

## Prepare your agent for assessment.

A practical evidence checklist for healthcare and financial deployments. For each control, identify the owner, the available evidence and the remaining gaps. Record only information you are authorised to share.

Agent, version and deployment scope

---

---

---

Completion of this checklist is not an assessment decision or a guarantee of insurance. An independent review and a defined scope are required.

---

# Governance and scope

Ownership, intended use, exclusions and a versioned inventory of the assessed agent.

---

G01 · 0.1

## ☐ **Named accountability**

Name an accountable owner, technical operator and independent decision reviewer; the implementer must not certify their own work.

Evidence to prepare

- Signed role matrix with decision rights.
- Conflict declarations and reviewer competence record.

Owner / evidence reference

---

---

G02 · 0.1

## ☐ **Intended use and exclusions**

State users, decisions, channels, data, actions and explicit exclusions in language that operations can enforce.

Evidence to prepare

- Approved use statement and user journey.
- Visible warnings, routing rules and blocked-use configuration.

Owner / evidence reference

---

---

G03 · 0.1

## ☐ **Versioned system inventory**

Bind the assessment to model, prompt, retrieval sources, tools, policies and deployment configuration with immutable versions.

Evidence to prepare

- Signed inventory and dependency bill.
- Hashes or release identifiers for assessed artefacts.

Owner / evidence reference

---

# Data and confidentiality

Data flows, minimisation, access, retention and protection against unintended disclosure.

---

D01 · 0.1

## ☐ Mapped and minimised data flows

Map each input, derived field, destination and purpose; remove fields not necessary for the declared task.

Evidence to prepare

- Data-flow diagram with declared locations.
- Field-level purpose and minimisation review.

Owner / evidence reference

---

---

D02 · 0.1

## ☐ Access and retention controls

Enforce role-based access, tenant separation and documented retention and deletion rules across prompts, traces, caches and backups.

Evidence to prepare

- Access matrix and tenant-boundary design.
- Retention schedule with deletion test records.

Owner / evidence reference

---

---

D03 · 0.1

## ☐ Disclosure resistance

Apply output filtering, retrieval authorization and context isolation so a requester receives only data they may access.

Evidence to prepare

- Retrieval authorization rules.
- Red-team cases for direct and indirect disclosure.

Owner / evidence reference

---

# Integration security

Untrusted input, tool permissions, secrets, dependencies and boundaries between tenants.

---

S01 · 0.1

## ☐ Untrusted input handling

Treat external content as data, preserve instruction priority and block content from granting itself authority.

Evidence to prepare

- Trust-boundary model and prompt policy.
- Injection test corpus with expected outcomes.

Owner / evidence reference

---

---

S02 · 0.1

## ☐ Least-privilege tools

Give each tool the narrowest scopes, records, actions and duration; keep tenant and environment credentials separate.

Evidence to prepare

- Tool catalogue with scopes and owners.
- Credential and tenant-isolation configuration.

Owner / evidence reference

---

---

S03 · 0.1

## ☐ Secrets, dependencies and execution boundaries

Keep secrets outside prompts and code, scan dependencies and isolate generated commands or code with explicit resource limits.

Evidence to prepare

- Secret-management and rotation records.
- Dependency inventory and sandbox policy.

Owner / evidence reference

---

# Reliability

Representative cases, unsupported output and regression evidence for the declared use.

---

R01 · 0.1

## ☐ Representative performance cases

Define a versioned test set covering normal, boundary and high-impact cases for the declared Swiss use, including supported languages.

Evidence to prepare

- Test-set rationale, size and composition.
- Per-case expected result and reviewed outcome.

Owner / evidence reference

---

---

R02 · 0.1

## ☐ Unsupported output handling

Require traceable sources where the task depends on facts and route low-support or conflicting answers to refusal or review.

Evidence to prepare

- Grounding policy and source-quality rules.
- Cases containing absent, stale and conflicting evidence.

Owner / evidence reference

---

---

R03 · 0.1

## ☐ Regression control

Run risk-based regression tests before release and define which changes trigger targeted or full reassessment.

