

RESPONSIBLE AI POLICY

AI Governance for Self-Hosted Deployments

How Neuron AI is built, what data it touches, who is accountable for its output, and the boundaries that keep it inside your environment.

Where Neuron AI Stands

Neuron AI is the optional generative module that supplements human authorship inside Neuron. It is licensed and deployed independently from the Core platform; Neuron Core operates without it. When licensed, Neuron AI runs inside the customer's environment and produces draft content that human operators review before it reaches a deliverable.

- **On-premises inference.** Models run on customer hardware. Prompts and completions never leave the customer's trust boundary.
- **Vendor-supplied models.** Models are delivered by PenTest.WS through the update channel. Customers do not supply or fine-tune models.
- **Human approval is structural.** AI output enters the same QA pipeline as manually authored content; visibility is gated by reviewer approval, not by the AI's confidence.
- **No training on customer data.** PenTest.WS does not collect, transmit, or train on customer prompts or completions, and no facility exists in the product to do so.
- **Air-gap supported.** The AI module operates with zero outbound connectivity. License activation and model updates are handled out-of-band when required.

What This Policy Covers

This policy describes the responsible-AI posture of Neuron AI as a product feature: what it is, where it runs, what data it processes, who reviews its output, and the responsibilities that fall to PenTest.WS versus the deploying organisation. It applies to Neuron 3.0 and later, on any supported deployment topology (single host, separate AI host, or air-gapped).

It does not cover the customer's own AI use outside Neuron, third-party AI tools running alongside Neuron, or content authored manually by operators. It also does not impose policy on the customer's internal procedures (testing scope, evidence handling, client communication) — those remain the deploying organisation's responsibility, with Neuron providing the technical controls described here.

Model and Runtime Provenance

Neuron AI runs language models that ship through PenTest.WS's signed update channel. Administrators activate or switch between supplied models from the Neuron admin interface; customers do not load arbitrary models from external sources.

Model family	Transformer language models supplied by PenTest.WS. Sizes are selected to fit a range of customer hardware profiles and may change between Neuron releases. The deployed instance lists its active model in the in-product Model Manager.
Model provenance	Distributed by PenTest.WS through the same signed update channel that delivers application updates. Each model is versioned and tied to a Neuron release.
Hardware	CPU-only inference is the supported default. GPU offload is supported on hardware that allows it. Sizing tiers are documented in the Neuron Server Specifications.
Training	No fine-tuning or model modification is performed by the customer or by PenTest.WS at customer sites. Model behaviour is defined by the published model.
Update lifecycle	New models are released through the update channel. Customers choose when to activate them; previous models can be retained and rolled back.

What the model is not used for

The model is not used to make autonomous decisions about engagement scope, finding severity, client communication, or release of evidence. Every AI output is a draft attached to an operator's record, surfaced to a reviewer before it can be released.

§ 4 — DATA HANDLING

What the Model Sees, and Where It Goes

Where prompts originate

An AI request is initiated by an operator action inside the Neuron application: requesting a finding draft, asking content-assist to rewrite a passage, generating an executive brief, or running post-import analysis on a freshly ingested scan. The application server constructs the prompt from the operator's input plus the engagement context the operator already has access to (the finding being authored, the imported host data, the brief outline). No data crosses an engagement boundary or a tenant boundary that the operator could not already see directly.

Where the prompt goes

The request is processed locally and the completion is returned to the application. Both the request and the completion stay inside the customer's trust boundary for their entire lifecycle.

There is no third-party model API behind Neuron AI. There are no API keys for hosted AI providers. There is no fallback path that escapes the local runtime. The entire inference loop is local-only by construction, not by configuration.

Retention

Completed AI output is stored as part of the Neuron record it was attached to and follows the lifecycle of the underlying engagement record — deleted when the engagement is deleted, exported when the engagement is exported. PenTest.WS does not receive or retain any copy of the request or the response.

Telemetry

The AI module emits no usage telemetry, prompt logs, or completion samples to PenTest.WS or to any third party. Connected installs may check for application updates against the PenTest.WS update server; air-gapped installs do not. Neither path includes any AI request data.

Training

PenTest.WS does not train on customer data. There is no facility in the product to opt into training, opt out of training, or contribute prompts to a shared dataset, because no such pipeline exists. Customer prompts and completions are not used to evaluate, fine-tune, or otherwise modify any model that PenTest.WS distributes to other customers.

§ 5 — HUMAN OVERSIGHT

Where Humans Are in the Loop

Neuron is built around the assumption that the human operator is the author of record, and the AI is a drafting aid that reduces the keystrokes required to express a finding the operator already understands. The product is structured so that AI output cannot bypass review.

Drafting endpoints

AI-drafted findings, custom field values, content rewrites, and brief paragraphs are written into the same fields a human would write into. They become part of the operator's working draft, displayed in the editor, and subject to the operator's edits before save. There is no auto-publish path for AI output.

