

READYFORROLE

Enterprise AI Agent Readiness Checklist Pack

Seven one-page design-review checklists, one for each part of the Enterprise AI Agent Engineering series: context and memory, tools and harness, loops and orchestration, guardrails and permissions, human-in-the-loop design, evals, and observability and tracing.

Read the full guide at readyforrole.com/blog/enterprise-ai-agent-engineering.html

Context and Memory Design Review

Use this checklist in your next design review. Tick each statement that is true for your agent today. Thresholds and intervals in the guide are recommended starting points; calibrate to your environment.

CONTEXT SOURCES

- ☐ Every context source the agent can load is named and has an owner.
- ☐ Retrieval has caps and ranking, so relevant items are not buried mid-context.
- ☐ Non-negotiable rules are placed at the start of the context.
- ☐ Sensitive data is loaded only when a task needs it.

MEMORY DESIGN

- ☐ Memory writes are gated by a rule or a human, and logged.
- ☐ Every memory entry carries provenance and an expiry or review date.
- ☐ Memory access is scoped by user and role.
- ☐ Prior decisions are stored as structured records that can be retrieved by identifier.

TESTING

- ☐ A cross-user test set confirms one user's memory never appears in another's answers.
- ☐ A stale-memory test confirms superseded policies and decisions are not cited.
- ☐ The assembled context is recorded for each step.

GOVERNANCE

- ☐ The design-review table has been scored and the result recorded with a go or no-go decision.
- ☐ A named owner can review and delete wrong or disputed memory entries.
- ☐ Any agent scoring 0 on memory scoping that touches personal or regulated data is stopped.

Score: ____ of 14. Any unticked box is a gap to close before the agent handles real work.

Full guide: readyforrole.com/blog/context-engineering-memory-architecture.html | Series hub: readyforrole.com/blog/enterprise-ai-agent-engineering.html | support@readyforrole.com

Tool and Harness Design Review

Use this checklist in your next design review. Tick each statement that is true for your agent today. Thresholds and intervals in the guide are recommended starting points; calibrate to your environment.

INVENTORY AND TIERS

- ☐ Every tool is listed and assigned a risk tier (read-only, reversible write, irreversible or external write).
- ☐ Any tool whose tier is unclear is treated as Tier 3 until it is classified.
- ☐ Read and write actions are separate tools.
- ☐ Tools expose only the actions the agent needs, not a whole API.

TOOL CONTRACTS

- ☐ Each tool description states when to use it and when not to use it.
- ☐ Input schemas are strict, with enumerated values instead of free text where possible.
- ☐ Long outputs are summarised by the tool and the full output is stored elsewhere.

HARNESS CONTROLS

- ☐ Tier 2 and Tier 3 tools use idempotency keys.
- ☐ The harness, not the model, controls retries for state-changing actions.
- ☐ Timeouts and a maximum number of calls per run are set.
- ☐ Permission scope and approval rules are checked before every execution.

OPERATIONS

- ☐ Success rate and error types are measured for each tool.
- ☐ A repeated-call test confirms no duplicate side effects.
- ☐ Every Tier 3 tool has a signed-off go or no-go decision.

Score: ____ of 14. Any unticked box is a gap to close before the agent handles real work.

Full guide: readyforrole.com/blog/tool-design-harness-engineering.html | Series hub:
readyforrole.com/blog/enterprise-ai-agent-engineering.html | support@readyforrole.com

Loop and Orchestration Design Review

Use this checklist in your next design review. Tick each statement that is true for your agent today. Thresholds and intervals in the guide are recommended starting points; calibrate to your environment.

PATTERN SELECTION

- ☐ Every agent task is mapped to a pattern: fixed workflow, workflow with a bounded loop, single-agent loop, planner and executor, or supervisor with specialists.
- ☐ The simplest pattern that fits was chosen, and each move to a more complex pattern is justified by a measured failure.
- ☐ Any multi-agent design is justified in writing, naming the problem the extra agent solves.
- ☐ Long tasks have a plan that is reviewed before execution.

LIMITS

- ☐ A maximum number of steps is enforced by the harness.
- ☐ A time limit is enforced by the harness.
- ☐ A cost limit is enforced by the harness.
- ☐ Alerts fire when a run approaches or exceeds its limits.

STOP AND HANDOFF

- ☐ Repeated actions are detected, with a stop after two or three repeats as a starting point.
- ☐ A defined handoff to a person happens when limits are reached or confidence is low.
- ☐ A deliberately stuck agent has been tested and stops with a useful summary.

COMPLETION AND HANDOFFS

- ☐ Completion is verified against a checklist or independent check, not only the agent's own claim.
- ☐ Handoffs between agents use structured records with required fields.
- ☐ Outputs are validated at each handoff so one agent's error does not cascade.

Score: ____ of 14. Any unticked box is a gap to close before the agent handles real work.

Full guide: readyforrole.com/blog/loop-engineering-orchestration-patterns.html | Series hub: readyforrole.com/blog/enterprise-ai-agent-engineering.html | support@readyforrole.com

Agent Guardrails and Permissions Review

Use this checklist in your next design review. Tick each statement that is true for your agent today. Thresholds and intervals in the guide are recommended starting points; calibrate to your environment.

IDENTITY AND PERMISSIONS

- ☐ Every agent runs under its own or a delegated identity, with no shared or administrator credentials.
- ☐ Each task has a written permission scope listing allowed actions and data.
- ☐ Elevated access is time-bound and expires by default.
- ☐ Delegated identity is used where the agent must not exceed the requesting user's rights.

