Choose a method for evaluating agents.

When creating a test suite, choose from different testing methods to evaluate agent responses. Each testing method has its own advantages and is suitable for different types of assessments.

When creating a test suite, choose from different testing methods to evaluate agent responses. Each testing method has its own advantages and is suitable for different types of assessments.

Testing method	Measurement	Type of test kit	Scoring	Configuration
General quality	How good is the response to a test case based on specific characteristics?	Single response or conversation	Score on a 100% scale	Are not
Connect oven	The degree of agreement between the meaning of the answer in the test case and the expected answer.	Single response	Score on a 100% scale	Passing score, expected answer
Capability use	Does the test case utilize all or any of the resources expected?	Single response	Pass/Fail	Predictability
Keyword match	Does the test case use all or any of the expected keywords or phrases?	Single response or conversation	Pass/Fail	Expected keywords or phrases
Text similarity	The degree of agreement between the content of the test answer and the expected answer.	Single response	Score on a 100% scale	Passing score, expected answer
Exact match	Do the results of the test case exactly match the expected results?	Single response	Pass/Fail	Expected answer
Custom	Does the response from the test case meet your defined criteria or expectations?	Single response or conversation	Pass/Fail (meets the defined label criteria)	Name, assessment guidelines, label

Add a test method

1. When creating or editing a test kit, select **Add test method** .

2. Select all the methods you want to test, then select **OK** . You can add more methods.

a. Some methods require passing scores. Pass scores determine which points lead to passing or failing. Set a score, then select **OK** .

b. Some testing methods require more criteria.

3. Select **Save** to save your changes to the test suite.

Choose an existing test method to modify its criteria or delete the method.

General quality

Available for both single-response and conversational testing suites, **General Quality** helps you determine if an agent's response meets your standards. It uses a large language model (LLM) to assess how effectively an agent responds to user questions.

General quality is particularly useful when no single, precise answer is expected. It provides a flexible and scalable way to evaluate feedback based on retrieved documents and conversational flows.

It uses these key criteria and applies a consistent prompt to guide the scoring process:

1. **Relevance** : To what extent does the agent's response address the question? For example, does the agent's answer stay on topic and directly answer the question?
2. **Contextuality** : To what extent is the agent's response based on the context provided? For example, does the agent's response reference or rely on information provided in the context, rather than providing irrelevant or unsupported information?
3. **Completeness** : To what extent does the agent's answer provide all the necessary information? For example, does the agent's answer cover all aspects of the question and provide sufficient detail?
4. **Effort:** Does the agent make an effort to answer the question?

To be considered high-quality, an answer must meet all of these key criteria. If a criterion is not met, the answer will be marked for improvement. This scoring method ensures that only answers that are both complete and well-supported receive the highest score. Conversely, answers that are incomplete or lack supporting evidence will receive lower scores.

When adding or editing test methods, select **General quality** . All test suites start with this method by default.

You don't need to add expected answers to test cases to complete the overall quality assessment.

Note : Reducing the number of knowledge sources for the agent does not guarantee an improvement in the overall quality score during agent evaluation. This limitation exists because the amount of knowledge retrieved (knowledge that the model deems relevant to a particular test case) may be too large.

Connect oven

Available for single-response test suites , **Compare Meaning** assesses how well the agent's response reflects the intended meaning of the expected response. Instead of focusing on exact wording, it uses similarity in intent, meaning it compares the ideas and meanings behind the words, to assess how well the response matches what you expect.

Similar to overall quality, Compare meaning is particularly useful when no single correct answer is expected. It provides a flexible and scalable way to evaluate responses based on retrieved documents and conversational flow.

You can set a threshold to determine the passing score for an answer. The default passing score is 50. The Compare Meaning testing method is useful when an answer can be expressed in many different correct ways, but the overall meaning or intent still needs to be conveyed.

1. When adding or editing test methods, select **Compare meaning** .

2. Set a target score for this method.

3. Add the expected response. Any test case that lacks an expected response will result in an invalid outcome for this testing method.

1. Choose a test case.
2. Add the answers you expect.
3. Select **Apply** to save your expected answer.
4. Repeat this process for all test cases you want to test using this method.

Tool use

Available for single-response testing toolkits, **Capability uses** to see if the agent uses specific tools or topics to generate a response. If it does, it passes. If not, it fails.

1. When adding or editing test methods, select **Tool use** .

2. Add the expected tools or themes. Any test cases that do not have the expected answers will produce an Invalid result for this testing method.

