Copilot for Microsoft 365

Risk Assessment QuickStart Guide

Contents

Disclaimer

Executive Summary

Copilot for Microsoft 365 is an intelligent assistant that helps users enhance productivity using relevant information and insights from their files, meetings, and communications whether they are stored in SharePoint, OneDrive, Outlook, Teams, or a third party solution connected through plugins or extensions. Copilot for Microsoft 365 uses natural language processing and machine learning to understand user queries and provide personalized results. Copilot for Microsoft 365 can also generate summaries, insights, and recommendations based on the content of user inputs and the Microsoft Graph.

This document is a QuickStart guide for organizations that want to perform a risk assessment of Copilot for Microsoft 365 as part of their due diligence, internal approval, or service curation process. It provides an overview of the potential AI risks and how these risks are mitigated for Copilot for Microsoft 365. The document is intended to be a starting reference point for risk identification, mitigation exploration, and discussion with various stakeholders involved in the assessment process. This is not a comprehensive or definitive assessment and should not be considered the final risk assessment product.

The document is structured as follows:

AI risks and mitigations framework. This section introduces the main categories of AI risks and how Microsoft addresses them at the company level and at the service level for Copilot for Microsoft 365. The framework covers topics such as security, resilience, bias, disinformation, hallucination, and privacy.

Sample risk assessment questions and answers. This section presents a sample set of questions and answers that can be used to assess the features, functionalities, and the surrounding security and compliance posture of Copilot for Microsoft 365, as well as the broader implications and impacts of using the service. The questions are derived from real enterprise customer inquiries and the answers are based on information from various Microsoft teams and sources. Some responses also include direct attestation from OpenAI, a Microsoft strategic partner from which we source our Copilot for Microsoft 365 foundation models.

Additional resources. This section provides links to additional materials and resources that can help organizations learn more about Copilot for Microsoft 365 and AI risk management generally.

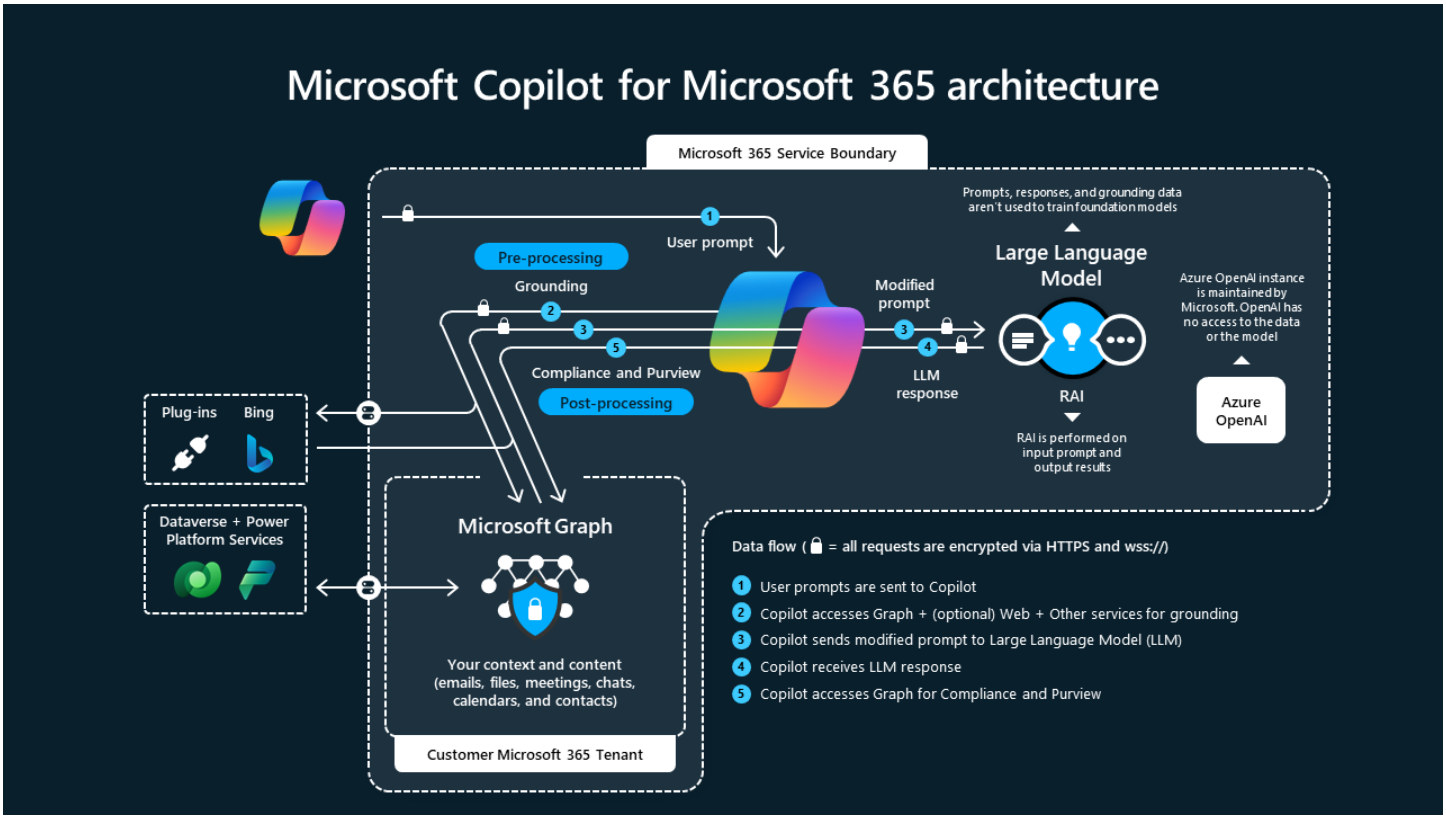
By reading this document, organizations can gain a better understanding of the potential benefits and challenges of using Copilot for Microsoft 365, as well as the best practices and safeguards that Microsoft has implemented to enable responsible and trustworthy AI. The document can also serve as a basis for further dialogue and collaboration between organizations and Microsoft to address any specific concerns or needs related to Copilot for Microsoft 365.

