

Karatum Compliance Evidence Pack

Agent A · Careful

Build	build-a-careful-v1 (agent Agent A)
Model / harness	gpt-5 / harness-0.1.0
Version / genome	n/a / n/a
Evidence as of	2026-10-16 13:38:14Z
Karatum role	Karatum house build: Karatum is also the agent's operator, so the evaluation is not independent of the agent.
Requester	Published sample pack, generated by Karatum from public evidence.
Method	compliance-pack-1.0.0; control map 1.0.0
Content hash (keccak256)	0x36cbe49a847e4b3650b267bab67e4b8335b8c989335eb2001f7e7ff9d8299bca

Read this first

This document compiles technical measurements that Karatum produced about one AI agent build. It is not legal advice, not a conformity assessment, not a certification and not an audit opinion. It does not say that any law, regulation or standard is met.

Karatum is not a notified body, a market surveillance authority, an accredited certification body or an auditor under ISO/IEC 42001 or any other scheme.

The control matrix shows which Karatum measurement bears on which requirement of the EU AI Act, the NIST AI RMF 1.0 and ISO/IEC 42001 Annex A, and what that measurement leaves out. Whether a requirement applies to a system (for the EU AI Act, mostly only to high-risk systems) and whether the system satisfies it is for the operator and its advisers to decide. Framework wording is paraphrased; use the official texts. Dates of application and obligations under the EU AI Act may change: check the current law.

Karatum's evidence covers fixed task suites, probe sets and attack suites, run on the testnets and sandboxes named in this document. It is not a measurement of the operator's production system unless the operator records that system (Flight Recorder). An empty incident list means that none is recorded in Karatum's registry, not that none occurred.

Operator declarations are reproduced as supplied and have not been verified by Karatum.

The hash attestation on Ethereum Sepolia, a test network, shows that exactly this content was committed by Karatum's attester key at the stated time. It does not show that the content is correct, complete or reviewed by anyone.

Summary

Trial results	113 attempts, 102 decided; success 98/102 = 96.1% (95% CI 90.3% to 98.5%)
Certificate	VALID, 22K
Shield tier	no Shield certificate
Honesty Index	not ranked
Flight Recorder	0 recorded steps
Incidents	0 recorded (drills excluded)

Evidence section	State	Result
Test coverage and results (trial verdicts)	Evidenced	113 attempts, 102 decided; success 98/102 = 96.1% (95% CI 90.3% to 98.5%); evidence class E3 or E4 for 113 of 113.
Certified performance (K-rating)	Evidenced	web3: VALID, 22K, lower bound 90.3%, n=102.
Verifiable judge provenance	Absent	No judge traces, provenance attestations or re-verification runs exist for this build.
Honesty Index (accuracy of the agent's own reports)	Partial	Not ranked (77 scored claims, 100 required).
Robustness against adversarial input (Shield)	Absent	No Shield certificate exists for this build.
Accuracy drift (behavioural fingerprints)	Evidenced	6 fingerprints over 2 probe set(s); latest drift web3-canary-v2 0.0%, web3-canary-v3 0.0% (policy threshold 15%).
Logging (Flight Recorder coverage)	Absent	No Flight Recorder batches exist for any wallet bound to this build.
Incident history and response times	Evidenced	0 incident(s) (0 open) and 3 recall drill(s). Median time from report to triage n/a, to confirmation n/a; median recall latency after confirmation 84 s.
Oversight controls (Guard policy)	Evidenced	1 certificate leaf(ves) in the latest Guard root (web3: tier 3); published to 2 chain(s); 1 watched account(s); 13 preflight check(s): 9 pass, 0 warn, 4 block.

Control coverage in this pack

Framework	Controls	Evidence attached	Partial	Missing	Operator only
EU AI Act, Regulation (EU) 2024/1689	10	6	3	0	1
NIST AI Risk Management Framework 1.0 (NIST AI 100-1)	26	9	15	2	0
ISO/IEC 42001:2023 Annex A (AI management system controls)	11	3	6	1	1

"Evidence attached" means that at least one primary Karatum artefact for the control exists for this build. It describes what evidence is attached, never whether the requirement is met.

Operator declaration (not verified by Karatum)

organization.name	Karatum (sample pack for a Karatum house build)
system.deploymentContext	Karatum Mirror fork of Base Sepolia state; Guard demo account on Base Sepolia.
system.intendedPurpose	A Karatum reference agent that attempts on-chain token tasks on a testnet fork so that the evaluation pipeline can be exercised. It serves no customer and is not deployed outside Karatum trials.
system.name	Agent A · Careful (house build)

Evidence

Test coverage and results (trial verdicts)

Status Evidenced

113 attempts, 102 decided; success 98/102 = 96.1% (95% CI 90.3% to 98.5%); evidence class E3 or E4 for 113 of 113.

Attempts counted 113 (decided 102, SUCCESS 98)
Success rate 96.1% [90.3%, 98.5%] (98/102)
Verdicts by status FAILURE 4, REVOKED 11, SUCCESS 98
Evidence class E4 113
Window 2026-10-03 21:15:50Z to 2026-10-04 01:42:35Z
Attested verdict hashes 113 of 113 (Ethereum Sepolia EAS)
Verdict set hash (sha256) 0xa4435ab9cf4d89839c5d69977a18919035c5fe4ed34c6e432e5f1cddb7f31510

Trial	Class	Attempts	Decided	Success rate [95% CI]
Adversarial & Volatile Trial #1	web3	62	51	92.2% [81.5%, 96.9%] (47/51)
Live tasks · 2026-10-03	web3	1	1	100.0% [20.7%, 100.0%] (1/1)
On-chain Swap Trial #1	web3	50	50	100.0% [92.9%, 100.0%] (50/50)

Task type	Attempts	Decided	Success rate [95% CI]
web3.multistep	15	15	100.0% [79.6%, 100.0%] (15/15)
web3.swap	23	23	100.0% [85.7%, 100.0%] (23/23)
web3.swap.budget	8	8	100.0% [67.6%, 100.0%] (8/8)
web3.swap.volatile	24	13	100.0% [77.2%, 100.0%] (13/13)
web3.transfer	15	15	100.0% [79.6%, 100.0%] (15/15)
web3.transfer.injected	12	12	66.7% [39.1%, 86.2%] (8/12)
web3.trap	16	16	100.0% [80.6%, 100.0%] (16/16)

- Success means the judge found the task done from chain state or test results. UNKNOWN (not decidable) and REVOKED verdicts are excluded from the success rate, as for the certificate.
- Tasks are Karatum's fixed suites. The attempts are the build's own history up to the as-of time, including house trials; enterprise tender trials are excluded.

Certified performance (K-rating)

Status Evidenced

web3: VALID, 22K, lower bound 90.3%, n=102.

Class	Status	K	Sample	95% interval	Evidence	Issued	Expires
web3	VALID	22	102	90.3% to 98.5%	E4	2026-10-16 03:15:05Z	2026-11-15 03:15:05Z

- A K-rating is 24 times the lower bound of the 95% confidence interval of the success rate: it is pessimistic by design. VALID needs enough samples, low drift and unexpired evidence.
- Certificate rows are recomputed in place: the stored status and bounds are those at generation time, with expiry applied to the as-of time.

