Agent Readiness Checklist

This is a pre-handoff gate, run once before you let an agent act on a task on its own, not a review of work it already did. An agent is a model that has been given tools and a goal and is allowed to loop on its own: read, decide, act, observe, repeat. The checklist forces a yes or no on one question. Is this specific task safe to hand off, and if so, with which guardrails and at which level of autonomy.

Scope the task

- [ ] The goal is written down, with a clear stopping condition the agent can recognize as "done".
- [ ] The task is narrow enough to describe in one sentence; it is not "handle everything in the inbox".
- [ ] You have named the worst single action the agent could take, and exactly what it would touch.
- [ ] You know how reversible the worst case is and roughly how long an undo would take.

Guardrails before the first run

- [ ] Every irreversible action (delete, send, charge, deploy) is either removed or behind a human approval.
- [ ] The agent holds only the tools and scopes this task needs, on a scoped credential, not a shared admin token.
- [ ] You know what data flows in, where outputs go, and that nothing crosses a boundary it should not.
- [ ] Each human approval point challenges intent, data lineage, and blast radius; it cannot be rubber-stamped.
- [ ] Every tool call, input, and output is logged from the first run, in a place a person can read later.
- [ ] A named human can roll back the agent's work using a path you have actually tested.

Set the autonomy tier

- [ ] You have chosen a tier on purpose: suggest only, act with approval, or act and report.
- [ ] You can state in one line why the task earns that tier today, not someday.
- [ ] You know the main ways the task fails (loops, wrong tool, prompt injection) and who gets paged.
- [ ] You have a trigger for widening autonomy later (for example, a count of clean logged runs).