Refer to the Microsoft Responsible AI Transparency Report 2024 for a holistic description of the Microsoft Responsible AI Program published to date.

How Copilot for Microsoft 365 Works

For an understanding of how Copilot for Microsoft 365 works, refer to Microsoft public documentation.

- The user enters an input prompt into Copilot for Microsoft 365.
- Copilot then pre-processes the input prompt through an approach called grounding, which improves the specificity of the prompt, to help users get answers that are relevant and actionable to their specific task. The input prompt can include an explicit reference to a Microsoft 365 entity like a person or a file. Copilot can also fetch Microsoft 365 content and/or web content, depending on admin/user settings, that may provide helpful context to support the user's request, e.g., "summarize my inbox" would fetch a user's recent emails. The prompt can also include text from input files or other content discovered by Copilot for Microsoft 365 and sends this prompt to the LLM for processing. Copilot for Microsoft 365 only accesses data that an individual user has existing access to, based on, for example, existing Microsoft 365 role-based access controls.
- Copilot for Microsoft 365 takes the response from the LLM and post-processes it. This post-processing includes other grounding calls to Microsoft Graph, responsible AI checks, security, compliance and privacy reviews, and command generation.
- Copilot for Microsoft 365 returns the response to the app, where the user can review and assess the response.

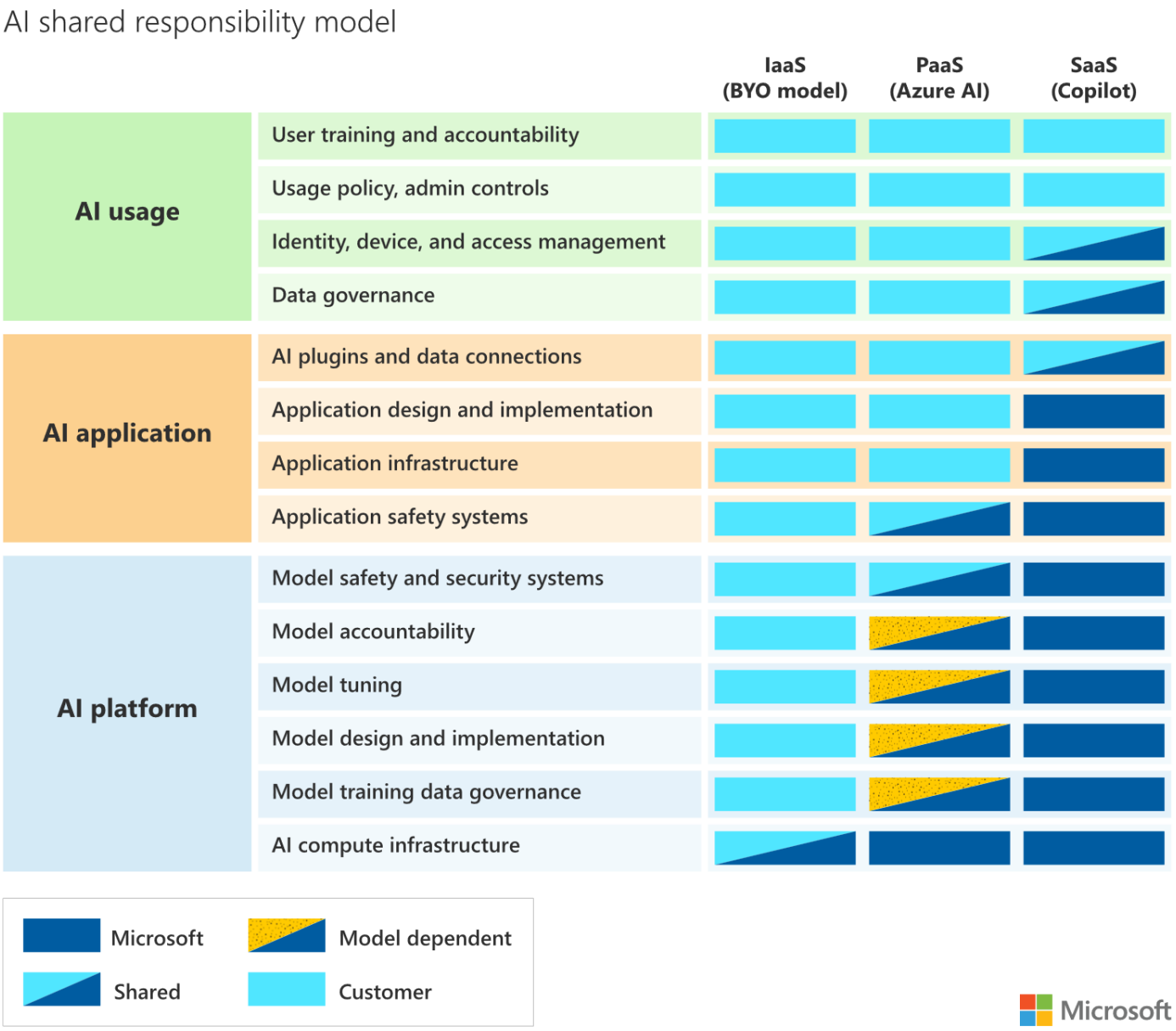


Mitigation Is a Shared Responsibility

Mitigation of AI risks is a shared responsibility between Microsoft and customers, as defined in the AI Shared Responsibility Model. Many risks can be mitigated by appropriate and responsible use, which is why Microsoft explicitly requires customers to comply with the Acceptable Use Policy within the Product Terms, as applicable, or the Azure OpenAI Code of Conduct. Microsoft also encourages customers to train their users in understanding the limitations and fallibility of AI.

As with cloud services, you have options when implementing AI capabilities for your organization. Depending on which option you choose, you take responsibility for different parts of the necessary operations and policies needed to use AI safely.

The following diagram illustrates the areas of responsibility between you and Microsoft according to the type of deployment. As you're performing your risk assessment, keep the shared responsibility model in mind as you identify risks to be mitigated by Microsoft and by your organization. Copilot for Microsoft 365 is a Software as a Service (SaaS) solution.



Copilot for Microsoft 365 Artificial Intelligence Risks & Mitigations Framework

The following are AI risks and their corresponding mitigations in the context of Copilot for Microsoft 365. Many of these mitigations are applicable beyond the Copilot for Microsoft 365 service to AI in the broader context of the Microsoft cloud.