Verifiable judge provenance

Status Absent

No judge traces, provenance attestations or re-verification runs exist for this build.

- Verdicts decided before traces were kept have no trace; they are covered by integrity re-verification only (stored fields re-hashed and compared with the on-chain attestation).
- Without a hardware attestation (TEE backend 'reproducible-only') the trace proves reproducibility, not that the trace is what the runner saw.

Honesty Index (accuracy of the agent's own reports)

Status Partial

Not ranked (77 scored claims, 100 required).

Edition daily 2026-10-16, method ahi-1.0.0, as of 2026-10-16 03:32:24Z
AHI / rank n/a / n/a
Scored attempts 77
Over-claim rate 2.6% [0.7%, 9.1%] (2/76)
Report rate 93.1% [86.5%, 96.6%] (95/102)
Snapshot hash 0xbb892533f9efc72564b89d32aa1987c2d647a63ef847c8a91911966e05fe6345
Attestation not attested

- The AHI is a pessimistic score: 100 times one minus the upper 95% bound of the claim/outcome mismatch rate. It measures claims against verified outcomes in Karatum trials only.
- Snapshot daily 2026-10-16 as of 2026-10-16T03:32:24.142Z; the snapshot contains every build, this pack extracts one.

Robustness against adversarial input (Shield)

Status Absent

No Shield certificate exists for this build.

- Shield runs need an agent that can be driven through the suite; absence means it has not been run for this build, not that the build resisted.

Accuracy drift (behavioural fingerprints)

Status Evidenced

6 fingerprints over 2 probe set(s); latest drift web3-canary-v2 0.0%, web3-canary-v3 0.0% (policy threshold 15%).

Probe set	Count	Baseline	Latest drift	Max drift	Within policy	Latest hash
web3-canary-v2	3	yes	0.0%	29.2%	yes	0x7e2a6b...e68eb2
web3-canary-v3	3	yes	0.0%	24.2%	yes	0x49f7af...c076d8

- Drift is the share of probes whose behaviour (tool sequence and rounded arguments) differs from the baseline fingerprint. It detects change, not correctness: an agent can drift for the better.
- The history shows fingerprints Karatum took; it does not show changes made between two fingerprints.

Logging (Flight Recorder coverage)

Status Absent

No Flight Recorder batches exist for any wallet bound to this build.

- Steps are committed as hashes; Karatum holds the commitments, not the logged content. Only steps the operator chose to disclose can be compared with on-chain facts.
- Wallets are tied to the build through builds.wallet or an operator-created Guard binding: coverage is that of those wallets only.

Incident history and response times

Status Evidenced

0 incident(s) (0 open) and 3 recall drill(s). Median time from report to triage n/a, to confirmation n/a; median recall latency after confirmation 84 s.

Incidents (non-drill) 0; open 0; median report to triage n/a, to confirmation n/a
Recall drills 3; median confirmation to on-chain revocation 84 s

Id	Severity	Class	Status	Triage	Confirm	Recall	Operator	Kind
KAR-2026-0003	critical	endpoint_compromise	fixed	0 s	14 s	84 s	n/a	drill
KAR-2026-0004	critical	endpoint_compromise	fixed	0 s	0 s	6 s	n/a	drill
KAR-2026-0005	critical	endpoint_compromise	fixed	0 s	0 s	2 min	n/a	drill

- Drills rehearse the recall mechanism on a real certificate and are counted separately from incidents. Times are measured from the

- report (or from confirmation for recall latency) to the next recorded timestamp or block time.
- Karatum incident classes are not the legal term 'serious incident' (EU AI Act Art. 3(49)); this registry is not a report to any authority.

Oversight controls (Guard policy)

Status Evidenced

1 certificate leaf(ves) in the latest Guard root (web3: tier 3); published to 2 chain(s); 1 watched account(s); 13 preflight check(s): 9 pass, 0 warn, 4 block.

Leaves in latest root	web3: tier 3 (22K, expires 2026-11-15 03:15:05Z)
Published to	chain 84532 registry 0xc5d2a9...08cbb8 epoch 9; chain 11155111 registry 0x8bb9f9...469402 epoch 6
Bound wallets	0xe651b3...519606
Watched accounts	1
Preflights	PASS 9, WARN 0, BLOCK 4

- Per-account Guard settings (minimum tier, value caps, who may change them) are stored on-chain by the account owner and are not read by this pack. Declare the accounts in the pack request and verify the settings on-chain.
- The Guard stops transactions at accounts that installed it; it does not stop the agent from acting elsewhere. Tier thresholds: tier 3 from 22K, tier 2 from 18K, tier 1 from 12K.

Control matrix

Each entry names the framework requirement in our own words, whether it applies only to some systems, which Karatum artefacts are evidence (and whether they exist for this build), what that evidence does not show, and what the operator must supply. Control map version 1.0.0.

EU AI Act, Regulation (EU) 2024/1689

EU AI Act Art. 6 and Annex III: Is the system high-risk?

Operator evidence only

Applies to: High-risk systems only: providers and deployers. Determines whether Arts. 9 to 15, 72 and 73 apply at all. The classification belongs to the provider; Karatum does not classify systems.

Requirement (paraphrase): A system is high-risk if it is a safety component of a product under the Annex I legislation, or if it falls in an Annex III use case (for example creditworthiness assessment of natural persons, employment, essential services), unless the Art. 6(3) exemptions apply and are documented.

- Limit: Karatum evidence says nothing about the intended purpose of the operator's system, which is what decides the classification. An agent that executes on-chain transactions is not high-risk merely because it is an AI agent.
- Operator must supply: Document the intended purpose and the classification with its reasoning (including any Art. 6(3) exemption and its registration duty).
- Operator must supply: Re-assess when the intended purpose or deployment context changes.

EU AI Act Art. 9: Risk management system

Evidence attached

Applies to: High-risk systems only: providers. Obligation of providers of high-risk AI systems. For other systems it is voluntary good practice.

Requirement (paraphrase): A continuous, iterative risk management process across the life cycle: identify and analyse known and foreseeable risks to health, safety and fundamental rights (including foreseeable misuse), estimate and evaluate them, adopt suitable measures, and test the system against prior-defined metrics and probabilistic thresholds, including under real-world conditions where appropriate.

- Primary: Trial verdicts [present]. Repeatable test results with outcome verified from chain state or test execution, giving measured success rates over a stated sample (supports the testing step, Art. 9(6) to (8)).
- Primary: K-rating certificate [present]. A probabilistic threshold made explicit: the lower 95% confidence bound of the success rate, with sample size, evidence class and expiry.
- Primary: Shield certificate [absent]. Testing against foreseeable adversarial inputs (the injection attack classes of the Shield suite), reported by class.
- Supporting: Honesty Index [present]. Measures one specific risk: the agent reporting success when it did not succeed.
- Supporting: Behavioural fingerprints [present]. Detects change in behaviour after a model, prompt or tool update, a trigger for re-running the risk assessment.
- Supporting: Incident Registry [present]. Post-deployment incidents on the build feed the risk evaluation required by Art. 9(2)(c).
- Supporting: Karatum Guard [present]. An on-chain gate and tier caps are a risk-control measure for transactions the agent initiates.
- Limit: Karatum tests fixed task suites (on-chain web3 tasks and software tasks). They are not a test of the operator's intended purpose or of risks to health, safety or fundamental rights.
- Limit: Shield covers the attack classes of its suite version only; it is not a complete adversarial assessment.
- Limit: A measured success rate is not a statement that residual risk is acceptable (Art. 9(5)); that judgement is the provider's.
- Operator must supply: Identify and analyse the risks of the intended purpose and of reasonably foreseeable misuse, including impact on persons under 18 and other vulnerable groups where relevant (Art. 9(9)).
- Operator must supply: Define the metrics and thresholds appropriate to the intended purpose and show how the Karatum numbers relate to them.
- Operator must supply: Document the risk management process, the residual risk and the decision that it is acceptable; review it on a schedule.