POLICY ENFORCEMENT

- ☐ Every hard rule exists as a check in the harness, not only in the prompt.
- ☐ Each action is classified by risk class, with minimum controls applied.
- ☐ Irreversible writes and access changes route to a human checkpoint.
- ☐ Output checks cover personal data and unapproved commitments.

INJECTION AND MISUSE

- ☐ All content from documents, tickets, emails and tool results is treated as untrusted data.
- ☐ An injection test set exists and was run before pilot, with zero unauthorised actions.
- ☐ Guardrails are tested with varied phrasing, not keywords alone.

AUDIT

- ☐ All allow, deny and flag decisions are recorded in the trace.
- ☐ Denials and false positives are reviewed on a schedule.
- ☐ Permission scopes are reviewed on a fixed schedule (quarterly as a starting point).

Score: ____ of 14. Any unticked box is a gap to close before the agent handles real work.

Full guide: readyforrole.com/blog/guardrails-and-permissions.html | Series hub:
readyforrole.com/blog/enterprise-ai-agent-engineering.html | support@readyforrole.com

Human-in-the-Loop Design Review

Use this checklist in your next design review. Tick each statement that is true for your agent today. Thresholds and intervals in the guide are recommended starting points; calibrate to your environment.

PLACEMENT

- ☐ Every agent action is classified, and each class has a defined oversight level.
- ☐ Irreversible, money-moving, access-changing and regulated actions require approval before action.
- ☐ Low-risk, reversible actions use sampled review instead of per-item approval.
- ☐ Triggers are fixed rules where possible, and confidence-based triggers are calibrated.

REVIEW PACKET

- ☐ The packet shows the proposed action, the evidence, the reason for the flag and the options.
- ☐ Reviewers can approve, edit, reject with a reason, escalate or take over.
- ☐ Rejection reasons are structured and recorded.

OWNERSHIP AND TIMING

- ☐ Each checkpoint has a named owner role with the authority to decide.
- ☐ Each checkpoint has a response target and a safe default on timeout.
- ☐ Regulated or high-value actions need two approvers, one independent.

OPERATION AND LEARNING

- ☐ Run state is saved at the pause and verified on resume.
- ☐ Decision time, override rate and agreement with later outcomes are measured.
- ☐ Rejections and overrides feed the evaluation set.

Score: ____ of 13. Any unticked box is a gap to close before the agent handles real work.

Full guide: readyforrole.com/blog/human-in-the-loop-design.html | Series hub:
readyforrole.com/blog/enterprise-ai-agent-engineering.html | support@readyforrole.com

Agent Evals Readiness Review

Use this checklist in your next design review. Tick each statement that is true for your agent today. Thresholds and intervals in the guide are recommended starting points; calibrate to your environment.

EVAL SET

- ☐ Built from real cases and covering every category the agent handles.
- ☐ Includes hard cases, past incidents and rejected checkpoints.
- ☐ Includes safety, permission and injection cases.
- ☐ Part of the set is held out and never used for tuning.

GRADERS

- ☐ Code checks are used for every item with an exact answer.
- ☐ Human review covers quality and edge cases on a sample.
- ☐ Any model-based judge has been compared with human ratings on your own data.

RUNNING EVALS

- ☐ Tools are mocked or sandboxed so runs are repeatable and safe.
- ☐ Each case is run several times and consistency is reported.
- ☐ Traces are stored so tool calls, steps, cost and handoffs can be graded.

RELEASE AND LEARNING

- ☐ Written release gates exist for pilot, limited rollout and full rollout.
- ☐ Every change to the prompt, model, tools or data triggers an eval run.
- ☐ Production runs are sampled and scored, and incidents become new cases.

Score: ____ of 13. Any unticked box is a gap to close before the agent handles real work.

Full guide: readyforrole.com/blog/evals-for-agents.html | Series hub: readyforrole.com/blog/enterprise-ai-agent-engineering.html | support@readyforrole.com

Agent Observability Readiness Review

Use this checklist in your next design review. Tick each statement that is true for your agent today. Thresholds and intervals in the guide are recommended starting points; calibrate to your environment.

TRACE DESIGN

- ☐ Every run has one identifier that follows it through every step, tool and agent.
- ☐ Every span records the prompt, model and tool versions.
- ☐ The assembled context is recorded for each step.
- ☐ Policy decisions and human approvals are recorded for every action that changes data.

SIGNALS AND ALERTS

- ☐ Quality, reliability, cost and speed, safety, and human oversight signals are all tracked.
- ☐ Every alert has a named owner and a defined action.
- ☐ Online eval scores are tracked next to operational metrics.

DATA PROTECTION

- ☐ Personal data in traces is redacted or stored as a reference.
- ☐ Access to traces is role-based, and who viewed a trace is logged.
- ☐ Retention periods are agreed with legal and security for each risk level.

REVIEW AND IMPROVEMENT

- ☐ Sampled and flagged traces are reviewed on a regular schedule.
- ☐ Every finding becomes a fix or a new eval case.
- ☐ Changes to prompts, models and tools are logged and linked to traces.

Score: ____ of 13. Any unticked box is a gap to close before the agent handles real work.

Full guide: readyforrole.com/blog/observability-and-tracing.html | Series hub:
readyforrole.com/blog/enterprise-ai-agent-engineering.html | support@readyforrole.com