Risk	Mitigation
Bias - AI technologies could unintentionally treat people unfairly or reinforce existing societal biases.	Mitigating unfairness starts with understanding the implications and limitations of AI predictions and recommendations. To follow this principle, Microsoft takes the following steps: (1) Microsoft AI systems are designed to provide a similar quality of service for identified demographic groups,

	including marginalized groups; (2) Microsoft AI systems that allocate resources or opportunities in essential domains are designed to do so in a manner that minimizes disparities in outcomes for identified demographic groups, including marginalized groups; and (3) Microsoft AI systems that describe, depict, or otherwise represent people, cultures, or society are designed to minimize the potential for stereotyping, demeaning, or erasing identified demographic groups, including marginalized groups. Copilot for Microsoft 365 service teams have measurements in place to account for and mitigate responsible AI harms such bias or unfairness. While mitigations may be scenario specific, teams have certain quantitative and qualitative measurements to mitigate responsible AI harms. Copilot for Microsoft 365 uses foundation models (i.e., OpenAI models), which implement mitigations against bias in their training as described in the GPT-4 System Card.
Disinformation - Disinformation is false information deliberately spread to deceive people.	Copilot for Microsoft 365 grounds its responses in customer data to help mitigate the risk of hallucination, which mitigates disinformation. User-in-the-loop is a concept that you will find across the board in Microsoft AI solutions, meaning that there is an end user involved in the actions Copilot for Microsoft 365 performs. Copilot for Microsoft 365 is designed to require explicit end user instruction. Copilot for Microsoft 365 cannot spread disinformation on behalf of a user on its own; users must take explicit action to spread disinformation manually, even if they evoke inaccurate or misleading information from Copilot for Microsoft 365 using tailored prompts.
Overreliance & automation bias – Automation bias is a human tendency to over-rely on information produced by automated artificial intelligence systems. AI systems may be used to support decision making, but there is a risk that any AI system generates incorrect or inaccurate information. This might affect pivotal decisions and could lead to the inadvertent spread of misinformation. The Microsoft Artificial Generative Intelligence	To mitigate the risk of overreliance and automation bias, Microsoft takes the following steps: (1) Microsoft AI systems are designed to inform people that they are interacting with an AI system; (2) Microsoft provides disclaimers in products conveying that AI generated content may be incorrect; (3) Microsoft AI systems that inform decision making by or about people are designed to support stakeholder needs for intelligibility of system behavior, which means we design the system so users can interpret the system effectively to make decisions and can protect themselves from "automation bias," which is over relying on outputs produced by the system; and (4) Microsoft provides information about the capabilities and limitations of our AI systems to support stakeholders in making informed choices about those systems. Copilot for Microsoft 365 interactions are clearly labeled to inform the user that they are interacting with an AI system such that the user can make an informed decision to use the provided information. These disclaimers may come in many forms such as "AI-generated content may be incorrect" or "Copilot uses AI. Check for mistakes." While the disclaimers may come in

Security team notes, "Overreliance may be naive (based in misunderstanding AI capabilities), rushed (failing to check because of time or overload), forced (situations where the user physically cannot validate responses) or motivated (users using the AI to provide justification for something they wanted to do). For Copilot for Microsoft 365, the primary concerns are naive and rushed overreliance."	different forms, the principle is the same: the user should check the accuracy of the information. Microsoft also encourages customers to train their users in understanding the limitations and fallibility of AI. Guidance for customers to reference: Building a foundation for AI success: Organization and culture Empowering responsible AI practices | Microsoft AI
Ungroundedness – also known as fabrication or hallucination, is another challenge faced by AI models. Ungroundedness occurs when the AI model generates information that is not based on provided input or pre-existing data, essentially inventing or hallucinating information.	Microsoft addresses ungroundedness with a defense-in-depth approach, including: • **Performance and effectiveness measures**: Microsoft measures various aspects such as response quality, groundedness, and link hallucination to ensure the effectiveness of Copilot for Microsoft 365. • **Metaprompt Engineering**: Microsoft provides Copilot for Microsoft 365 additional instructions beyond the user prompt to ensure Copilot responds in a manner compliant with the Microsoft Responsible AI Standard. • **Information governance**: Microsoft recommends implementing information governance best practices and using Microsoft information governance and security controls to protect and ground in sensitive data. • **Abuse monitoring and content filtering**: Copilot for Microsoft 365 employs abuse monitoring and content filtering to protect against unwanted output, including hallucination and prompt injections. • **Retrieval augmentation generation (RAG):** This technique is used to ensure the accuracy of responses by grounding them to enterprise data. • **Continuous improvement**: Microsoft is committed to continuously improving Copilot for Microsoft 365 by replacing models following a rigorous process that includes benchmarking and performance analysis. • **User oversight**: Microsoft encourages end user reviews of prompt and response outliers to minimize negative consequences from inaccurate responses. All Copilot for Microsoft 365 components are specifically engineered and tested to maximize groundedness whenever they operate in "grounded