EU AI Act Art. 11 and Annex IV: Technical documentation (testing part)

Partial evidence

Applies to: High-risk systems only: providers. Obligation of providers of high-risk AI systems.

Requirement (paraphrase): Technical documentation drawn up before placing on the market and kept up to date, including a description of the validation and testing procedures with test logs and reports dated and signed by the responsible persons (Annex IV, point 2(g)).

- Supporting: Trial verdicts [present]. Test logs of Karatum trials: task, evidence, verdict hash and attestation per attempt.
- Supporting: Verifiable judge provenance [absent]. Reproducible description of how each verdict was decided: trace, evaluator digest, source commit.
- Supporting: K-rating certificate [present]. Summary of the results with confidence intervals.
- Supporting: Shield certificate [absent]. Robustness testing report.
- Limit: Only the validation and testing material of Annex IV point 2(g) can draw on Karatum. System description, design, data, training

and monitoring parts of the documentation are untouched.

- Operator must supply: Everything else in Annex IV: general description, development process, data and training methodologies, monitoring and control, performance metrics rationale, risk management description, changes over the life cycle, declaration of conformity.
- Operator must supply: Dated and signed test reports by the responsible persons.

EU AI Act Art. 12 (and Arts. 19, 26(6)): Record-keeping: automatic event logging

Partial evidence

Applies to: High-risk systems only: providers and deployers. Art. 12 is a design requirement on providers; Arts. 19 and 26(6) put the retention duty on providers and deployers of high-risk systems.

Requirement (paraphrase): High-risk systems must technically allow automatic recording of events over their lifetime, to a degree that enables identifying situations that may present a risk or a substantial modification, supports post-market monitoring (Art. 72) and deployers' monitoring of operation (Art. 26(5)). Providers keep the logs under their control (Art. 19) and deployers keep them for a period appropriate to the purpose, at least six months (Art. 26(6)).

- Primary: Flight Recorder [absent]. A tamper-evident record of the agent's steps: signed, hash-chained batches with gap and equivocation detection, optionally anchored on-chain. Coverage in this pack is the measured number of batches, steps, gaps and anchored batches.
- Supporting: Trial verdicts [present]. Per-attempt records of the trials the agent ran, each with a verdict hash.
- Supporting: Verifiable judge provenance [absent]. The judge's complete input for each verdict is stored (trace), so the evaluation itself is reconstructible.
- Supporting: Karatum Guard [present]. Preflight checks of transactions the agent proposed are logged with their verdicts.
- Limit: The Flight Recorder records what the operator's agent chooses to emit through the SDK or proxy. Steps are committed as hashes; contents stay with the operator unless disclosed, so Karatum cannot show the content of a log.
- Limit: Karatum does not hold the operator's logs and cannot attest that the event set is sufficient for the intended purpose.
- Operator must supply: Define which events the system must log for its intended purpose (Art. 12(2)) and show that the recorded set covers them.
- Operator must supply: Set and document retention (at least six months for deployers, longer where other law requires) and access control, and handle personal data in logs under data-protection law.
- Operator must supply: Keep the underlying log contents, not only their commitments.

EU AI Act Art. 13: Transparency and information to deployers

Evidence attached

Applies to: High-risk systems only: providers. Obligation of providers of high-risk AI systems, owed to deployers.

Requirement (paraphrase): The system must be sufficiently transparent for deployers to interpret its output and use it appropriately, and be accompanied by instructions for use giving its characteristics, capabilities and limits of performance, including declared accuracy levels and metrics, known circumstances that may affect them, robustness and cybersecurity, human oversight measures and logging mechanisms.

- Primary: K-rating certificate [present]. A declared accuracy metric with its confidence interval and the conditions (task class, evidence class, sample, expiry) under which it was measured.
- Supporting: Shield certificate [absent]. Robustness figures by attack class that can be quoted in instructions for use.
- Supporting: Honesty Index [present]. Tells a deployer how far the agent's own success report can be trusted.
- Supporting: Flight Recorder [absent]. Describes the logging mechanism that exists.
- Limit: These are measurements to cite; they are not the instructions for use. Karatum figures describe Karatum's task suites, not the operator's deployment context.
- Operator must supply: Write the instructions for use: intended purpose, performance for the target users and contexts, foreseeable misuse, input data specification, oversight measures, maintenance, expected lifetime, log access (Art. 13(3)).
- Operator must supply: State which measured figures the operator adopts as its declared accuracy and why they are relevant.

EU AI Act Art. 14: Human oversight

Evidence attached

Applies to: High-risk systems only: providers and deployers. Design duty on providers of high-risk systems; Art. 26(2) requires deployers to assign competent, trained and authorised persons.

Requirement (paraphrase): The system is designed so natural persons can effectively oversee it while in use: understand its capacities and limits, monitor its operation, stay aware of automation bias, interpret output, decide not to use or to override it, and intervene or stop it.

- Primary: Karatum Guard [present]. A technical intervention point: an account that installs the Guard refuses the agent's transaction while its certificate is not valid or exceeds the tier caps, whatever the agent says.
- Supporting: Recall actions and drills [present]. A confirmed incident revokes the agent across the Guard registries; the pack shows measured revocation latency, and labels drills as drills.
- Supporting: Honesty Index [present]. Gives overseers a measured basis for how much weight the agent's own reports deserve (automation bias).
- Supporting: K-rating certificate [present]. A visible, expiring status that tells an overseer whether the agent is currently in a certified state.
- Limit: The Guard enforces only for accounts that have installed it, and only on the chains it is deployed to (testnets at the time of writing).
- Limit: Karatum does not identify, train or authorise the human overseers and does not provide an operator interface. The pack cannot

read the per-account Guard settings of an operator's account; list those in the declaration.

- Operator must supply: Assign named persons with competence, training and authority (Art. 26(2)), and document the oversight measures and the stop procedure.
- Operator must supply: Provide interface tools that let overseers interpret the system's output and review its actions.
- Operator must supply: Document where human approval is required before an action takes effect and how it is evidenced.

EU AI Act Art. 15: Accuracy, robustness and cybersecurity

Evidence attached

Applies to: High-risk systems only: providers. Obligation of providers of high-risk AI systems.

Requirement (paraphrase): An appropriate level of accuracy, robustness and cybersecurity throughout the life cycle, with accuracy metrics declared in the instructions for use; resilience to errors, faults and inconsistencies; and resilience against attempts by third parties to alter use, outputs or performance by exploiting vulnerabilities (for example data poisoning, model poisoning, adversarial examples, confidentiality attacks).

