

Phase 1 Requirements

The accountable owner can either progress to user testing and regularly have a human expert review answers to measure the following statistics or can have a human create test questions and acceptable answers. In either case, an expert human's judgment—not an AI—should be used to measure the following statistics and provide the results below:

- 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 or document containing inputs and reviewed outputs:

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

Accountable owner signature: _____