



MISSION // SMARTCLOUD 365 - 2026

HR Companion Reimagined

Enterprise-ready AI agents with Microsoft Foundry

Nicole Enders

Microsoft MVP since 2020, 7× in a row · Die Agentin

24 NOVEMBER 2026 · ONLINE · AI AGENTS



Nicole Enders

Microsoft MVP since 2020. I work at the intersection of what is technically possible and what actually helps people in their working day.

MISSION

“Connecting people and organisations with intelligent solutions, so they can work future-proof, productive and smarter.”

● Strategy

Clear. Focused.

● Architecture

Secure. Scalable.

● Delivery

Pragmatic. Efficient.

● Adoption

Change. Acceptance.

● Enablement

Knowledge. Skills.

● Impact

Results. Value.



The assignment for the next 35 minutes.

01

Two dates

What retired between the abstract and today?

02

Four routes

Which one survives contact with your tenant?

03

Proof

How do you know the agent is actually grounded?

04

Consequences

What it costs, and when it expires.



MISSION

01

THE CALENDAR

This session was submitted in June. Two dates since then changed the answer.



The API this HR bot was built on retired three weeks ago.

- 26 August 2026: the Assistants API is gone. Threads, runs and create_agent() with it.
- 31 March 2027: classic agents follow. Two retirement dates for one stack, both from Microsoft.
- 1 October 2026: gpt-4o 2024-05-13 retires. Standard deployments auto-upgrade to gpt-5.1.
- Three renames on top: Azure AI Studio, Azure AI Foundry, Microsoft Foundry, same resource type.

KEY POINT

Renames are cosmetic. Retirement dates are not. A demo agent does not get old, it stops answering.

MISSION CALENDAR



SPACING NOT TO SCALE



Three questions separate a demo agent from one in production.

01

WHO IS ASKING

A demo runs as you. In production the agent runs as itself, or on behalf of the employee, and those are different systems with different limits.

02

WHAT MAY THEY SEE

Trimming is not a switch you turn on. It is a property of the retrieval path you picked, and some paths do not have it at all.

03

WHO PAYS PER ANSWER

Seats, tokens, retrieval calls and governance licences run at the same time. Most business cases count exactly one of the four.

None of these are model questions. All three are settled before you write a prompt.



This session promised two things that exclude each other today.

- Ground the agent in SharePoint, so employees get answers trimmed to what they may read.
- Publish it into Microsoft Teams, so people ask where they already work.
- The SharePoint tool is documented as not working once the agent is published to Teams.
- It needs a delegated user token on every call. A published agent has none to pass on.

KEY POINT

I wrote that abstract in June and it was already wrong. Say the uncomfortable part out loud, then fix the architecture.

THE PROMISE THAT DOES NOT HOLD

REQUIREMENT A

Grounded in SharePoint

Identity passthrough, trimmed to what each employee may read.

PREVIEW, NO SLA

REQUIREMENT B

Published to Microsoft Teams

Azure Bot Service, activity protocol, its own identity.

GA

NOT SUPPORTED TODAY

WHY

The SharePoint tool needs a delegated user token on every call.
A published agent has no user token to pass on.

MICROSOFT LEARN, 05 AUG 2026



MISSION

02

FOUR ROUTES

Grounding and distribution constrain each other. Pick both at once.



Foundry now offers three shapes of agent, not one.

01

PROMPT AGENT

Instructions, a model, tools. No code and no container. You pay for inference and tool calls. This is where an HR companion starts.

02

HOSTED AGENT

Your container and your framework, Agent Framework or LangGraph or your own. Managed endpoint, dedicated Entra identity, compute on the bill.

03

RESPONSES API

No agent resource at all. The definition lives in your application code and versions with it. Foundry supplies the models and the tools.

The choice is an operating model, not a difficulty level. Ask who owns this in a year.



Four routes to the same HR policies, and only two of them reach Teams.

- SharePoint tool: live query, nothing copied, permission-trimmed. Preview, and not in Teams.
- Foundry IQ indexed: a copy in AI Search with scheduled ACL sync. Works when published.
- Foundry IQ remote SharePoint: the same retrieval path, wrapped. Preview, same open question.
- File search: you copy the documents, so you also own the trimming and the drift.

KEY POINT

Every route that copies your documents hands you a second permission model to keep in sync with the first.

FOUR WAYS TO REACH THE HR POLICIES

	STATUS	IDENTITY	IN TEAMS
SharePoint tool Live query, nothing is copied.	PREVIEW	delegated	NO
Foundry IQ, indexed Indexed copy, scheduled ACL sync.	GA IN REST	ACL sync	YES
Foundry IQ, remote SharePoint, queried at run time.	PREVIEW	delegated	UNTESTED
File search Your own store, your own trimming.	GA	yours	YES



Microsoft publishes two different answers to whether Foundry IQ is GA.

- The Foundry migration page lists Foundry IQ as Preview, next to memory, workflows and A2A.
- The Azure AI Search docs call knowledge bases GA in the 2026-04-01 REST API.
- Both are current, both are Microsoft. GA depends on which API you call, not on the feature.
- Agent publishing to Teams, hosted agents and the Responses API are GA on both pages.