	circumstances," i.e. are expected to produce outputs that are based in some ground truth as opposed to purely creative outputs.
Privacy – Data is a quintessential component for the operation of an AI system and lack of protections can leave this data vulnerable. Customers need to be assured that their data is secure while using an AI system.	At Microsoft, we have a long-standing practice of protecting our customers' information. Our approach to responsible AI is built on a foundation of privacy, and we remain dedicated to upholding core values of privacy, security, and safety in all our generative AI products and solutions as described in Julie Brill's blog (Microsoft Chief Privacy Officer). Microsoft customers can obtain assurance that the privacy commitments they rely on for our enterprise cloud products also apply to our enterprise generative AI solutions, including Copilot for Microsoft 365. • **We will keep your organization's data private.** Your data remains private and is governed by our applicable privacy and contractual commitments, including the commitments we make in the Microsoft's Data Protection Addendum, Microsoft's Product Terms, and the Microsoft Privacy Statement. • **You are in control of your organization's data.** Your data is not used and processed in undisclosed ways or without your permission in accordance with the Microsoft Data Protection Addendum. • **Your access control and enterprise policies are maintained.** To protect privacy within your organization when using enterprise products with generative AI capabilities, your existing permissions and access controls will continue to apply to ensure that your organization's data is displayed only to those users to whom you have given appropriate permissions. • **Your organization's data is not shared.** Microsoft does not provide access to your data to third parties without your permission and in accordance with the terms of the Microsoft Data Protection Addendum • **Your organization's data privacy and security are protected by design.** Security and privacy are incorporated through all phases of design and implementation of Copilot for Microsoft 365. As with all our products, we provide a strong privacy and security baseline and make available additional protections that you can choose to enable. As external threats evolve, we will continue to advance our solutions and offerings to enable world-class privacy and security, and we will continue to be transparent about our approach. • **Your organization's data is not used to train foundation models**. Copilot for Microsoft 365 services do not use your organization's data to train foundation models without your permission. Your data is not available to OpenAI or used to train OpenAI models. • **Our products and solutions continue to comply with global data protection regulations**. The Microsoft AI products and solutions you deploy continue to be compliant with today's global data protection and privacy regulations. As we continue to navigate the future of AI together, including the implementation of the EU AI Act and other laws globally, organizations can be certain that Microsoft will be transparent

	about our privacy, safety, and security practices. We will seek to comply with laws globally that govern AI, and back up our promises with clear contractual commitments. Please review the privacy and security controls for Copilot for Microsoft 365 here: Data, Privacy, and Security for Microsoft Copilot for Microsoft 365 | Microsoft Learn
Model training on personal data – Use of personal data in training large language models poses a risk of that data being disclosed to third parties.	Copilot for Microsoft 365 does not use customer data to train foundation models, as indicated in our commitments. See, for reference, Microsoft Product Terms. Copilot for Microsoft 365 uses OpenAI provided foundation models. The foundation models used by Copilot for Microsoft 365 do not learn dynamically based on usage. During development of these foundation models, OpenAI maintains high standards for data anonymization prior to training models. OpenAI continues to improve these systems iteratively and maintain quality control through human training data review to remove false negatives. Refer to the Sample Risk Assessment: Questions and Answers section for direct responses from OpenAI.
Resiliency - Organizations heavily reliant on AI may face operational disruptions if the AI system fails, malfunctions, or encounter unexpected issues.	Resiliency is not a novel risk related to AI; Microsoft maintains mitigations against service disruption to ensure resiliency of Copilot for Microsoft 365 alongside our other services. Microsoft online services achieve service resilience through redundant architecture, data replication, and automated integrity checking. Redundant architecture involves deploying multiple instances of a service on geographically and physically separate hardware, providing increased fault-tolerance for Microsoft online services. Data replication ensures there are always multiple copies of customer data in different fault-zones, allowing critical customer data to be recovered if corrupted, lost, or even accidentally deleted by the customer. Automated integrity checking increases data availability by automatically restoring data impacted by many kinds of physical or logical corruption. Refer to Microsoft official documentation for more details, including Data Resiliency in Microsoft 365. Moreover, Microsoft has mitigations in place for security threats that may impact availability of our systems. As an example, Microsoft mitigates the risk of distributed denial of service (DDoS) attacks using a mechanism called OneDDoS. From the Office 365 FedRAMP SSP, control SC-5, Denial of Service Protection, "Azure implements OneDDOS for the Office 365 service teams as defense against single point and distributed network flooding denial of service attacks. Individual service teams request OneDDOS deployment from Azure."