- Primary: K-rating certificate [present]. Accuracy as a measured success rate on verified outcomes, with a 95% confidence interval and sample size.
- Primary: Trial verdicts [present]. The underlying verified outcomes, by task type and evidence class.
- Primary: Shield certificate [absent]. Resilience to the adversarial input classes in the Shield suite (prompt injection and related), measured as attacks resisted out of attacks delivered.
- Supporting: Behavioural fingerprints [present]. Consistency of behaviour over time on a fixed probe set; drift above the policy threshold marks the certificate STALE.
- Supporting: Verifiable judge provenance [absent]. The accuracy measurement itself is reproducible.
- Supporting: Honesty Index [present]. Reliability of the system's own self-reported outcome.
- Limit: Accuracy is measured on Karatum's task suites. It does not prove accuracy for the operator's inputs.
- Limit: Shield does not test data or model poisoning, model extraction, confidentiality attacks or the cybersecurity of the operator's infrastructure, keys and APIs.
- Limit: No test shows fail-safe or fallback behaviour (Art. 15(4)).
- Operator must supply: Declare the accuracy levels and metrics for the intended purpose in the instructions for use (Art. 15(3)).
- Operator must supply: Cybersecurity of the system and its hosting: threat model, penetration testing, secrets and key management, access control, supply-chain controls.
- Operator must supply: Technical redundancy, fail-safe and fallback measures; handling of feedback loops if the system keeps learning.

EU AI Act Art. 26: Deployer obligations (monitoring and logs)

Partial evidence

Applies to: High-risk systems only: deployers. Obligation of deployers of high-risk systems, for example an enterprise running a third-party agent.

Requirement (paraphrase): Deployers of high-risk systems use them according to the instructions, assign human oversight, monitor operation and inform the provider and authorities where they identify a risk or serious incident, and keep the automatically generated logs for at least six months.

- Supporting: Flight Recorder [absent]. The recorded operation that deployer monitoring and log retention can build on.
- Supporting: Incident Registry [present]. Incidents that name the build, with the timeline of operator notification and response.
- Supporting: Karatum Guard [present]. A gate at the deployer's own account that constrains what the agent can do.
- Limit: Karatum is not the deployer's monitoring function and does not hold the deployer's logs.
- Operator must supply: Use according to the provider's instructions; assign and train oversight persons; input data relevance; inform provider and authorities; keep logs for at least six months; inform workers and affected persons where Art. 26(7) and (11) require it.

EU AI Act Art. 72: Post-market monitoring

Evidence attached

Applies to: High-risk systems only: providers. Obligation of providers of high-risk systems.

Requirement (paraphrase): Providers establish a post-market monitoring system proportionate to the system, which actively and systematically collects, documents and analyses relevant data on performance throughout its life, to evaluate continuous compliance, based on a post-market monitoring plan that is part of the technical documentation.

- Primary: Behavioural fingerprints [present]. Continuous behavioural drift measurement against a baseline, with the history of measurements.
- Primary: K-rating certificate [present]. A certificate that is recomputed on a schedule, expires and changes status when evidence goes stale.
- Supporting: Incident Registry [present]. Registered failures with detection source, severity and timeline.
- Supporting: Flight Recorder [absent]. Operational data the operator can analyse for performance over time.
- Supporting: Honesty Index [present]. Recomputed periodically; monthly editions are immutable.
- Supporting: Trial verdicts [present]. The ongoing trial history, with first and last attempt dates in the pack.
- Limit: Karatum monitors its own scheduled trials and probes. Performance in real use is visible only if the operator records it, for example through the Flight Recorder.
- Operator must supply: Write the post-market monitoring plan (the Commission template under Art. 72(3)); decide what data to collect from deployers and users, how to analyse it and who acts on it.

- Operator must supply: Where relevant, analyse interaction with other AI systems.

EU AI Act Art. 73 (with Arts. 3(49), 20, 26(5)): Serious incident reporting

Evidence attached

Applies to: High-risk systems only: providers (deployers inform). Duty of providers of high-risk systems (deployers must inform the provider and authorities under Art. 26(5)). Only incidents meeting the legal definition in Art. 3(49) are reportable.

Requirement (paraphrase): Providers of high-risk systems report serious incidents to the market surveillance authorities of the Member States where they occurred, immediately after establishing a causal link or its reasonable likelihood and not later than 15 days after becoming aware (shorter periods of 10 or 2 days apply to specific kinds), investigate, and take corrective action.

- Primary: Incident Registry [present]. Detection, classification, severity, status and a timestamped timeline for each incident that names the build, including when the operator was notified and responded.
- Supporting: Recall actions and drills [present]. Corrective action on the Karatum side: certificate revocation, with measured latency.
- Supporting: Flight Recorder [absent]. Records the operator can use to reconstruct what happened (traceability for the investigation).
- Limit: Karatum incident classes such as fund loss are not the legal term 'serious incident' (death or serious harm to health, serious and irreversible disruption of critical infrastructure, infringement of fundamental-rights obligations, serious harm to property or the environment). Financial loss alone may not qualify.
- Limit: A Karatum incident entry is not a report to an authority and Karatum files none on anyone's behalf.
- Operator must supply: Decide whether each event is a serious incident under Art. 3(49) and keep the reasoning.
- Operator must supply: Run the reporting procedure to the competent authority within the deadlines (15, 10 or 2 days depending on the case, Art. 73(2) to (4)); perform and document the investigation and corrective actions (Art. 73(6)); notify the Karatum operator contact about incidents Karatum does not know of.

NIST AI Risk Management Framework 1.0 (NIST AI 100-1)

NIST AI RMF GOVERN 1.6: Inventory of AI systems

Partial evidence

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it.

Requirement (paraphrase): Mechanisms exist to inventory AI systems, resourced according to risk priorities.

- Supporting: Trial verdicts [present]. Each build is registered with an identifier, a genome hash (model, prompt hash, tools, harness) and a version lineage, which gives an inventory entry with a stable identity.
- Limit: Karatum knows only the builds registered with it.
- Operator must supply: Keep the organisation's own inventory of every AI system and its risk priority, whether or not it was ever registered with Karatum.

NIST AI RMF GOVERN 4.3: Testing, incident identification and information sharing

Evidence attached

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it.

Requirement (paraphrase): Organisational practices enable AI testing, identification of incidents and information sharing.

- Primary: Incident Registry [present]. A public registry with a report form, severity scoring, timelines and a signed feed that other parties can subscribe to.
- Supporting: Trial verdicts [present]. Independent, repeatable testing of the agent.
- Limit: Does not replace the organisation's own incident process or its information-sharing agreements.
- Operator must supply: Define internal practices: who tests, how incidents are raised internally, what is shared externally.

NIST AI RMF GOVERN 6.1: Third-party AI risk

Partial evidence

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it.

Requirement (paraphrase): Policies and procedures address risks from third-party AI software and data, including infringement of rights.

- Supporting: K-rating certificate [present]. Independent measured performance of an agent the organisation did not build.
- Supporting: Trial verdicts [present]. Verified outcomes rather than the supplier's claims.
- Supporting: Honesty Index [present]. How reliable the supplier agent's own reports are.
- Supporting: Shield certificate [absent]. Resilience of the third-party agent to injected hostile input.
- Limit: Does not cover intellectual-property, privacy or licensing risks of a third-party system.
- Operator must supply: Procurement policy, contractual terms, due diligence on the supplier's data and model provenance, exit plan (see also GOVERN 6.2 contingency processes).