QA gate to external visibility

Findings, methodology pages, and other outward-facing content reach the customer's Delivery Portal only after passing Neuron's QA workflow. A finding in QA is visible in the portal by name only; the body, severity, evidence, and recommendation are released only after a manager-tier reviewer fully approves the finding. This applies identically to AI-drafted content and human-drafted content. The QA gate is not aware of, and does not differ based on, the origin of the draft.

Executive briefs

Executive briefs are tracked as deliverable content with their own QA status. A draft generated by Neuron AI is held to the same review process as a manually authored brief.

Post-import analysis

Post-import analysis triages freshly ingested scan data and suggests tags on imported hosts and services. These suggestions are advisory: they appear in the operator's review queue and require operator confirmation before they take effect on engagement data. The rule set is editable per instance, so the deploying organisation controls which patterns trigger which tags.

§ 6 — OUTPUT QUALITY & LIMITATIONS

Known Failure Modes and Mitigations

A self-hosted language model is a useful drafting aid; it is not infallible. The product treats output quality as a question of process design, not an in-model promise.

Failure modes the operator should expect

- **Factually incorrect statements.** The model can describe a vulnerability inaccurately, misattribute a CVE, or hallucinate a remediation step. The operator is responsible for verifying technical content against the underlying evidence.

- **Plausible but generic recommendations.** Drafts may default to general best-practice language when specific context would be more useful. Operators are expected to revise toward the engagement's actual environment.
- **Stale knowledge.** The model's knowledge is bounded by its training data; emerging vulnerabilities, new CVEs, and recent vendor advisories may not be represented. Verification against authoritative sources is the operator's responsibility.
- **Prompt sensitivity.** Output quality varies with the quality of the input the operator provides. A finding draft generated against an empty title and no evidence will be lower quality than one generated with structured context.

Product-level mitigations

- Every AI output is presented as a draft inside an editable field, never as a final deliverable.
- The QA workflow gates release; AI-drafted content cannot reach the Delivery Portal without explicit reviewer approval.
- Post-import tagging is advisory; tags do not silently change finding state.
- The audit log records state-changing actions on engagement data, providing visibility into who accepted what content into the record.

Bias and fairness

The models distributed with Neuron AI inherit the strengths and limitations of their upstream provenance. PenTest.WS does not claim that any model is free of cultural, linguistic, or technical bias. Because Neuron AI's outputs are technical drafts authored by a human reviewer before publication, the operator is the bias-mitigation control of last resort: text that does not represent the engagement accurately is rewritten or rejected at the desk where it is reviewed.

§ 7 — ACCEPTABLE USE

What Neuron AI Is For

Neuron AI is intended to support authorised offensive-security engagement work performed under contract or under explicit internal authorisation. The product is sold and licensed to security teams operating in those contexts. The deploying organisation controls which users can invoke AI features and what data those users can submit to it.

Out of scope

Neuron AI is not a general-purpose assistant. It is shaped around drafting and triage tasks for security findings, methodology pages, executive briefs, and post-import analysis. PenTest.WS does not market the AI module for purposes outside that scope and does not warrant its behaviour for unrelated tasks.

Customer responsibilities

- Restrict access to the AI module to operators authorised under the customer's own engagement-handling policies.
- Review AI drafts against engagement evidence before approving content for release.
- Apply the customer's own review and retention policies to AI-drafted content as it would to manually authored content.
- Ensure any data submitted into Neuron, including data that may end up in an AI prompt, is data the operator is authorised to handle on the customer's infrastructure.

Boundaries Around the AI Runtime

The AI module inherits Neuron's overall security posture, which is described in detail in the Neuron Technical Whitepaper. Two points matter most for AI governance: the AI runtime makes no outbound network calls during inference, and air-gapped installs operate with zero outbound connectivity at any phase, including initial install.

Vendor and Customer Responsibilities

The matrix below distinguishes responsibilities that fall to PenTest.WS as the vendor from those that fall to the deploying organisation as the operator of the platform. ● indicates the primary owner of the activity; ○ indicates a contributing role.

ACTIVITY	PENTEST.WS	CUSTOMER
Curating, packaging, and signing the AI runtime and models	●	
Distributing model and runtime updates through the update channel	●	
Activating, switching, or rolling back models in the deployed instance		●
Provisioning hardware that meets the published AI sizing tiers		●
Operating the network boundary that contains the AI runtime		●
Providing product-level controls for human review of AI output	●	
Reviewing AI-drafted content before release		●
Configuring import-analysis rules used to triage scan data	○	●
Restricting AI access to authorised operators		●
Verifying technical accuracy of AI-drafted findings against evidence		●
Auditing and retaining records of accepted AI output	○	●
Investigating and remediating security defects in the AI runtime	●	

When to Contact PenTest.WS

Customers should contact PenTest.WS support if they observe behaviour in the AI module that may indicate a security defect, a confidentiality concern, or a runtime failure that the deployment cannot recover from. Reports should include the Neuron version, the AI model version, and a sanitised description of the observed behaviour. Sample prompts and completions should not be transmitted unless the customer has independently verified that they contain no engagement-sensitive material.