1. Select a test case. To add the same expected tool and theme to all test cases, select the **Edit icon in the Tool use** column header .
2. In the **Edit test case** pane , select the tools you expect your agent to use for that test case.
3. Select **OK** .
4. Select **Apply** to save the changes.
5. Repeat this process for all test cases where you want to verify the use of the tool.

Keyword match

Available for single-response and conversational test suites, **Keyword Match** checks whether the agent's response contains some or all of the words or phrases from the expected response that you define. If it does, it passes. If not, it fails.

You can choose whether a test requires any keywords or all keywords. Selecting **"Any"** means that if at least one word or phrase matches, the test case passes. Selecting **"All"** means that all expected words or phrases must match for the test case to pass.

1. When adding or editing test methods, select **Keyword match** .

2. Choose whether a test case requires any or all keywords to match.

3. Add the expected keywords. Any test cases that lack the expected keywords will produce invalid results for this testing method.

1. Choose a test case.
2. In the **Edit test case** pane , add a keyword or phrase that you expect the answer to that case to include.
3. Select **+ Add** to add more keywords or phrases. To delete a keyword or phrase, select the **Delete** icon .
4. Select **Apply** to save the proposed keywords.
5. Repeat this process for all test cases where you want to check for keyword matching.

Text similarity

Available for single-response test suites, the Text similarity test compares the similarity of responses from the agent to expected responses that you define in your test suite. This method is useful when the response can be expressed in many different correct ways, but the overall meaning or purpose still needs to be conveyed.

It uses a cosine similarity index to assess the degree of similarity between the agent's response and the wording and meaning of the expected response, and to determine a score. The score ranges from 0 to 1, where 1 indicates a close match and 0 indicates a no match. You can set a passing score threshold to determine the passing score for a response.

1. When adding or editing test methods, select **Text similarity** .

2. Set a target score for this method.

3. Add expected responses. Any test cases that lack expected responses will result in an invalid outcome for this testing method.

1. Choose a test case.
2. Add the answers you expect.
3. Select **Apply** to save your expected answer.
4. Repeat this process for all test cases you want to test using this method.

Exact match

Available for single-response test suites, **Exact Match** checks whether the agent's response exactly matches the expected response in the test: character by character, word by word. If they match, the test passes. If there are any discrepancies, the test fails. Exact match testing is useful for short, precise responses such as numbers, codes, or fixed phrases. It is not suitable for responses that people can express in multiple correct ways.

1. When adding or editing test methods, select **Exact match** .

2. Add expected answers. Any test cases without expected answers will produce invalid results for this testing method.

1. Choose a test case.
2. Add the answers you expect.
3. Select **Apply** to save your proposed answer.
4. Repeat this process for all test cases you want to test using this method.

Custom

Custom is a customizable testing methodology. It allows you to test and label agent responses using your own criteria. For example, you could create a compliance test for an HR agent to label test responses as compliant or non-compliant with your HR problem description.

A custom test has two components that you can configure:

Assessment instructions : Describe the goals you want to achieve with this test. What do you want the test to learn about the agent's responses?

Good review guidelines require:

1. Aim for a goal.
2. Use only the characters that are allowed.
3. Use bullet points and headings to organize.

For example:

```
?ánh giá câu tr? l?i c?a agent v? vi?c tuân th? chính sách nhân s?. Nh?ng ?
i?u c?n ki?m tra: - Xác ??nh xem câu tr? l?i có b?o v? quy?n riêng t?
 và tránh ti?t l? ho?c yêu c?u d? li?u nh?y c?m hay không. - Tránh phân bi?
t ??i x?, thiên v? ho?c phán xét không phù h?p. - Cung c?p h??ng d?
n an toàn, trung l?p và phù h?p v?i chính sách nhân s?. - Không ??a ra l?
i khuyên pháp lý ho?c ??a ra tuyên b? d?t khoát.
```

Labels : Describe the results assigned to each answer using the custom test. Labels also include pass/fail designations, which are factored into the pass rate of the test set for this testing method.

Labels have a name and a description. A good description should:

1. Brief.
2. The relevant answers contain the attributes you are looking for.

One strategy for labeling is to use two labels: one for answers that successfully meet the criteria you're looking for, and another for answers that don't. For example, a custom HR policy compliance test might have two labels: *Compliant* and *Non-Compliant* .

1. When adding or editing test methods, select **Custom** .

2. Enter a name for this custom test.

3. Add evaluation guidelines.

4. Add two or more labels. Each label should have a name and a description.

To add more labels, select **Add label** .

Label titles can only use letters, numbers, spaces, hyphens -, underscores _, forward slashes /, ampersands &, plus signs +, and periods .

5. Assign a Pass or **Fail** result to each label.

6. Select **OK** .