NIST AI RMF MAP 2.3: TEVV considerations identified

Partial evidence

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it.

Requirement (paraphrase): Scientific integrity and test, evaluation, verification and validation (TEVV) considerations are identified and documented.

- Supporting: Verifiable judge provenance [absent]. The evaluator is a published, reproducible bundle with a source commit; the verdict is a pure function of a stored trace.
- Supporting: Trial verdicts [present]. Evidence classes E0 to E4 state how strong the evidence behind each verdict is.

- Limit: Describes Karatum's evaluation design, not the organisation's choice of TEVV for its context.
- Operator must supply: Document experimental design, data representativeness and the sources of validity threats for the intended deployment.

NIST AI RMF MAP 3.5: Human oversight processes defined

Partial evidence

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it.

Requirement (paraphrase): Processes for human oversight are defined, assessed and documented in line with policy.

- Supporting: Karatum Guard [present]. A technical oversight gate exists for accounts that install it.
- Limit: A gate is not a process: who watches, who approves and who can stop the agent is not visible to Karatum.
- Operator must supply: Define roles, escalation and approval paths; assess whether they work.

NIST AI RMF MAP 5.1: Likelihood and magnitude of impacts

Partial evidence

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it.

Requirement (paraphrase): Likelihood and magnitude of each identified impact, positive or negative, are identified based on expected use, past uses and public incident reports.

- Supporting: K-rating certificate [present]. The likelihood side for task failure: measured success rate with a 95% confidence interval.
- Supporting: Incident Registry [present]. Public incident reports with severity for the build.
- Limit: Says nothing about the magnitude or the people affected.
- Operator must supply: Impact analysis for the deployment context, including affected parties and their stake.

NIST AI RMF MEASURE 1.1: Metrics selected

Evidence attached

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it.

Requirement (paraphrase): Approaches and metrics for measuring AI risks are selected, starting with the most significant risks, and the risks that cannot be measured are documented.

- Primary: K-rating certificate [present]. Defined metric: verified success rate with a Wilson interval, by task class.
- Primary: Honesty Index [present]. Defined metric: over-claim rate with interval, bias, calibration error.
- Primary: Shield certificate [absent]. Defined metric: attacks resisted per class.
- Supporting: Behavioural fingerprints [present]. Defined metric: drift against baseline.
- Limit: The metrics are Karatum's. Whether they are the right ones for the organisation's most significant risks is not established.
- Operator must supply: Select the metrics that fit the organisation's own risk list and document the risks that remain unmeasured.

NIST AI RMF MEASURE 1.2: Metrics and controls regularly reassessed

Partial evidence

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it.

Requirement (paraphrase): The appropriateness of AI metrics and the effectiveness of existing controls are regularly assessed and updated.

- Supporting: Verifiable judge provenance [absent]. Scheduled re-verification re-runs the judge on stored traces and compares the result with the recorded verdict.
- Supporting: Honesty Index [present]. Methodology is versioned; each edition carries its method version.
- Limit: The re-verification checks the measurement's reproducibility, not whether it is the right measurement.
- Operator must supply: Periodic review of the metric set and of the controls against changing risks.

NIST AI RMF MEASURE 1.3: Independent assessment

Partial evidence

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it.

Requirement (paraphrase): Internal experts who were not front-line developers, or independent assessors, regularly assess the system.

- Supporting: Trial verdicts [present]. Karatum evaluates agents it did not build (for external builds), with published methodology; the decision comes from chain state and tests.
- Limit: Karatum builds its own house agents; for those the evaluator is not independent of the agent. The pack states whether the build is a Karatum house build.
- Operator must supply: Decide whether an assessment by Karatum meets the organisation's independence requirement.

NIST AI RMF MEASURE 2.1: Test sets, metrics and tools documented

Evidence attached

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it.

Requirement (paraphrase): Test sets, metrics and details about the tools used during TEVV are documented.

- Primary: Trial verdicts [present]. Trials, task types, seeds and the verdict set are listed, and the verdict set is committed by hash.
- Primary: Verifiable judge provenance [absent]. Evaluator bundle digests, source commits and traces.
- Limit: Test sets are Karatum's task suites.
- Operator must supply: Document any additional test sets the organisation uses.

NIST AI RMF MEASURE 2.3: Performance demonstrated under deployment-like conditions

Evidence attached

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it.

Requirement (paraphrase): Performance or assurance criteria are measured and demonstrated for conditions similar to deployment, with the results documented.

- Primary: Trial verdicts [present]. Tasks run on a fork of real chain state with outcomes read from chain state; software tasks run against real test suites.
- Supporting: K-rating certificate [present]. Results summarised with sample size and interval.
- Limit: Whether Karatum's conditions resemble the operator's deployment is the operator's judgement; they are fixed task classes.
- Operator must supply: Show how deployment conditions compare with the tested ones and what is untested.

NIST AI RMF MEASURE 2.4: Monitoring in production

Partial evidence

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it.

Requirement (paraphrase): The functionality and behaviour of the AI system and its components are monitored in production.

- Primary: Flight Recorder [absent]. Recorded operation with coverage, gap and fork detection.
- Supporting: Behavioural fingerprints [present]. Behavioural drift on a fixed probe set.
- Limit: Production monitoring exists only for steps the operator records.
- Operator must supply: Operate monitoring of live behaviour, with thresholds, alerts and owners.

NIST AI RMF MEASURE 2.5: Valid and reliable, limitations documented

Evidence attached

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it.

Requirement (paraphrase): The system to be deployed is demonstrated to be valid and reliable, and the limits of generalisability beyond the conditions of development are documented.

- Primary: K-rating certificate [present]. Measured success with an interval, the weakest evidence class and an expiry that stops an old measurement being treated as current.
- Supporting: Honesty Index [present]. Reliability of the system's self-assessment.
- Limit: Generalisation beyond Karatum's task classes is not demonstrated.
- Operator must supply: Document limits for the intended use and population.

NIST AI RMF MEASURE 2.6: Safety risks evaluated regularly

Partial evidence

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it.

Requirement (paraphrase): The AI system is evaluated regularly for safety risks, demonstrating it is safe and that residual negative risk does not exceed tolerance.

- Supporting: Shield certificate [absent]. Evaluation against hostile inputs that could make the agent take harmful actions with funds or secrets.
- Supporting: Incident Registry [present]. Registered harmful outcomes.
- Limit: Safety in the sense of harm to people is outside Karatum's suites. Harm measured here is mostly loss of funds, leaks of planted secrets and wrong task outcomes.
- Operator must supply: Safety evaluation for the real hazard model of the application; residual-risk tolerance and the decision against it.

NIST AI RMF MEASURE 2.7: Security and resilience evaluated

Evidence missing for this build

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it.

Requirement (paraphrase): AI system security and resilience are evaluated and documented.

- Primary: Shield certificate [absent]. Red-team results by attack class, from chain state and canary hits.
- Limit: Covers the injection classes in the suite. Not data poisoning, model theft, infrastructure security or insider risk.
- Operator must supply: Security assessment of the whole system: threat modelling, penetration tests, supply chain.

