

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