REDHUB.AI · FREE CHECKLIST

The AI Inspectability Checklist

Ten questions to ask any vendor or team that says its AI is explainable. Send them ahead of the call and write the answers in the right-hand column, in the vendor's own words.

01	**What can you see inside the model you run?** Inspection inside a model needs its weights. A product built on a hosted API sees outputs and nothing else. **Evidence:** A plain answer naming the level: hosted API, outputs and logs only; or open weights, activations readable. *Not evidence: "Our platform is fully transparent."*	*Their answer*
02	**When your system explains a decision, what is it describing?** Stated reasoning can be faithful to the computation or produced after the fact, and it reads the same either way. **Evidence:** The retrieved sources and the rule that fired, or the model's own reasoning treated as a claim. *Not evidence: "The model explains its reasoning."*	*Their answer*
03	**Has the explanation ever been tested against behavior?** A description becomes evidence when it predicts what the system does under a change. **Evidence:** The inputs the explanation names were varied, the output moved with them, and the test set is shown. *Not evidence: "Users find the explanations helpful."*	*Their answer*
04	**What is your first-line check, and does it run on every output?** Oversight that scales starts with a cheap signal that sees everything. **Evidence:** A named check with a cost per run: schema, policy-word filter, citation-present check, or a probe. *Not evidence: "We monitor quality."*	*Their answer*
05	**What escalates, to whom, and how fast?** A first-line flag with no owner is a log entry. **Evidence:** A named grader and reviewer, a time for each, and last month's escalation count. *Not evidence: "Issues are escalated appropriately."*	*Their answer*
06	**Which actions cannot be undone, and who approves them?** Model quality does not change the list of irreversible actions. A person owns those regardless of score. **Evidence:** The list itself, with the approver for each. *Not evidence: "The model is highly accurate on those."*	*Their answer*
07	**Can the system reach the thing that grades it?** An agent that can edit its tests or close its own ticket has been handed the answer key. **Evidence:** The grader lives where the agent has no write access, and the permission map shows it. *Not evidence: "The agent is instructed not to modify tests."*	*Their answer*
08	**What record exists after a decision?** An explanation that cannot be checked later is not one anyone can act on. **Evidence:** Four things joined: the output, the inputs it saw, the rule in force that day, and who approved it. *Not evidence: "Everything is logged."*	*Their answer*
09	**What changes without telling us?** A hosted model can be retired, repointed or retrained, and every check was tuned on the old one. **Evidence:** The model and version pinned, a notice period in the contract, and the right to re-run your evaluation. *Not evidence: "We always use the latest model."*	*Their answer*
10	**Who can stop it, and have they practiced?** A stop that has never been pulled is a hope. **Evidence:** A named person with the authority, the mechanism, and the date of the last drill. *Not evidence: "There is a kill switch."*	*Their answer*

This is a buyer and operator checklist, not an interpretability certification. It does not rate any vendor, and a good answer to every question does not make a system safe. RedHub does not read model internals; its instruments grade the decisions around a model.

The full guide: blog.redhub.ai/ai-inspectability-checklist