



# Agents Series: Deep reasoning in Copilot agents

Aline Tognini  
Principal Cloud Solution Architect

Presented on 4/28/2025



# Problem Statement

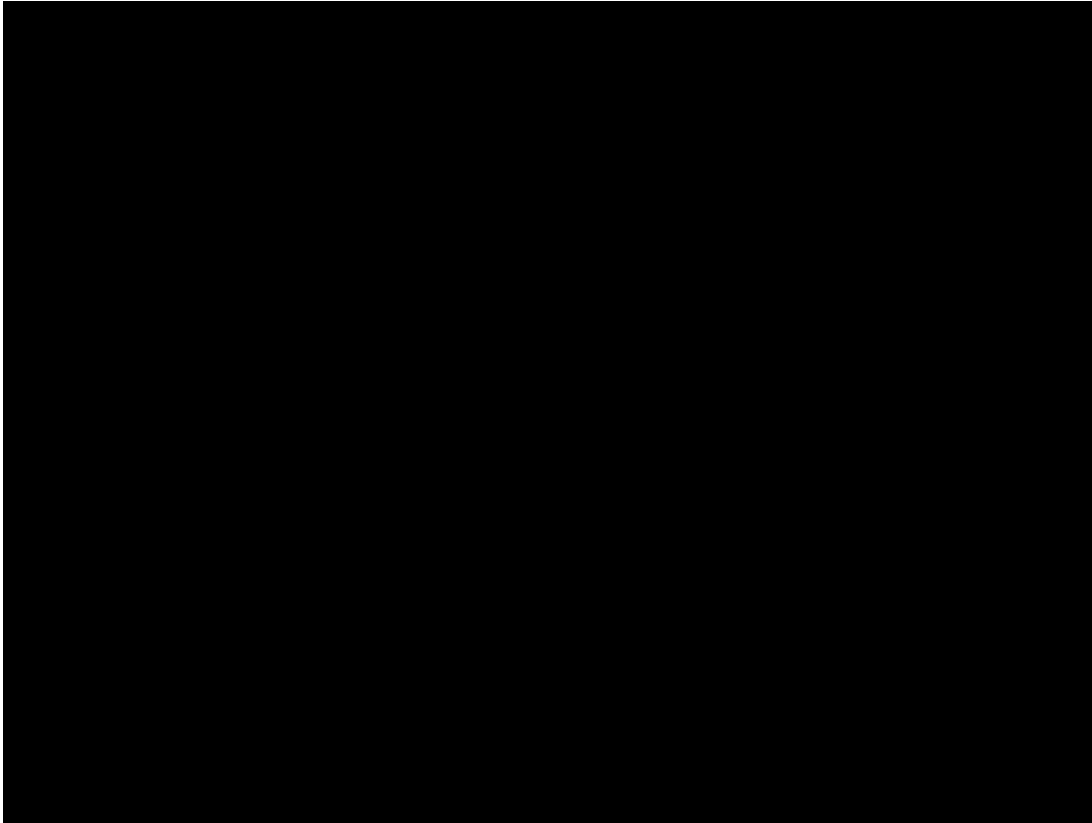
Professionals and academics in healthcare and life sciences need gen AI assistants **trained specifically for research tasks**, able to **emulate research methodologies** through **iterative reasoning** and **distilling extensive data**.

**The answer:** Copilot agents that can perform advanced reasoning\* by leveraging Azure OpenAI o1 series models and complete tasks requiring **logical reasoning, problem solving, and step-by-step analysis**.

*\*In preview*



# Importance of Visibility




- **Transparency in AI Decisions**
  - Visibility into Copilot agent's thought process is crucial for understanding and trusting its decisions in complex scenarios.
- **Documentation in Research**
  - Documenting every step and decision in the research process ensures thorough and credible outcomes, akin to AI transparency.
- **Ensuring Reliability**
  - Visibility in thought process assures users of its reliability and accuracy, similar to transparent workflows in research.

# Speed vs. Accuracy: A Balancing Act

- Speed can enhance productivity in various tasks.
- Accuracy ensures quality and correctness of outcomes.
- Too much speed can compromise accuracy in scenarios that require precision.
- What is the cost associated with it?





