance on privacy and the use of commercially available AI products (updated 17 January 2025) https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/guidance-on-privacy-and-the-use-of-commercially-available-ai-products

S15 Digital Transformation Agency, Policy for the responsible use of AI in government, version 2.1 https://www.digital.gov.au/ai/ai-in-government-policy

S16 Digital Transformation Agency, Using public generative AI tools safely and responsibly https://www.digital.gov.au/policy/ai/staff-guidance-public-generative-ai

S17 Department of Home Affairs, Protective Security Policy Framework https://www.protectivesecurity.gov.au/pspf-annual-release

S18 Australian Signals Directorate, Infosec Registered Assessors Program https://www.cyber.gov.au/irap

S19 Department of Home Affairs, Hosting Certification Framework https://www.hostingcertification.gov.au/framework

S20 The White House, America's AI Action Plan (July 2025) https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

S21 Prime Minister of Australia, Australian Government partners with Amazon Web Services to bolster national defence and security (4 July 2024) https://www.pm.gov.au/media/australian-government-partners-amazon-web-services-bolster-national-defence-and-security

S22 Google Cloud, Google Cloud Awarded Landmark Sovereign Cloud Contract with UK Ministry of Defence (11 September 2025) https://www.googlecloudpresscorner.com/2025-09-11-Google-Cloud-Awarded-Landmark-Sovereign-Cloud-Contract-with-UK-Ministry-of-Defence

S23 Microsoft Azure Government, Microsoft AI, database, and identity products receive authorization for Top Secret workloads (16 January 2025) https://devblogs.microsoft.com/azuregov/unlocking-new-possibilities/

S24 Google Cloud, Google Public Sector achieves Top Secret and Secret cloud authorization (10 April 2024) https://cloud.google.com/blog/topics/public-sector/google-public-sector-achieves-top-secret-and-secret-cloud-authorization/

S25 Bundeswehr, Partnerschaft von BWI und Google: Die Bundeswehr bekommt eine eigene Cloud (5 June 2025, in German) https://www.bundeswehr.de/de/meldungen/partnerschaft-bwi-google-bundeswehr-eigene-cloud-5952950

S26 US General Services Administration, GSA, Meta Collaborate to Accelerate AI Adoption Across the Government (22 September 2025) https://www.gsa.gov/about-us/newsroom/news-releases/gsa-meta-collaborate-on-ai-adoption-09222025

S27 Oman Observer, First large language model Mu'een launched (8 September 2025) https://www.omanobserver.om/article/1176266/oman/first-large-language-model-mueen-launched