KEY POINT

Read the status per API version, not per product name. Then write that version into the architecture decision record.

TWO MICROSOFT PAGES, TWO ANSWERS

SOURCE A

Migrate from the Foundry classic portal

MICROSOFT LEARN, 19 JUN 2026

Foundry IQ: **PREVIEW**

SOURCE B

Create a Knowledge Base

AZURE AI SEARCH DOCS, AUG 2026

Knowledge bases: **GA**

BOTH ARE TRUE. THE ACCESS PATH DECIDES.

REST API 2026-04-01

GA

knowledge bases, core sources

REST API 2026-05-01-preview

PREVIEW

answer synthesis, permissions

Foundry and Azure portal

PREVIEW

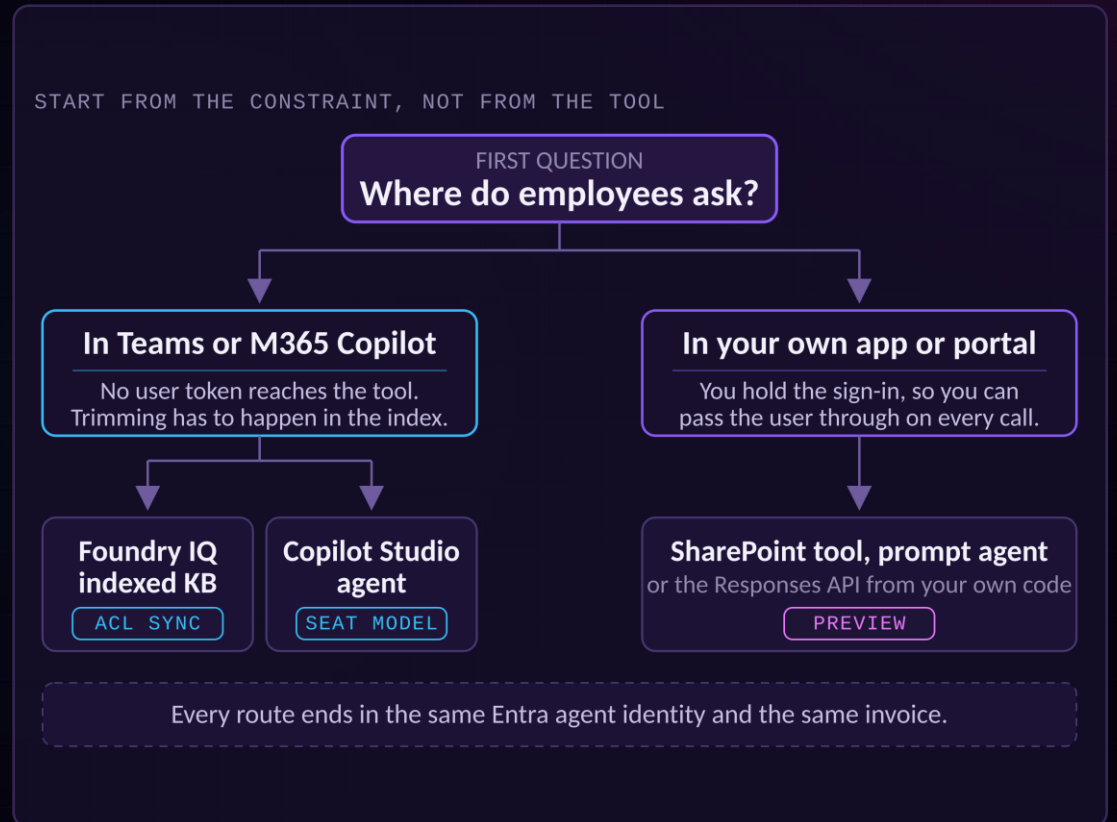
every agentic retrieval feature

Start from where people ask. That decides the retrieval path, not the other way round.

- Asked in Teams? Then no user token reaches the tool and trimming has to live in the index.
- Asked in your own app? Then you hold the sign-in and can pass the user through per call.
- Both routes land in the same Entra agent identity, the same traces and the same invoice.
- Copilot Studio is a legitimate answer on the left branch. Foundry is not always the answer.

KEY POINT

Choosing the tool first is how teams rebuild retrieval after the pilot instead of after the outage.



MISSION

03

PROOF

An agent is only grounded when two people get two different answers.



Demo



PERMISSION TRIMMING, LIVE, TWO IDENTITIES

Watch two things: the citation, and whether it disappears for the second person.

FALLBACK: SCREENSHOTS OF BOTH ANSWERS ARE READY



Two users is a spot check. It does not survive the next document.

- Evaluations, tracing and monitoring went GA in March 2026, on OpenTelemetry into Azure Monitor.
- Built-in evaluators cover groundedness and retrieval quality. Custom evaluators are possible.
- Continuous evaluation of quality.
- Run it against ground truth.

Evaluation runs

Select multiple runs to compare results statistically or use AI to analyze failed tests.