Demo

# Utilization rates depend on type of agent and prompt

\*Rate changes in effect as of April 1, 2025

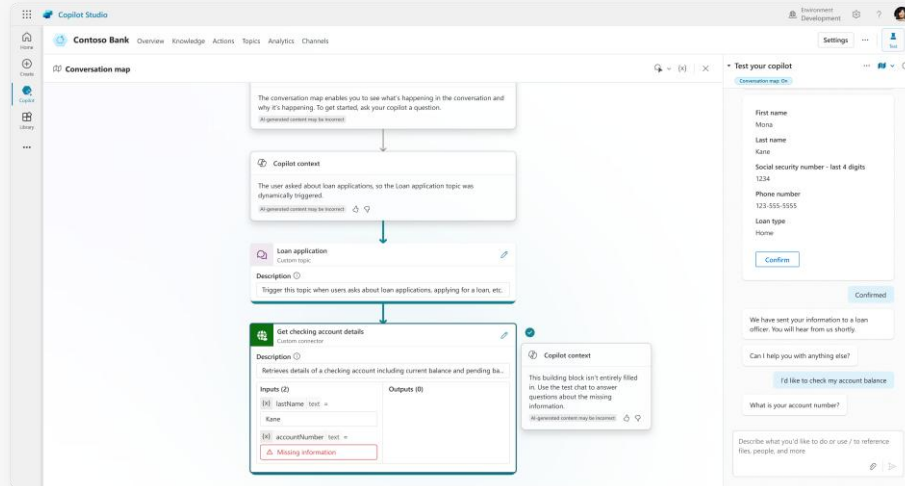
	Orchestration Mode	M365 Copilot Users	Copilot Chat Users	Use of Other Agents Built w/ Copilot Studio
<b>Web-grounded answers</b> Dynamically-generated responses based on the web as a knowledge source.	Classic & Generative	0	0	2 messages
<b>Classic answers</b> Predefined responses manually authored by makers through topics (includes messages, connectors, flows etc.) that are static unless manually updated in Classic Orchestration mode. Used when a precise or controlled response is desired output. Each action (not each topic) counts as an answer. Not available in agent builder.	Classic only	0	1 message	1 message
<b>Generative answers 1,2</b> Dynamically-generated responses based on knowledge sources and context that provide flexible and natural interactions.	Classic & Generative	0	2 messages	2 messages
<b>Tenant graph grounding for messages 1,2</b> Grounding to enhance AI agents with up-to-date, context-aware knowledge from Microsoft 365 and external data, offering built-in security and inheriting data access governance policies.	Classic & Generative	0	10 messages*	10 messages*
<b>Agent actions 1,2</b> AI-led orchestration for triggers, topics, agent flows, text & generative AI tools, Power Platform premium connectors and custom connectors to automate complex business processes. Not available in agent builder.	Generative only	0 <sup>4</sup>	5 messages*	5 messages*
<b>Text &amp; generative AI tools</b> Specialized tools that extend agents capabilities by teaching them to perform specific tasks, leveraging a combination of AI prompt engineering, model configuration, code execution, and knowledge retrieval	-	-	-	-
<b>Basic</b> (Message rate per 10 responses <sup>3</sup> )	Classic & Generative	1 message*	1 message*	1 message*
<b>Standard</b> (Message rate per 10 responses <sup>3</sup> )	Classic & Generative	15 messages*	15 messages*	15 messages*
<b>Premium</b> (Message rate per 10 responses <sup>3</sup> ) For deep reasoning prompts	Classic & Generative	100 messages*	100 messages*	100 messages*
<b>Agent flow actions</b> (Message rate per 100 agent flow actions) Agent flow actions are used to create agent flows. Agent flows are rules-based automations in Copilot Studio that follow a predefined sequence of flow actions to perform repetitive tasks.	Classic & Generative	13 messages*	13 messages*	13 messages*

## Notes

- Each interaction with an agent could utilize multiple utilization rates simultaneously i.e., an agent grounded in Tenant graph could use 12 messages (10 for the graph grounding and 2 for Generative Answer) to respond to a single complex prompt from the user. Most agents built natively in SharePoint or Copilot Chat will have tenant graph grounding enabled by default.
- Generative answers, tenant graph grounding for messages, web-grounded answers and agent actions apply to both declarative agents and custom engine agents.
- 1 response = 1,000 tokens for LLM models, 1 image for image processing, 1,000 characters for text processing and 1 row when processing rows for prediction. Billing will be prorated to exact number of responses.
- Agent actions are included at no additional cost for interactive use only. Autonomous use will incur a 5 message charge