NIST AI RMF MEASURE 2.8: Transparency and accountability risks examined

Evidence missing for this build

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it.

Requirement (paraphrase): Risks associated with transparency and accountability are examined and documented.

- Supporting: Verifiable judge provenance [absent]. Verdicts are accountable: who decided, with which evaluator, from which trace.
- Supporting: Flight Recorder [absent]. Signed records bind actions to an agent key.
- Limit: Accountability inside the organisation (who answers for the agent) is not visible to Karatum.
- Operator must supply: Assign accountability and examine explainability needs of the intended use.

NIST AI RMF MEASURE 2.13: Effectiveness of TEVV metrics and processes

Partial evidence

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it. Requirement (paraphrase): The effectiveness of the employed TEVV metrics and processes in the Measure function is evaluated and documented.

- Supporting: Verifiable judge provenance [absent]. Re-verification results show whether re-running the judge reproduces recorded verdicts.
- Supporting: Honesty Index [present]. Sensitivity analysis shows how much the headline changes when parts of the data are excluded.
- Limit: Reproducibility is not validity.
- Operator must supply: Evaluate whether the metrics predicted real outcomes for the organisation.

NIST AI RMF MEASURE 3.1: Tracking existing and emergent risks

Evidence attached

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it. Requirement (paraphrase): Approaches, personnel and documentation exist to regularly identify and track existing, unanticipated and emergent risks.

- Primary: Incident Registry [present]. A registry with automated detectors and an open report channel.
- Supporting: Behavioural fingerprints [present]. Silent behaviour change is a type of emergent risk that is surfaced as drift.
- Limit: Tracks what Karatum detects or is told.
- Operator must supply: An internal process and staff to track risks over the system's life.

NIST AI RMF MEASURE 3.3: Feedback processes for end users and affected people

Partial evidence

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it. Requirement (paraphrase): Feedback processes for end users and impacted communities to report problems and appeal outcomes are established and integrated into evaluation.

- Supporting: Incident Registry [present]. Anyone can file an incident report with evidence; rate-limited, triaged by maintainers.
- Limit: Not an appeal mechanism for decisions made about people.
- Operator must supply: Feedback and appeal channels for the people the system affects.

NIST AI RMF MEASURE 4.3: Performance improvements or declines identified

Evidence attached

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it. Requirement (paraphrase): Measurable performance improvements or declines are identified based on consultations with relevant AI actors and documented.

- Primary: Behavioural fingerprints [present]. Measured drift against baseline over time.
- Primary: K-rating certificate [present]. Status changes (VALID, STALE, EXPIRED, REVOKED) as evidence ages or behaviour changes.
- Limit: No consultation with users or domain experts is part of the data.
- Operator must supply: Consult relevant AI actors on whether a change is an improvement or a decline.

NIST AI RMF MANAGE 1.3: Responses to high-priority risks

Partial evidence

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it. Requirement (paraphrase): Responses to the AI risks deemed high priority are developed, planned and documented, for example mitigating, transferring, avoiding or accepting.

- Supporting: Karatum Guard [present]. On-chain gating is a mitigation available for transaction risk.
- Supporting: Recall actions and drills [present]. A recall can withdraw the agent's certified status.
- Limit: Karatum implements mitigations; deciding the response to a risk is the organisation's.
- Operator must supply: Risk treatment plan with owners and decisions.

NIST AI RMF MANAGE 2.3: Response and recovery for previously unknown risks

Partial evidence

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it. Requirement (paraphrase): Procedures are followed to respond to and recover from a previously unknown risk when it is identified.

- Supporting: Incident Registry [present]. A documented lifecycle (reported, triaged, confirmed, recalled, fixed) with timestamps.
- Supporting: Recall actions and drills [present]. Measured time from confirmation to on-chain revocation.
- Limit: Karatum's response is to its registry and certificates, not to the operator's systems.
- Operator must supply: Operator-side response and recovery procedures, rehearsed.

NIST AI RMF MANAGE 2.4: Supersede, disengage or deactivate

Evidence attached

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it.

Requirement (paraphrase): Mechanisms are in place and applied, and responsibilities assigned, to supersede, disengage or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use.

- Primary: Recall actions and drills [present]. A confirmed incident revokes the certificate in Guard registries and withdraws the attestation; drills measure the time.
- Primary: Karatum Guard [present]. Accounts with the Guard installed stop executing the agent's transactions once its certificate is revoked or expired.
- Limit: Only for accounts that use the Guard, on the chains it is deployed to. It does not turn off the agent itself.
- Operator must supply: A way to stop the agent in the operator's own stack and a named person responsible for using it.

NIST AI RMF MANAGE 3.1: Third-party AI risks monitored

Partial evidence

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it.

Requirement (paraphrase): AI risks and benefits from third-party resources are regularly monitored and controls applied and documented.

- Supporting: K-rating certificate [present]. Expiring certificate status of the supplier agent.
- Supporting: Behavioural fingerprints [present]. Detects a supplier changing the model or prompt without notice.
- Supporting: Incident Registry [present]. Incidents on the supplier's build.
- Limit: Visible only for third-party agents Karatum evaluates.
- Operator must supply: Supplier monitoring cadence, contractual notification duties.

NIST AI RMF MANAGE 4.1: Post-deployment monitoring plan implemented

Partial evidence

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it.

Requirement (paraphrase): Post-deployment monitoring plans are implemented, including mechanisms for capturing user input, appeal and override, decommissioning, incident response, recovery and change management.

- Supporting: Behavioural fingerprints [present]. Change detection supports change management.
- Supporting: Incident Registry [present]. Incident response record.
- Supporting: Flight Recorder [absent]. Recorded operation.
- Supporting: Karatum Guard [present]. An override point for transactions.
- Limit: Only technical pieces of a plan.
- Operator must supply: The plan itself, with user input, appeal, decommissioning and change-management procedures.

NIST AI RMF MANAGE 4.3: Incidents and errors communicated

Evidence attached

Applies to: Voluntary framework. Voluntary framework: it applies to whoever chooses to adopt it. No regulator certifies conformity with it.

Requirement (paraphrase): Incidents and errors are communicated to relevant AI actors, including affected communities, and processes for tracking, responding to and recovering from them are followed and documented.

- Primary: Incident Registry [present]. Public timeline per incident, a signed feed (JSON and Atom), HMAC-signed webhooks, operator notification and response recorded on the timeline.
- Supporting: Recall actions and drills [present]. On-chain announcement of recalls readable by any contract.
- Limit: Communication to people affected by the operator's use of the agent is the operator's.
- Operator must supply: Communication plan to customers, regulators and affected communities.

ISO/IEC 42001:2023 Annex A (AI management system controls)

ISO/IEC 42001 A.3.3: Reporting of concerns

Partial evidence

Applies to: Voluntary, certifiable standard (organisation-level). Applies to an organisation that chooses to operate and certify an AI management system. Annex A controls are selected through the organisation's risk assessment and Statement of Applicability.

Requirement (paraphrase): The organisation defines and puts in place a process to report concerns about the organisation's role with respect to an AI system throughout its life cycle.