☐	Name	Target	Dataset	Status	Created by	Created on
☐	test2	test2: 1	eval-data-2026-03-... Version 1	↓ Completed		4/21/26, 11:16:32 AM

Evaluated system tokens896

Query666

Response230

Evaluation tokens18,053

Query17,013

Response1,040

ToolCallSuccessEvaluator

Evaluated system tokens: 896

Evaluation tokens: 18,053

KEY POINT

If you cannot say what your groundedness score was last Tuesday, you are not operating the agent, you are hoping.

← eval-target-model

▼ Evaluation details

Create time

Created by

See all properties

Evaluated system tokens896

Query666

Response230

Evaluation tokens18,053

Query17,013

Response1,040

ToolCallSuccessEvaluator

Evaluated system tokens: 896

Evaluation tokens: 18,053

Token usage

Evaluated system tokens: 300

Evaluation tokens: 14,734

Evaluators

Evaluators stay consistent across all runs in this group. To use different evaluators, exit this group and start a new evaluation.

Name

Violence

SelfHarm

GroundednessPro

MISSION

04

CONSEQUENCES

What this costs on Monday, and what expires before Christmas.



Four meters can run on a single answer, and most business cases count one.

- Model tokens on every turn, including the turns the agent takes before it replies.
- Retrieval at 0.10 USD per API call for anyone without a Copilot licence. Preview pricing.
- Container compute, but only for hosted agents. A prompt agent has none.
- Agent 365 is licensed per human, not per agent. Ten agents do not cost ten times.

KEY POINT

A refusal costs the same as an answer. Price the questions your agent cannot answer, not only the ones it can.

ONE ANSWER, FOUR METERS RUNNING

MODEL

Per token, on every turn

Including the turns the agent takes on its own before it answers.

RETRIEVAL

0.10 USD per API call

Pay-as-you-go for anyone without a Copilot licence. Preview, no SLA.

PER ANSWER

COMPUTE

Per container session

Hosted agents only. A prompt agent has no container to pay for.

GOVERNANCE

Per human, per month

Agent 365 is licensed per user, not per agent. Ten agents, same bill.



Your model has an expiry date. Your endpoint does not have to.

- GA models carry a retirement date from launch. After it, inference returns 410 Gone.
- Standard deployments auto-upgrade. Pinned and reserved ones are exactly the ones that do not.
- Agent versions snapshot automatically, and the published endpoint stays stable across them.
- Set the version selector deliberately. Always-latest ships to Teams the moment you save.

KEY POINT

Pinning a model version is not caution, it is a calendar entry. Put the retirement date in the backlog on day one.

The screenshot shows the Microsoft Foundry interface for the 'gpt-4o' model. The left sidebar contains navigation options: Create (Agents, Models, Services, Tools, Knowledge, Memory, Guardrails, Data), Optimize (Evaluations, Fine-tune), and a search bar. The main content area is titled 'gpt-4o' and has tabs for Playground, Details (selected), Monitor, and Evaluation. The Details tab shows a table of model specifications:

Model specifications	
Tokens per Minute Rate Limit	15000000
Requests per Minute Rate Limit	90000
Model name	gpt-4o
Date created	Dec 3, 2024, 1:00 AM
Model version	2024-11-20
Date updated	Dec 3, 2024, 1:00 AM
Lifecycle status	Legacy
Retirement date	Apr 14, 2027, 2:00 AM

Below the table, there is a section for 'Model guardrails' with a single entry: Name: Microsoft.DefaultV2.

Demo



OPERATE: TRACE AND COST OF THE RUN YOU JUST SAW

The refusal was not caution. Watch the tool call return empty, and what it cost.

FALLBACK: TRACE AND ASSET SCREENSHOTS ARE READY



Three things that stick.

1

Distribution decides retrieval

Where people ask constrains how you are allowed to ground. Decide both in the same meeting, or build retrieval twice.

2

Two users, every release

Ask one question as someone with access and someone without. If both get the same answer, nothing is trimmed.

3

Read status per API version

GA and preview belong to the API you call, not to the product name. Write the version into the decision record.

Next step: run the two-user check against whatever agent already runs in your tenant this week.

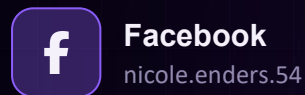
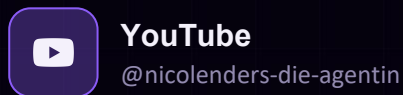
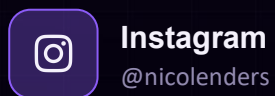
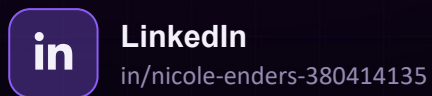




THANK YOU.

Questions, disagreement and field reports are explicitly welcome.

CONTACT · NO LOOSE ENDS



WEBSITE

nicolenders.com

Identities, missions, briefings and dispatches in one place.



SLIDES FOR THIS TALK

nicolenders.com/briefings

All briefings with slides, recordings and dates.

The slides go online after the talk. Scanning is enough, no need to take notes.