

Phase 4 Suggestions (Long-Term Goals: Continuous Improvement to Best Practices)

The following represent additional elements to move agents to the ideal state.

1. Link to an internally performed or externally contracted study, using quasi-experimental or experimental research evidence, that provides evidence for the agent's effectiveness on the measurable outcomes you are seeking to improve. If results are not favorable, considering retiring the agent:
2. Regularly collect and summarize the key feedback received, improvements made (or backlogged for future versions), and any outstanding major issues that were not addressed or only partially mitigated:
3. Regularly update these statistics using human judgment combined with LLM-as-a-judge approaches:
 - a. **Correctness:** %. The share of responses that are factually accurate and do not introduce errors.
 - b. **Groundedness:** %. The share of responses that are supported by the allowed sources or context, cite sources, and do not invent unsupported claims.
 - c. **Explainability and clarity:** %. The share of responses that provide evidence for the agent's output that is meaningful to the intended user, accurately reflects the system's process for generating the answer, and clearly communicates the agent's knowledge limits.
 - d. **Completeness:** %. The share of responses that correctly cover all necessary elements of the request—no key steps, constraints, or required details are omitted.
 - e. **Reliability under variation:** **percentage points**. The difference in correctness across reasonable variations in phrasing, order, multi-turn context, or scenario paths.
 - f. **Cost per completed agentic task:** \$. The total cost for the agent to complete all necessary actions for the given task.
 - g. **Usage:** # . The total number of completed tasks.

Link to the test set document:

4. Expand the number of subgroups analyzed for fairness testing, based on the potential for plausible harm for the agent's intended use. Regularly update these statistics using human judgment combined with LLM-as-a-judge approaches, appropriately validated and regularly reviewed:
 - a. **Outcome gap:** **percentage points ([highest subgroup]: [yy]%, [lowest subgroup]: [zz]%)**. Report the difference between subgroups who were able to complete the task they came to the agent for: they found the needed information, resolved their issue, submitted their information, etc. Measured as the subgroup with the highest rate minus the subgroup with the lowest rate.
 - b. **Decision gap:** **percentage points ([highest subgroup]: [yy]%, [lowest subgroup]: [zz]%)**. If the agent helps a human make, or itself makes, a consequential decision, report the difference between subgroups provided with materially different decisions. Measured as the subgroup with the highest rate minus the subgroup with the lowest rate.
 - c. **Bias gap:** **percentage points ([highest subgroup]: [yy]%, [lowest subgroup]: [zz]%)**. Ask each subgroup tester whether the agent perpetuated stereotypes, prejudices, or assumptions about

their identity or background. Report the share of participants by subgroup and at least one biased or stereotyped response. Measured as the subgroup with the highest rate minus the subgroup with the lowest rate.

Link to the document(s) containing scenarios, subgroups, and reviewed answers:

5. Describe the key points in the agent process where human intervention is needed, if any. If human intervention is necessary:
 - a. Describe how the appropriate staff members have been identified and trained, and how they will continue to be identified and upskilled as model changes are made:
 - b. Describe how you plan to monitor and measure the rate of “rubber-stamping” (insufficient human review):
 - c. Describe how users can challenge the decisions or information in the system, as appropriate:
6. List the accountable owner name and title. Consider naming an evaluation lead, transparency lead, and responsible agentic AI lead as capacity and need permits:
7. Regularly review and update the vetted agentic cybersecurity environment:
8. Check the boxes below to affirm the following:
 - ☐ I have made this document, apart from security information deemed sensitive by myself and the Security Lead available to all users of the agent.
 - ☐ I have a process for regularly releasing updates of this document when changes are made.