

2026 Responsible AI Transparency Report

Foundry Agent Service

Addressing novel memory capabilities for agents

In fall 2025, Microsoft released memory capabilities in Foundry Agent Service as a preview, allowing developers to test the ability to create and manage long-term memory for their agents. This preview capability was designed to support continuity across agent conversations, devices, and workflows, with the goal of allowing agents to deliver more personalized experiences over time.¹

Because memory (preview) can store information that may be sensitive or personal, this feature raised novel risks that warranted additional oversight. As a result, prior to the release of memory (preview), the team underwent a Sensitive Uses review where they collaborated closely with the Sensitive Uses team in the Office of Responsible AI to map, measure, and manage risks associated with agent memory (preview).

Map

The team identified ungrounded memory as a primary risk. This occurs when the AI agent might rely on a memory that is inaccurate, outdated, misattributed, or insufficiently tied to the current context. For example, an agent may hallucinate or confidently recall information that was never provided; over time, such erroneous memories could compound and lead to misaligned behavior. The team also identified risks related to the storage of sensitive or personal information in agent memory (preview), including data such as precise location, biometric identifiers, or political opinions.

Measure and manage

To help mitigate risks associated with sensitive data, the team implemented an editable meta-prompt, enabled by default, to help prevent the memorization of sensitive attributes. These attributes include, for example, health information, religious beliefs, and immigration status. The team can continuously update the list of sensitive attributes as new ones emerge, allowing for flexibility in mitigation measures.

To address risks of ungrounded memory, the team conducted adversarial testing to evaluate the robustness of agent memory (preview) against threats such as memory poisoning, indirect prompt injection attacks, and other jailbreak attempts. These evaluations helped test and strengthen the security of agent memory (preview) and its resistance to manipulation that could result in inaccurate or misleading recalls.

In addition, the Sensitive Uses team advised the Foundry team to implement user controls and transparency features, including:

- Visibility into stored memories
- The ability for developers to enable deletion of individual memories
- The ability to delete all memories associated with a user
- Options for developers to offer opt-in and opt-out controls for memory functionality

2026 Responsible AI Transparency Report

Release

Prior to the release of memory (preview), the Sensitive Uses and Foundry teams collaborated to update transparency documentation, clearly describing feature capabilities, limitations, and recommended best practices. This documentation includes guidance on securing agents against jailbreak attacks and evaluating the impact of memory (preview) in AI applications.² Memory was released as a public preview, enabling customers to test and use the capability. This approach has allowed the team to gather early feedback and monitor performance with a smaller audience that will help inform any future product launch stages for memory (preview).



2026 Responsible AI Transparency Report

Endnotes

- 1 Introducing Memory in Foundry Agent Service - Microsoft Foundry Blog.
<https://devblogs.microsoft.com/foundry/introducing-memory-in-foundry-agent-service/>
- 2 What Is Memory? - Microsoft Learn.
<https://learn.microsoft.com/en-us/azure/foundry/agents/concepts/what-is-memory>