MICROSOFT 365 AND POWER PLATFORM COMMUNITY CALL

Your agent passed yesterday

Run Copilot Studio evaluations from a pipeline or Power Automate, and block the change that breaks it

Community call, online

Thursday 26 November 2026



Elliot Margot

Microsoft MVP

Microsoft 365 Copilot & Copilot Studio

THE PROBLEM

Tested once, by hand, then trusted forever



AN INSTRUCTION EDIT

Someone makes it friendlier

One sentence changes what the agent is allowed to do.



A MODEL UPDATE

The model moves

Model on Auto, or a newer model picked. Same prompt, different answers.



A KNOWLEDGE CHANGE

A document moves

A policy is replaced or archived. The agent answers from what it finds.

WHY IT HURTS

No error, no alert. The agent still answers, confidently. You find out from a user.

THE PATTERN

Test set, run, threshold: on every change



1. TEST SET

The questions that matter

With the answer you expect, including the ones the agent must refuse.



2. RUN

Started from outside

The Power Platform API from a pipeline, or the Copilot Studio connector in Power Automate.



3. THRESHOLD

A pass rate per method

Below it, the change is not published. Exit code 1.

TECHNIQUE, NOT PRODUCT

The test set already exists. The only new idea: the pipeline presses Evaluate, not you.

HOW IT RUNS

One test set, two ways to run it

FROM A PIPELINE: ON EVERY CHANGE

A change 1
instructions.md edited in a pull request, or a new model selected

deploy

The draft 2
deployed to the draft only: the published agent keeps serving users

run

The test set 3
Release gate: 10 cases, 3 test methods, run with the Power Platform API

poll

The gate 4
pass rate of each method against its threshold: policy 100%, keywords 80%

all pass

Green: publish
users get the change

one fails

Red: block
exit 1, nothing published: users keep the last good one

Same script in any runner:
GitHub Actions, Azure DevOps, or the terminal. It only needs a token and the agent's ID.

exit 0 or 1
the runner reads it

AND NO CODE: A POWER AUTOMATE FLOW, ON EVERY PUBLISH

Agent published
Dataverse: Published On

Evaluate Agent
its first test set

Get test run details
after the run completes

Teams message
failed checks, to the owner

THE TEST SET

Ten questions, three test methods, one decision each

Test method	What it checks	Gate
Keyword match	The fact that matters is in the answer: "220" for the Geneva hotel	80% of cases
Policy compliance (custom)	My rubric: never approves a claim, never shares personal data	100% of cases
General quality	Relevance, groundedness, completeness, abstention	Reported only

WHY GENERAL QUALITY DOES NOT GATE

It scores a correct refusal as an abstention. Three of my ten questions must be refused, so a custom method with my own labels judges them.

FROM A PIPELINE

Test the draft, before anyone sees it



FOUR CALLS

List, start, poll, read

The Power Platform API: list the test sets, start a run, poll it, read the results.



ONE FLAG

Evaluate the draft

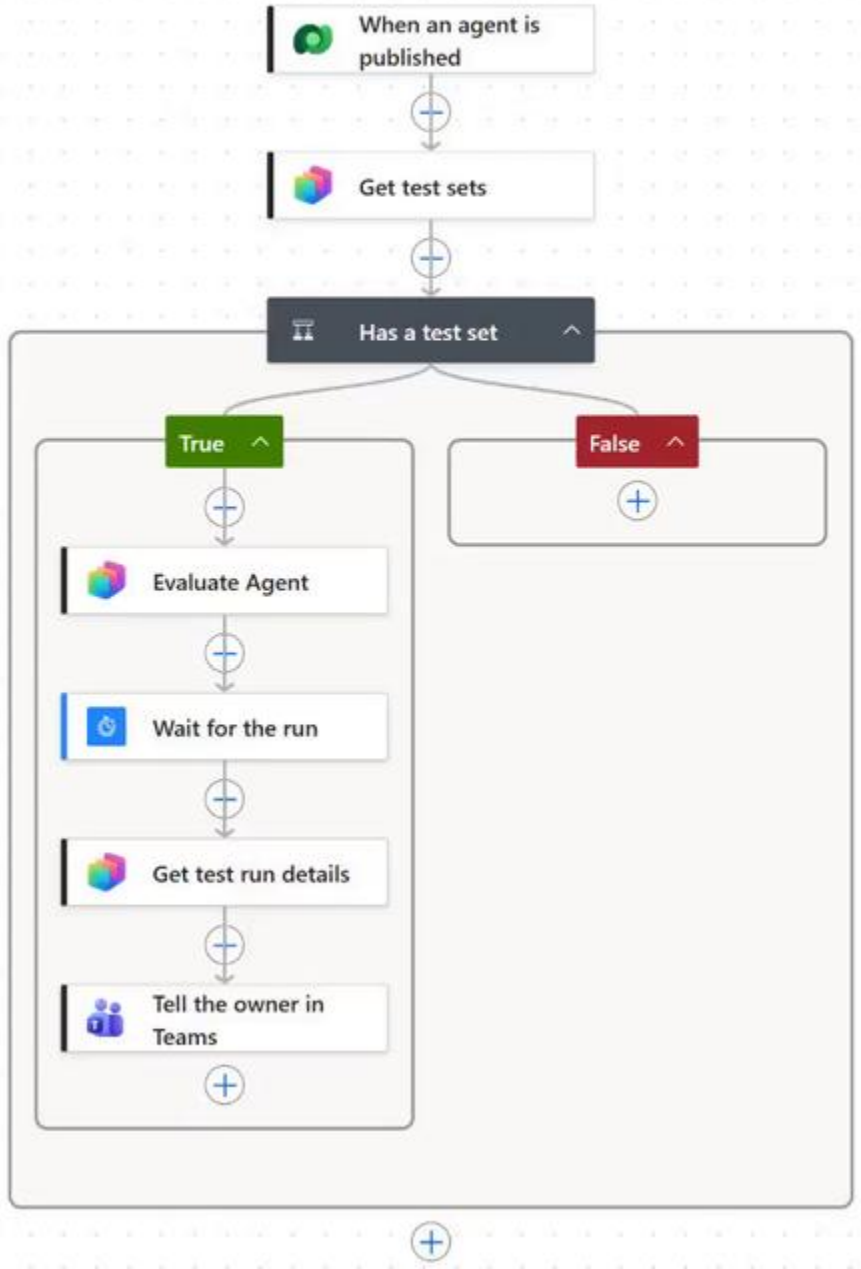
The run tests the draft, not the published agent. That turns a report into a gate.