# Orchestration modes for your agent

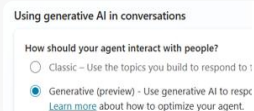
## Generative mode



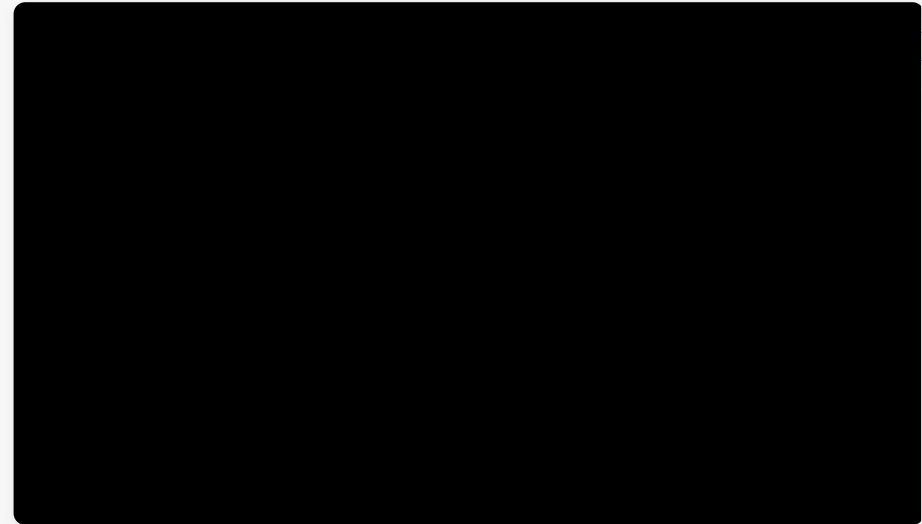
**WHAT** Uses LLM-based engines to pick the best actions, knowledge, and topics to respond to queries or event triggers.

**WHY** Agents needing to be dynamic, flexible and respond to ambiguous scenarios deciding the best course of action at runtime.

**HOW** Select "Generative" radio button in Generative AI tab of the settings page.



## Classic mode



**WHAT** Responds by triggering topics through specific trigger phrases that are the closest match to the user's query.

**WHY** Traditional bots that restrict the agents to a limited set of predefined responses and processes.

# Classic answers

## WHAT

Predefined responses manually authored by makers through topics (includes messages, connectors etc.) that are static unless manually updated in Classic Orchestration mode. Used when a precise or controlled response is desired output. Each action (not each topic) counts as an answer.

---

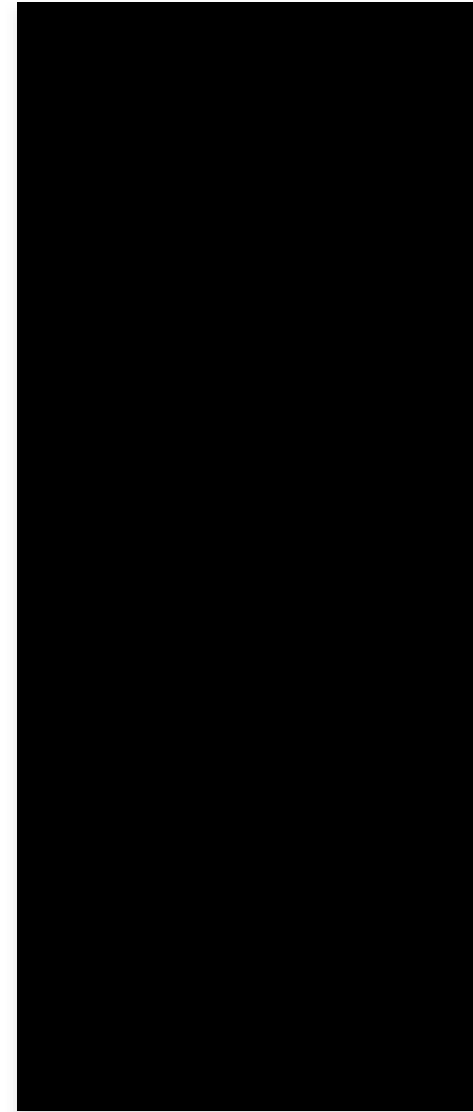
## WHY

Ensures consistent and accurate responses for frequently asked questions while handling the specific set of steps the agent needs to take.

---

## HOW

Configure the agents' topics in Copilot Studio to follow the specific set of actions and responses required.



# Generative answers

## WHAT

Dynamically-generated responses based on large language models that provide flexible and natural interactions using knowledge sources and context.

---

## WHY

Offers more natural and contextually relevant answers to enhance user experience.

---

## HOW

Configure a “conversational boosting” topic or add a generative answers node in the topic.