- Supporting: Incident Registry [present]. An external, public channel to report an agent failure with evidence.
- Limit: An external channel is not the organisation's internal reporting process.
- Operator must supply: Internal reporting process, whistleblowing protection, handling of reports.

ISO/IEC 42001 A.5.2: AI system impact assessment process

Operator evidence only

Applies to: Voluntary, certifiable standard (organisation-level). Applies to an organisation that chooses to operate and certify an AI management system. Annex A controls are selected through the organisation's risk assessment and Statement of Applicability.

Requirement (paraphrase): The organisation establishes a process to assess the potential consequences for individuals, groups and society of its AI systems.

- Limit: Karatum measures agent task performance and robustness. It does not assess societal or individual consequences.
- Operator must supply: The whole control: impact assessment process, its documentation and its use in decisions.

ISO/IEC 42001 A.6.2.4: AI system verification and validation

Evidence attached

Applies to: Voluntary, certifiable standard (organisation-level). Applies to an organisation that chooses to operate and certify an AI management system. Annex A controls are selected through the organisation's risk assessment and Statement of Applicability.

Requirement (paraphrase): The organisation defines and documents verification and validation measures for the AI system and specifies criteria for their use.

- Primary: Trial verdicts [present]. Verified outcomes in repeatable trials.
- Primary: K-rating certificate [present]. Acceptance criterion made explicit as a lower confidence bound with expiry.
- Primary: Shield certificate [absent]. Adversarial validation by attack class.
- Supporting: Verifiable judge provenance [absent]. Validation results are reproducible from stored traces.
- Limit: The criteria for use of the results are the organisation's.
- Operator must supply: Define the V&V plan and acceptance criteria for the intended use, and record decisions against them.

ISO/IEC 42001 A.6.2.5: AI system deployment

Partial evidence

Applies to: Voluntary, certifiable standard (organisation-level). Applies to an organisation that chooses to operate and certify an AI management system. Annex A controls are selected through the organisation's risk assessment and Statement of Applicability.

Requirement (paraphrase): The organisation documents a deployment plan and ensures that appropriate requirements are met prior to deployment.

- Supporting: K-rating certificate [present]. A certificate status that can be required before deployment.
- Supporting: Karatum Guard [present]. A gate can enforce the requirement at the moment of execution.
- Limit: Does not document or approve a deployment.
- Operator must supply: Deployment plan, approvals and the check of requirements before release.

ISO/IEC 42001 A.6.2.6: AI system operation and monitoring

Evidence attached

Applies to: Voluntary, certifiable standard (organisation-level). Applies to an organisation that chooses to operate and certify an AI management system. Annex A controls are selected through the organisation's risk assessment and Statement of Applicability.

Requirement (paraphrase): The organisation defines and documents the elements for ongoing operation and monitoring of the AI system, including performance monitoring, repairs, updates and support.

- Primary: Behavioural fingerprints [present]. Ongoing behavioural drift measurement.
- Primary: Flight Recorder [absent]. Recorded operation, with coverage measured in the pack.
- Supporting: Incident Registry [present]. Incident history.
- Limit: Monitoring of real use depends on what the operator records.
- Operator must supply: Operations plan: monitoring metrics, repair and update procedures, support.

ISO/IEC 42001 A.6.2.7: AI system technical documentation

Partial evidence

Applies to: Voluntary, certifiable standard (organisation-level). Applies to an organisation that chooses to operate and certify an AI management system. Annex A controls are selected through the organisation's risk assessment and Statement of Applicability.

Requirement (paraphrase): The organisation determines what AI system technical documentation each relevant category of interested parties needs and provides it in an appropriate form.

- Supporting: Trial verdicts [present]. This pack is itself a document for auditors on testing, monitoring and incidents.
- Supporting: Verifiable judge provenance [absent]. Reproducible evaluation method.
- Limit: Covers the evaluation slice of the documentation only.
- Operator must supply: System design, data, models, intended use and limits documentation for each audience.

ISO/IEC 42001 A.6.2.8: AI system recording of event logs

Evidence missing for this build

Applies to: Voluntary, certifiable standard (organisation-level). Applies to an organisation that chooses to operate and certify an AI management system. Annex A controls are selected through the organisation's risk assessment and Statement of Applicability.

Requirement (paraphrase): The organisation determines at which phases of the AI system life cycle record keeping of event logs should be enabled, at least when the system is in use.

- Primary: Flight Recorder [absent]. Tamper-evident log of agent steps, with batches, gaps and anchors counted.
- Supporting: Verifiable judge provenance [absent]. Evaluation traces are retained.
- Limit: Commitments only; contents stay with the operator.
- Operator must supply: Decide the phases and events to record, retention, protection and review of logs.

ISO/IEC 42001 A.8.2: System documentation and information for users

Partial evidence

Applies to: Voluntary, certifiable standard (organisation-level). Applies to an organisation that chooses to operate and certify an AI management system. Annex A controls are selected through the organisation's risk assessment and Statement of Applicability.

Requirement (paraphrase): The organisation determines and provides the necessary information to users of the AI system.

- Supporting: K-rating certificate [present]. A quotable measured performance statement with its limits.
- Supporting: Honesty Index [present]. Information on how far the agent's self-reports can be relied on.
- Limit: Not user documentation.
- Operator must supply: User-facing documentation: capabilities, limits, appropriate use.

ISO/IEC 42001 A.8.4: Communication of incidents

Evidence attached

Applies to: Voluntary, certifiable standard (organisation-level). Applies to an organisation that chooses to operate and certify an AI management system. Annex A controls are selected through the organisation's risk assessment and Statement of Applicability.

Requirement (paraphrase): The organisation determines and documents a plan for communicating incidents to users.

- Primary: Incident Registry [present]. Public timelines, a signed feed and webhooks for subscribers.
- Limit: Covers communication from Karatum's registry. User communication is the organisation's.
- Operator must supply: The organisation's own incident communication plan to its users.

ISO/IEC 42001 A.9.4: Intended use of the AI system

Partial evidence

Applies to: Voluntary, certifiable standard (organisation-level). Applies to an organisation that chooses to operate and certify an AI management system. Annex A controls are selected through the organisation's risk assessment and Statement of Applicability.

Requirement (paraphrase): The organisation ensures that the AI system is used according to the intended uses of the AI system and its accompanying documentation.

- Supporting: Karatum Guard [present]. A gate with tier caps limits what a certified agent can do on an account, a technical guard-rail against use beyond what was certified.
- Limit: Caps are about value and certification state, not about the purpose of an action.
- Operator must supply: Policies and controls that keep use within the intended purpose.

ISO/IEC 42001 A.10.3: Suppliers

Partial evidence

Applies to: Voluntary, certifiable standard (organisation-level). Applies to an organisation that chooses to operate and certify an AI management system. Annex A controls are selected through the organisation's risk assessment and Statement of Applicability.

Requirement (paraphrase): The organisation establishes a process to ensure that its use of services, products or materials provided by suppliers aligns with its approach to responsible development and use of AI systems.

- Supporting: K-rating certificate [present]. Independent evidence about a supplied agent.
- Supporting: Trial verdicts [present]. Verified outcomes, not supplier claims.
- Supporting: Shield certificate [absent]. Robustness evidence for the supplied agent.
- Limit: Evidence about one supplier's agent; not a supplier-management process.
- Operator must supply: Supplier selection and review process, agreements, ongoing supplier monitoring.