Evidence to prepare

- Release gate and regression history.
- Change-to-test impact matrix.

Owner / evidence reference

---

# Human oversight

Approval, escalation and stop mechanisms for consequential actions.

---

H01 · 0.1

## ☐ Approval of consequential actions

Bind a named human approval to the exact action, target and current payload before bookings, external sends, deletions and clinical case progression.

Evidence to prepare

- Action inventory and approval policy.
- Anonymised traces for approved, denied and expired requests.

Owner / evidence reference

---

---

H02 · 0.1

## ☐ Escalation to a competent person

Define observable triggers, destination, context package and response ownership for escalation.

Evidence to prepare

- Escalation matrix and service ownership.
- Sample escalations with context and resolution.

Owner / evidence reference

---

---

H03 · 0.1

## ☐ Stop and containment

Provide tested controls to disable actions, revoke credentials, isolate a tenant and preserve evidence without relying on the agent.

Evidence to prepare

- Kill-switch and containment runbook.
- Exercise record with roles and elapsed steps.

Owner / evidence reference

---

# Transparency and impact

AI disclosure, understandable decisions and impacts on affected people.

---

T01 · 0.1

## ☐ Disclosure of AI interaction

Identify the AI interaction at the relevant moment and explain the purpose, main data uses and route to a person.

Evidence to prepare

- Interface copy for every supported channel.
- Disclosure timing and accessibility review.

Owner / evidence reference

---

---

T02 · 0.1

## ☐ Understandable outcomes

Present the material basis, uncertainty, next step and review route at the level needed by the affected user or operator.

Evidence to prepare

- Outcome templates and explanation standard.
- Reviewed examples for different audiences and languages.

Owner / evidence reference

---

---

T03 · 0.1

## ☐ Impact assessment and affected groups

Identify affected groups, credible harms, severity, likelihood and mitigations before high-impact use, then revisit material changes.

Evidence to prepare

- Impact assessment with consultation inputs.
- Risk owners, mitigations and acceptance decisions.

Owner / evidence reference

---

# Operations

Tamper-evident traces, incident response and controlled change after release.

---

O01 · 0.1

## ☐ Tamper-evident event trails

Record identity, version, inputs references, policy decisions, tool calls, approvals and outcomes in append-only or alteration-evident storage.

Evidence to prepare

- Event schema, integrity design and access rules.
- Sample trace linked across agent, approval and tool systems.

Owner / evidence reference

---

---

O02 · 0.1

## ☐ Incident response for agent failures

Define agent-specific severity, containment, evidence preservation, notification assessment and learning responsibilities.

Evidence to prepare

- Incident playbook and contact tree.
- Exercise or real incident record with follow-up actions.

Owner / evidence reference

---

---

O03 · 0.1

## ☐ Change and quarterly review

Review evidence at least quarterly under this draft scheme and reassess immediately after defined material changes or incidents.

Evidence to prepare

- Review calendar and material-change criteria.
- Quarterly records, release history and open exceptions.

Owner / evidence reference

---

# Evidence

Provenance, reproducibility, assessment boundaries and residual risk.

---

E01 · 0.1

## ☐ Evidence provenance and integrity

Label each artefact with origin, collector, timestamp, environment, version and integrity reference, while redacting unnecessary data.

Evidence to prepare

- Evidence register with chain of custody.
- Original artefacts or hashes in access-controlled storage.

Owner / evidence reference

---

---

E02 · 0.1

## ☐ Reproducible verification

Record preconditions, inputs, procedure, expected result, observed result and evaluator so another qualified reviewer can repeat the test.

Evidence to prepare

- Executable test cases or precise protocols.
- Environment manifest and retained result files.

Owner / evidence reference

---

---

E03 · 0.1

## ☐ Scope and residual-risk statement

Publish the assessed version, environment, use, exclusions, failed or untested areas, residual risks and review triggers beside the result.

Evidence to prepare

- Signed scope and residual-risk section.
- Consistency check across report, register and public claims.

Owner / evidence reference

---