Pre-Phase 2 Requirements (External Users Pilot)

1. Based on the results of phase 1, describe how the agent has met the needs of the key populations targeted:

2. List the names and e-mails of external testers and any additional internal testers that need access to the agent at this stage:

3. Define the theory of action and logic model of the AI agent. Be sure to include major key assumptions (if x and y are true about the tool, and we have z in place, we can achieve the measurable outcomes mentioned above):

4. Briefly describe the user testing plan, how that plan effectively measures the measurable outcomes the agent plans to achieve, the organizational and user risks that the tool presents, and planned mitigations for those risks:

5. Describe the key points in the agent process where human intervention is needed, if any. If human intervention is necessary, describe how the appropriate staff members have been identified and trained:

6. If any items from the phase 1 documentation have changed, including performance targets, describe those changes here:

7. The accountable owner should set performance targets according to the following fairness metrics. These should represent the minimum values necessary for continued external use on an ongoing basis.

    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.

    To calculate these statistics, you'll create two or three structured scenarios based on the AI agent's purpose and intended use. Choose one to three sets of relevant subgroups, based on the potential for plausible harm from using the agent. Ensure that a few members of each subgroup test the AI agent and collect all responses. 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.

8. Describe the baseline agentic cybersecurity tests run, vulnerabilities found, and mitigations implemented in the following areas. 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):

9. Name and provide the title of the cybersecurity official that conducted or oversaw the agentic cybersecurity tests and made any necessary mitigations:

10. Check the box below to affirm the following:

    - ☐ I have attached the Phase 1 documentation to this document.
    - ☐ I will make this document, with the exception of any sensitive security information, available to all users of the agent ahead of testing.
    - ☐ I have notified pilot testers and they understand that this is a limited term pilot and testing is done at their own risk.

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: _____

As the responsible agentic AI lead, I and the relevant internal governance committee approve the release of this agent for no more than six months as of

Responsible agentic AI lead signature: _____