GitHub Copilot Evals enters public preview: testing your AI setup like software

18 November 2026 · 4 min read · blogs.iictp.org/github-copilot-evals-public-preview-testing-ai-setup

GitHub announced on 17 November 2026 that GitHub Copilot Evals is in public preview. GitHub says the tool helps users check how the skills, custom agents and MCP tools in their Copilot setup behave, and whether a change made them better or worse.

TL;DR The short version

- GitHub Copilot Evals entered public preview on 17 November 2026, letting teams test how their Copilot skills, custom agents and MCP tools behave.
- An eval is a repeatable test: a realistic task, a description of a good result, and checks that show whether a change made things better or worse.
- Start with one customisation your team relies on, define a good result before changing it, and run each eval more than once.

Read online, listen to it, or change the text size: https://blogs.iictp.org/github-copilot-evals-public-preview-testing-ai-setup

What was reported

- GitHub describes an eval as a repeatable test for AI behaviour: give an agent a realistic task, describe a good result, and check both the result and how the agent reached it.
- Users can ask Copilot in their own words or use the /eval command to create an eval, run it, see which checks passed or failed, investigate failures and compare runs from before and after a change.
- GitHub says each run is approved by the user before it starts, and that the evaluation engine is Vally, which it describes as an extensible local eval tool.
- It is offered for Visual Studio Code, Copilot CLI and the GitHub Copilot app, on Copilot Individual, Business and Enterprise plans.
- These are GitHub's descriptions of a preview product. This post has not tested it or measured its results.

What a practitioner should do

- Start with one customisation your team relies on, and write down what a good result looks like before you change it.
- Run an eval more than once. GitHub itself notes that AI output can change between runs, so one good answer proves little.
- Build checks from real tasks and past failures, and check the steps as well as the final answer, especially where an agent can call tools.

- Rerun the same eval after changing a prompt, model, skill or tool, and keep the results so a regression can be traced.
- Read each run’s scope before approving it, and keep secrets and production access out of eval environments.

The wider lesson

Evals bring a familiar engineering habit to a less predictable kind of software: define the expected behaviour, test it, and test it again after every change. Whichever tool you use, the judgement about what counts as good remains yours. Verifying AI output in this way is central to the Applied AI domain of the TPBOK.

> Facts in this post were read from the sources listed below on 18 November 2026; this is a developing story or product and details may have changed since.

Sources and further reading

1. GitHub Changelog. GitHub Copilot Evals is now in public preview. 17 November 2026. github.blog/changelog/2026-11-17-github-copilot-evals-is-now-in-public-preview

Published by IICTP, an independent technical credential body, on 18 November 2026. The current version of this post is at https://blogs.iictp.org/github-copilot-evals-public-preview-testing-ai-setup. More guides: blogs.iictp.org.