LUMINOS.AI

WHITEPAPER

COPILOT STUDIO · PRODUCTION READINESS

5 Gaps to Close Before Your Copilot Studio Agents Go Live

How to get the agents you build in Copilot Studio ready for production.

AUTHORED BY
LuminosAI

FOR
Copilot Studio builders

READING TIME
~5 minutes

Copilot Studio has made building agents easier than ever. According to Microsoft, **nearly 90 percent of the Fortune 500** now have agents built with its low-code tools. But for agents to have an impact, the next step is scaling them, and that starts with making sure each agent is ready for production.

If you're using Copilot Studio, you may already be testing your agents. Copilot Studio even offers built-in evaluation tools that Microsoft keeps improving, meant to help you measure an agent's performance. So you try a few prompts, read the answers, and if they look good, you ship. That's a great start, but production takes a few more steps.

Getting an agent ready for production also means **testing for the risks unique to your business**, because what works in a controlled environment may not work in the real one.

FIVE GAPS TO CLOSE BEFORE YOU GO LIVE

01 **Deciding which risks matter.** The hardest part, and it takes specific expertise.

02 **Gaps in your test data.** A test only finds what its data touches.

03 **Testing once isn't enough.** Agents change after launch.

04 **Treating evals as a team sport.** The builder can't sign off alone.

05 **Fixing what you find.** A failed test should tell you what to change.

01

Deciding which risks matter

THE GAP

Choosing the right risks is the hardest part of testing an agent, because it requires a very specific type of expertise. Copilot Studio gives you a foundation of quality and safety checks. It's up to you to decide which risks matter most.

Copilot Studio's default grader, General quality, checks four things: relevance, groundedness, completeness, and whether the agent attempted an answer. Those measure how good an answer is. But production also requires knowing whether the agent should have given it.

Some agents also get content safety checks for common risks like hate, violence, and self-harm. Microsoft's own documentation notes that these checks don't guarantee an agent is safe or appropriate for every scenario.

The risks that matter most are specific to your business:

> Does the **HR agent** reveal another employee's information?
>
> Does the **finance agent** approve outside policy?
>
> Does the **contracts agent** give advice it shouldn't?

For these, Copilot Studio offers a flexible Custom grader, where you name each risk and write the instructions for detecting it. That works well if you already know which risks matter, and which ones auditors, regulators, and customers will ask about.

But that's not your job. You weren't hired to be a risk expert. Without that expertise, it's hard to know where to start.

02

Gaps in your test data

> THE GAP
>
> **Your tests are only as good as your data. If your test data doesn't cover each risk, a passing score can't tell you much.**

An evaluation runs your agent against a set of test questions. If those questions never touch a risk, the test can't find it. A perfect score on the wrong data tells you very little.

In Copilot Studio, test data comes from one of two places. Either you write it, or Copilot Studio generates it from your agent's own description and instructions. Generated data is a fast way to test the agent on what it was built to do. But real users will also try edge cases, off-topic requests, and attempts to get around the rules.

To understand a risk, you need data built for that risk. If the risk is an HR agent oversharing salaries, you need questions that probe for salaries, asked by employees who shouldn't see them. You also need to know whether your test set covers the risks that matter, and which ones it misses. If you're testing with data that has major gaps, you need to know that, too.

If your data has gaps, so does your test.

03

Testing once isn’t enough

> THE GAP
>
> **A test is a snapshot, not a safeguard. Copilot Studio tests your agent when you click Evaluate. Once the agent goes live, you need testing that keeps up.**

Copilot Studio evaluations run when someone starts them. Microsoft also offers APIs and Power Automate connectors to run evaluations automatically, even inside CI/CD pipelines. Setting those up takes technical work that many no-code business builders aren’t set up to do.

Meanwhile, agents change. Models get updated. Someone edits a knowledge source. A connector changes. An agent that passed in March can fail in June, and no one finds out until it says the wrong thing to a user.

Nobody tests payroll software once at launch and walks away. Agents behave differently every time they run and can drift over time, so **they need more testing after launch, not less.**

04

Treating evals as a team sport

> THE GAP
>
> **Testing an agent takes more than the person who built it. Copilot Studio’s evals are designed for the builder. Getting to production usually takes the whole team.**

The person who builds an agent is rarely the only person who should test it. Each of these people sees risks the builder can’t, and each needs to see the results before the agent goes live:

> **Business owner** knows what a correct answer looks like.
>
> **Compliance** knows the rules the agent has to follow.
>
> **Security** knows what data it shouldn’t touch.
>
> **Legal** knows what it shouldn’t say.

Copilot Studio’s evaluation tools live where the builder works. The rest of the team needs an easy way to weigh in on which risks matter, review the results, and sign off. Without one, the review either turns into a long chain of emails and meetings, or it doesn’t happen at all.

05

Fixing what you find

> THE GAP
>
> **Copilot Studio shows you when a test fails. It doesn't tell you how to fix it.**

A failed test is only useful if you know what to do next. Most builders don't. The agent gave the wrong answer, but was the problem in its instructions? Its knowledge sources? Its topics? Something else entirely? Without that answer, you're left guessing, changing things one at a time and hoping the next test passes.