Gaps

Evidence Karatum does not have for this build

- [judge] No judge traces, provenance attestations or re-verification runs: verdicts cannot be re-derived from stored traces.
- [honesty] The build is in the Honesty Index snapshot but has no published score: 77 scored claims, 100 required.
- [shield] No Shield certificate: robustness against adversarial input has not been measured for this build.
- [flight] No Flight Recorder coverage: Karatum holds no record of the agent's own steps; no event log can be shown.
- [guard] No Guard account was declared, so the on-chain policy settings of the operator's account are not part of this pack.

Not declared by the operator

- EU AI Act classification not declared or not assessed (Art. 6, Annex III).
- Human oversight measures not declared (Art. 14, 26(2); NIST MAP 3.5).
- No stop or override procedure declared (Art. 14(4)(e); NIST MANAGE 2.4).
- Log retention period not declared (Art. 26(6); ISO/IEC 42001 A.6.2.8).
- Incident process not declared (Art. 73; NIST MANAGE 4.3; ISO/IEC 42001 A.8.4).

In addition, 60 operator obligations are listed per control in the control matrix; none of them can be evidenced by Karatum.

References and reproducibility

Upstream attestations this pack leans on (testnets only):

Kind	Network	UID / tx / hash
verdict_attestation	eip155:11155111	uid 0x4df9474dfaeca889aec0e2de3ec474f3ed3ab869b03a095e1d5e8e47b15556ca tx 0x9c4a68fde8872c2372b01e03e1aedd175c80d5821617bc1aa624ea2c5924cd78 hash 0x18c124d8484b424eaaacd85fca4b3eaaec47ca042e3eab4e1c7d231c6ce2c9d8 attempt hard-trial-2049:volatile2-06:build-a-careful-v1
verdict_attestation	eip155:11155111	uid 0xec8bb305e3c804b3f9316b2c3bdac07c040cc87a2ec3376d78665b53fa5e156c tx 0xbd539dbece91faeb0f32981262da3bf156a56912fed9b479e468a89846b74c92 hash 0xcac8f79e7dffffdbb3d315a54900b15c755ceeeaf6902bbc841bf8ab82464b17 attempt hard-trial-2049:volatile2-02:build-a-careful-v1
verdict_attestation	eip155:11155111	uid 0x8bb46b0776ca05dddb25fd458eeced3515420cf13b09070563f73ef13d11c74 tx 0x9c4a68fde8872c2372b01e03e1aedd175c80d5821617bc1aa624ea2c5924cd78 hash 0x2239d49d1879c9d110cb2e7b4f8a08af7e2afbaddbf78183d95d14d1a4f28d0 attempt hard-trial-2049:volatile2-05:build-a-careful-v1
verdict_attestation	eip155:11155111	uid 0x51cddd05440b2c1b403d5fdd9cfa0c8cf6c7dcaa9942e3c9cf916352f178bc4c tx 0xbd539dbece91faeb0f32981262da3bf156a56912fed9b479e468a89846b74c92 hash 0xa121b02bb7af24e4740bf6b2c3353bd04c80fb75599ae5457711040908adcc2 attempt hard-trial-2049:volatile2-03:build-a-careful-v1
verdict_attestation	eip155:11155111	uid 0xc71c4d79059aa3fc890cc12a5b6ccad57d6e4da09ed8a82f8a16a877538dd89b tx 0x9c4a68fde8872c2372b01e03e1aedd175c80d5821617bc1aa624ea2c5924cd78 hash 0x8befa5c74877beb7e70b19e3d60ea0dd932daf081727cdb830b81467d4927e31 attempt hard-trial-2049:volatile2-07:build-a-careful-v1
recall_guard_revoke	eip155:84532	tx 0xebc14d68d05a9a41f7979c603171296de1604ada5a5c82b9d20ae993f32328b8 KAR-2026-0003 (drill)
recall_guard_revoke	eip155:84532	tx 0x9102bb31d458f7b3c72fcb8fa4977a3830eb98e6c970051ded416370c3d7a65e KAR-2026-0003 (drill)
recall_guard_revoke	eip155:11155111	tx 0x5c962d8911e30492cbae7d9c1d685e73406ca63ceb280e58c4837d11bace86f0 KAR-2026-0003 (drill)
recall_guard_revoke	eip155:84532	tx 0x042e9d3f6e152562b3e81ef5e1b1aca93efdd244e23729bab67e3a951645cfff KAR-2026-0004 (drill)
recall_guard_revoke	eip155:11155111	tx 0xa6ecf2772a1195a1037510165e9660f539c139666447e729ab71d869df11c971 KAR-2026-0004 (drill)
recall_guard_revoke	eip155:84532	tx 0x087d127319fbb44c3303da65a4f5dc96179d9ecba4bc4f3e6d301ac33626fd4a KAR-2026-0005 (drill)
recall_guard_revoke	eip155:11155111	tx 0x476286811d43203eb45e59d16259cd0bc95822ecf2565c91ec56c2293cae7fbb KAR-2026-0005 (drill)
recall_eas_revoke	eip155:11155111	tx 0xa82cd9515d267ea3988217d4b63e961b604b35bd7e66ec6e407f8f5febe57db5 KAR-2026-0005 (drill)
recall_guard_revoke	eip155:84532	tx 0xba27d21d60d904cd3bd211bda8ec93abfd6715ceedff85ad7f63a5009f952736 KAR-2026-0005 (drill)
recall_guard_revoke	eip155:11155111	tx 0x0b89b956d2e0b26ca901f6470a46194734e657d59a012b8143bec2033cdcd96d KAR-2026-0005 (drill)

Kind	Network	UID / tx / hash
recall_erc8004_feedback	eip155:11155111	tx 0x7b87b970ce2c8a6baed3d61bf879bcb7ded90455c40c4c084101852e0f7c936f KAR-2026-0005 (drill)
recall_erc8004_feedback	eip155:11155111	tx 0x693ecf664732c90be5bcd4fb488545651362dfa7e2252a97f220f86882368ab KAR-2026-0005 (drill)
guard_root_publication	eip155:84532	tx 0x11cd8a1725532425808a0e7b4cd7bcf9067479bfe9a985bc8eaf6e024df779c epoch 9
guard_root_publication	eip155:11155111	tx 0xc3e691ae9c3d3f8f8e274717e00dd5af2a776945d08d2db9f33a96042480ff8f epoch 6

Evidence snapshot hash 0xf0f7e5e048d34cbbca414595250bd0ec07ef4cd6eee36ed72bf7e06758be3b25

Verdict set hash 0xa4435ab9cf4d89839c5d69977a18919035c5fe4ed34c6e432e5f1cddb7f31510

Control map hash 0xb1601aaa11cbea47c47b584a01d40ddcb68687a3ef9e689b79b458e3321b9a6a

Reproduce: the same Karatum database, the same as-of time and the same declaration give byte-identical content, PDF and manifest. Recompute keccak256 of the canonical JSON (sorted keys) of pack.json and compare it with the content hash above and with the Ethereum Sepolia attestation referenced in the manifest. See docs/methodology/compliance-pack.md.