Overview of agent evaluation

As AI agents take on a critical role in business processes, the need for reliable and repeatable testing becomes essential. Agent evaluation allows you to create tests that simulate real-world scenarios for your agents.

As AI agents take on a critical role in business processes, the need for reliable and repeatable testing becomes essential. Agent evaluation allows you to create tests that simulate real-world scenarios for your agents.

These tests involve multiple questions and conversations more quickly than manual, case-by-case testing. You can then measure the accuracy, relevance, and quality of responses from agent interactions, based on the information the agent has access to. By using the results from the testing tool, you can optimize agent behavior and verify that the agent meets your quality and business requirements.

Why should we use automated testing?

Agent evaluation provides automated, structured testing. It helps detect problems early, reduces the risk of false positives, and maintains quality as the agent evolves. This process provides a form of automated, repeatable quality assurance for agent testing. It ensures the agent meets your business's accuracy and reliability standards and provides transparency regarding its performance. It has distinct advantages over testing using chatbots.

You run reviews and view results using the Copilot Studio interface , via the Power Platform REST API, or by adding actions in the tool, flow, or Power Automate.

Agent reviews measure accuracy and performance, not ethical or safety issues, of the AI. An agent might pass all review tests but still give an inappropriate answer to a question. Clients should still use responsible AI reviews and content safety filters; these reviews do not replace those reviews and filters.

Limitations of Government Community Cloud

Evaluating agents in a Government Community Cloud (GCC) environment has the following limitations:

1. Creators cannot add user profiles to their test suite. However, creators can still run reviews without user profiles.
2. The creator cannot use similarity testing methods for reviews. All other testing methods are available.

How the agent rating feature works.

Copilot Studio uses one test case for each agent evaluation. Each test case is a unique interaction that simulates how a user interacts with your agent. The interaction can be a single question or an entire conversation.

A test case could also include the answer you expect your agent to give. For example:

1. Question: What are your working hours?
2. Expected answer: We are open from 9 a.m. to 5 p.m., Monday through Friday.

By using evaluation agents, you can create, import, or write a group of test cases yourself. This group of test cases is called a test suite. A test suite allows you to:

1. Run multiple test cases that include various possibilities simultaneously, instead of asking your agent each question individually.
2. Agent performance analysis with an easy-to-understand composite score can also be considered in detail for each individual test case.
3. Test the changes to your agent using the same testing toolset, so you have an objective benchmark to measure and compare performance changes.
4. Quickly create new test toolkits or modify existing ones to include the agent's changing capabilities or requirements.

Each test can evaluate your agent using multiple testing methods simultaneously.

You can also select a user profile to act as the simulated user. The system can be configured to respond to different users in different ways, or allow access to resources in different ways.

When you select a test suite and run a system review, Copilot Studio will send questions in test cases, record the system's responses, compare those responses to expected responses or quality standards, and assign scores to each test case. You can also view details, logs, and activity maps for each test case and the resources the system used to generate the responses.

Develop a comprehensive evaluation strategy.

Before running the evaluation, define success for the system and decide which scenarios are most important to your business results. A clear strategy helps you choose the right testing methods, prioritize high-impact test cases, and interpret results in the appropriate context.

1. Utilize the System Solution Architecture: Evaluation Framework to map business objectives to measurable evaluation dimensions and scoring methodologies.
2. Utilize System Design and Operational Assessment to build repeatable assessment processes that support continuous quality improvement.

Integrate evaluation into automated flows.

Agent evaluation supports automation so creators can run evaluations without manual intervention. By using the REST API or the Power Platform connector, you can programmatically trigger evaluation runs and integrate testing into automated workflows such as continuous integration and continuous deployment (CI/CD). This approach allows you to run test suites at scale and validate agent behavior as changes are introduced, without manual intervention in Copilot Studio.

Chat test versus agent rating

Each testing method provides you with different insights into the agent's qualities and behavior:

Test chat:

1. Receive and answer one question at a time. It's unlikely you'll repeat the same test multiple times.
2. Allows you to view an entire session containing multiple messages.
3. This allows you to interact with your agent as a user through the chat interface.

Agent evaluation:

1. You can create and run multiple test cases simultaneously using a test suite. You can repeat the tests using the same test suite.
2. You can test one question and one answer per test case, or one conversation per test case. However, you have less control over the conversations compared to using the test chat feature.
3. Choose different user profiles to simulate different users without having to complete the interactions yourself.

When testing agents, use both chat testing and agent rating features to get a comprehensive view of your agent.

You finished reading the article "**Overview of agent evaluation**" edited by the **TipsMake** team. We hope this article has provided you with many useful tech tips and tricks. You can search for similar articles on tips and guides. Thank you for reading and for following us regularly.

© 2019 TipsMake.com