You shouldn't have to guess, and you shouldn't have to become an expert to fix your own agent. What you need are instructions you can hand straight to the agent:

> *"Decline questions about other employees' salaries and direct them to HR."*
>
> *"Only answer refund questions using the current refund policy."*

Paste them in, and the agent does the rest. No need to know which setting to change or why.

Then test again. **If the fix worked, you'll know. If it didn't, you'll know that too, before your users do.**

06

Close the gaps and scale your agents with LuminosAI

Every other kind of business software is tested before it ships. Nobody launches a payroll system on faith. Your agents shouldn't launch on faith either.

That's why we built LuminosAI, a no-code testing platform for AI agents built to work alongside Copilot Studio. It answers one question: **does your agent do what it's supposed to do?**

And it does the hardest part for you. LuminosAI custom generates the most important risks for your agent, based on what it does and who it serves. You don't have to decide which risks matter. **That's not your job. It's ours.**

It closes the other four gaps too. LuminosAI generates test data around each risk and flags any gaps, so you know which risks you're testing and which you aren't. You can retest in minutes whenever your agent changes. Every test produces a complete report showing what was tested, against which risks, and what to fix, and you can customize it and assert legal privilege if you want. And you can share evals with your team, even before they're finished. For every failure, you get instructions you can give directly to your agent to fix it.

HOW IT WORKS WITH COPILOT STUDIO

1. **Generate test data** with the tools already in Copilot Studio, or let LuminosAI generate it for you. We'll help you understand whether the data is good enough to test.
2. **Drag and drop that data into LuminosAI.** LuminosAI generates the risks that matter most for your agent and tests for each one.
3. **Get a complete report** that shows where your agent performs as intended and where it doesn't.
4. **Improve your agent** back in Copilot Studio, using the instructions in the report.

It's fast, and it requires no code. Watch the 5-minute demo to see it run on a real Copilot Studio agent.

SIDE BY SIDE

GAP	BUILT INTO COPILOT STUDIO	COPILOT STUDIO + LUMINOSAI
1. Deciding which risks matter	Quality and content safety checks, plus a Custom grader where you define other risks yourself	**The most important risks for your agent, custom generated for you**
2. Gaps in your test data	Data you write, or data generated from your agent's description	**Data generated around each risk, with gaps flagged so you know what you're testing and what you aren't**
3. Testing once isn't enough	Manual runs, or pipelines you build	**Retest in minutes whenever your agent changes**
4. Evals as a team sport	Tools designed for the person building the agent	**A report the whole team can review before launch, and evals you can share while they're in progress**
5. Fixing what you find	Shows which tests failed	**Instructions you give directly to your agent to fix each failure, plus a retest in minutes to confirm it worked**

A NOTE TO EVERYONE BUILDING AGENTS WITH MICROSOFT

Keep Copilot Studio. Keep Agent 365. They're excellent at what they do. LuminosAI doesn't replace them. **It fills the gap between building an agent and trusting it, so more of the agents you build make it to production.**

The future of AI is not a few projects run by specialists. It's thousands of agents built by the people closest to the work. That's you. You need to move fast, and your proof has to move just as fast.

SEE IT ON YOUR OWN AGENTS

Test your first agent. Free.

LuminosAI is open to everyone with a free tier. If your agent passes, you'll know. If it fails, you'll know that too, before anyone else does.

Test your first agent Watch the 5-minute demo →

REFERENCES *Documentation accessed October 28, 2026.*

1. Microsoft Corporation. "Transcript: Microsoft FY26 Third Quarter Earnings Conference Call." April 29, 2026. https://www.microsoft.com/en-us/investor/events/fy-2026/earnings-fy-2026-q3
2. Microsoft Corporation. "About Agent Evaluation." Copilot Studio documentation. https://learn.microsoft.com/en-us/microsoft-copilot-studio/analytics-agent-evaluation-intro
3. Microsoft Corporation. "Choose Evaluation Methods." Copilot Studio documentation. https://learn.microsoft.com/en-us/microsoft-copilot-studio/analytics-agent-evaluation-overview
4. Microsoft Corporation. "Create a Test Set for an Agent (Preview)." Copilot Studio documentation (GitHub Copilot harness). https://learn.microsoft.com/en-za/microsoft-Copilot-studio/agents-experience/analytics-agent-evaluation-create
5. Microsoft Corporation. "Harnesses Overview." Copilot Studio documentation. https://learn.microsoft.com/en-us/microsoft-copilot-studio/harnesses-overview
6. Microsoft Corporation. "Create a Conversational Test Set." Copilot Studio documentation. https://learn.microsoft.com/en-za/microsoft-Copilot-studio/analytics-agent-evaluation-multi-turn
7. Microsoft Corporation. "Create a Single Response Test Set." Copilot Studio documentation. https://learn.microsoft.com/en-us/microsoft-copilot-studio/analytics-agent-evaluation-create#generate-a-single-response-data-set-from-knowledge-or-topics

LuminosAI is not affiliated with or endorsed by Microsoft. Microsoft, Copilot Studio, Microsoft 365 Copilot, and Agent 365 are trademarks of Microsoft Corporation. This paper reflects Microsoft's public documentation and existing features as of October 2026, and both may have changed since publication. We're also human, and we make mistakes. Some features may be tucked away where we didn't find them. If we couldn't find them, chances are you can't either. Either way, if we missed something or a statement no longer applies, please let us know at contact@luminos.ai. We review all feedback promptly and will issue updates as we learn of them.