Phase 2 Requirements (Production)

1. The accountable owner should ensure that the following is measured and reported to users of the tool at least once per year. Human judgment combined with an LLM-as-a-judge approach may be used at this stage if deemed appropriate, only if human and LLM-as-a-judge metrics have been appropriately tested and validated:

   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.

   h. **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.

   i. **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.

   j. **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 subgroups, inputs, and reviewed outputs:

2. Describe the baseline automated agentic cybersecurity tests run, vulnerabilities found, and mitigations implemented. If not relevant for this agent, say so:

    a. Unauthorized actions/excessive agency (if the agent can be manipulated into doing things it shouldn't have permissions to do):

    b. Sensitive information disclosure (if the agent contains any private or sensitive information and can be manipulated to expose that information):

    c. Knowledge source poisoning/indirect injection (if the agent uses external sources and treats those external sources as prompt instructions):

    d. Red teaming (have the internal team try to creatively elude safeguards):

As the accountable owner, I affirm the above to be true and accurate as of

Accountable owner signature: _____

As the security lead, I affirm the above to be true and accurate as of

Security lead signature: _____