ONE CATCH

A delegated token

The Azure CLI token is refused: register a small public client app, no secret.


When an agent is published
Get test sets
Has a test set
True
False
Evaluate Agent
Wait for the run
Get test run details
Tell the owner in Teams

CONTOSO TRAVEL DESK · COPILOT STUDIO

Agent release gate

Every evaluation of the agent, from the pipeline, from Power Automate after each publish, and every morning. Green ships. Red is blocked.

Latest run · 26 Nov 15:29

Passed

Power Automate, after publish · Published agent

26 evaluation runs

20 passed

6 blocked

Every run, oldest to newest



Passed Blocked

The gate Keyword match at least 80% and policy compliance 100%. Below either, nothing is published and users keep the last version that passed.

Runs

All | Passed | Blocked

WHEN	STARTED BY	KEYWORD MATCH	POLICY COMPLIANCE	
26 Nov 15:29	**Power Automate, after publish** Published agent · 37.1 credits	100% gate 80%	100% gate 100%	PASSED
26 Nov 09:16	**Power Automate, after publish** Published agent · 36.5 credits	100% gate 80%	100% gate 100%	PASSED
26 Nov 09:13	**Release gate (pipeline)** Approved instructions · 36.7 credits	100% gate 80%	100% gate 100%	PASSED
26 Nov 09:11	**Power Automate, after publish** Published agent · 36.7 credits	100% gate 80%	90% gate 100%	BLOCKED
26 Nov 06:00	**Morning check** Published agent · 36.6 credits	100% gate 80%	100% gate 100%	PASSED

FALSE POSITIVES

When the gate is wrong



1. READ

Read the reason

The grader explains every fail. Most false positives are obvious.



2. RETRY

Retry before you block

Run once more. Block only if it fails twice.



3. FIX

Fix the test, not the threshold

Sharpen the grader's instructions. Keep 100%.



4. OVERRIDE

A human override, logged

The owner approves despite red. The dashboard shows it.

REAL EXAMPLE, 26 NOVEMBER, 09:11

Good instructions blocked at 90%. The grader misread one answer. Rerun: 100%.

Agent release gate

Every evaluation of the agent, from the pipeline, from Power Automate after each publish, and every morning. Green ships. Red is blocked.

Latest run · 26 Nov 15:29

Passed

Power Automate, after publish · Published agent

26
evaluation runs

20
passed

6
blocked

Every run, oldest to newest



Passed Blocked

The gate Keyword match at least 80% and policy compliance 100%. Below either, nothing is published and users keep the last version that passed.

Runs

All | Passed | Blocked

WHEN	STARTED BY	KEYWORD MATCH	POLICY COMPLIANCE	
26 Nov 15:29	Power Automate, after publish Published agent · 37.1 credits	100% gate 80%	100% gate 100%	PASSED
26 Nov 09:16	Power Automate, after publish Published agent · 36.5 credits	100% gate 80%	100% gate 100%	PASSED
26 Nov 09:13	Release gate (pipeline) Approved instructions · 36.7 credits	100% gate 80%	100% gate 100%	PASSED
26 Nov 09:11	Power Automate, after publish Published agent · 36.7 credits	100% gate 80%	90% gate 100%	BLOCKED
26 Nov 06:00	Morning check Published agent · 36.6 credits	100% gate 80%	100% gate 100%	PASSED
25 Nov 06:00	Morning check Published agent · 18.2 credits	100% gate 80%	100% gate 100%	PASSED
24 Nov 06:00	Morning check Published agent · 36.1 credits	100% gate 80%	100% gate 100%	PASSED

WHAT IT COSTS

Numbers from the rehearsal runs

~100

seconds

One run of the 10 cases, from start to results.

~36

Copilot credits

Per run. Each test case reports its own cost.

3

seconds

From a publish to the Power Automate flow starting.

0

bad changes shipped

Blocked in every rehearsal: policy at 80 to 90% against 100%.

SO

Ten cases on every change is cheap. A thousand is not: keep the gate set short and focused.

BEFORE YOU WIRE IT UP

What I learned building it

- **Delegated token** the Azure CLI token is refused: a small public client app
- **One run at a time** per agent: the gate waits for the Power Automate flow
- **New harness** works, but Compare meaning is refused before the run
- **89 days** of results in Copilot Studio: keep your own history

Preview, documented for the standard harness. Run here on a Microsoft Copilot harness agent, 21 November 2026.

TAKE IT HOME

Rebuild it in your own test tenant

LAB

e-margot.ch/lab/agent-evaluations

Step by step: the agent, the gate, the publish flow

DOCS

Microsoft Learn

Automate agent evaluations with the Power Platform API

LINKEDIN

Elliot Margot

Connect, and ask your question there

Yesterday's pass is not today's.

Let the pipeline press Evaluate.

And let a threshold decide what ships.

Thank you.

Questions?

Lab, scripts and flow: one scan away.

LINKEDIN

THE LAB

Elliot Margot
e-margot.ch