	See also, the recent third party penetration test report for Copilot for Microsoft 365 in which OWASP Top 10 vulnerabilities for LLM applications were assessed, including denial of service. Microsoft makes credit-backed uptime commitments in our Service Level Agreements.
Data leakage - shared resources of a cloud environment imposes a risk of sharing of information to unauthorized or unintended parties. Potential risks are data leakage outside the customer tenant boundary, within the tenant boundary between users and groups, or through the integration of third-party tools via plugins.	The permissions model within a Microsoft 365 tenant helps to enable that controls are applied so that data will not unintentionally leak between users, groups, and tenants. Copilot for Microsoft 365 presents only data that an individual can access using the same underlying controls for data access used in other Microsoft 365 services. AI technologies leverage the user identity-based access boundary so that the grounding process only accesses content that the current user is authorized to access. Multiple forms of protection have been implemented throughout Microsoft 365 to prevent customers from compromising Microsoft 365 services or applications or gaining unauthorized access to the information of other tenants or the Microsoft 365 system itself, including: • Logical isolation of customer content within each tenant for Microsoft 365 services is achieved through Entra ID (formerly known as Azure Active Directory) authorization and role-based access control. • Microsoft uses rigorous physical security, background screening, and a multi-layered encryption strategy to protect the confidentiality and integrity of customer data. • Microsoft 365 uses service-side technologies that encrypt customer data at rest and in transit, including BitLocker, per-file encryption, Transport Layer Security (“TLS”) and Internet Protocol Security (“IPsec”). Moreover, Microsoft also provides administrators with the ability to control integration with any third-party through the use of plugins. Please visit the following documentation that provides information on these controls: Manage access to web content in Microsoft Copilot for Microsoft 365 responses and Manage Apps with Plugins for Copilot in Integrated Apps - Microsoft 365 admin. As data governance is a shared responsibility, customers should also have appropriate governance controls (e.g., access management) in their environment to prevent data leakage. Here are some helpful resources to get customers started: Securing data in an AI-first world with Microsoft Purview Learn about Microsoft Purview Microsoft Syntex - SharePoint Advanced Management overview Delete your Copilot for Microsoft 365 interaction history Learn about retention for Microsoft Copilot for Microsoft 365 Customer Lockbox requests Overview of Customer Key

Security vulnerabilities – Vulnerabilities in AI service development can compromise security.	From Brad Smith's article, Microsoft has taken measures to support new voluntary commitments crafted by the Biden-Harris administration to help ensure that advanced AI systems are safe, secure, and trustworthy. By endorsing all of the voluntary commitments presented by President Biden and independently committing to several others that support these critical goals, Microsoft has expanded its safe and responsible AI practices, working alongside other industry leaders. Microsoft's commitments to protect, detect, and respond are directly related to protecting OpenAI assets within the Microsoft environment. ## Security Development Lifecycle Practices All Copilots, including Copilot for Microsoft 365, and Azure products/services (including the Azure OpenAI service and the Azure Machine Learning Platform in which OpenAI models reside) follow the Security Development Lifecycle (SDL) practices through all phases of the development process, to ensure a baseline level of security assurance. SDL Practices are summarized below (full detail: Microsoft SDL Practices): 1. **Provide training** – Enable developers, service engineers, and program and product managers understand security basics and know how to build security into software and services to make products more secure while still addressing business needs and delivering user value. 2. **Define security requirements** – The need to consider security and privacy is a fundamental aspect of developing highly secure applications and systems and regardless of development methodology being used, security requirements must be continually updated to reflect changes in required functionality and changes to the threat landscape. 3. **Define metrics and compliance reporting** – Define the minimum acceptable levels of security quality and to hold engineering teams accountable to meeting that criteria. 4. **Perform threat modeling** – Applying a structured approach to threat scenarios helps a team more effectively and less expensively identify security vulnerabilities, determine risks from those threats, and then make security feature selections and establish appropriate mitigations. 5. **Establish design requirements** – Apply security features (e.g. cryptography, authentication, logging, etc.) consistently and with a consistent understanding of the protection they provide. 6. **Define and use cryptography standards** – It's critically important that all data, including security-sensitive information and management and control data, is protected from unintended disclosure or alteration when it's being transmitted or stored, via encryption.

7. **Manage the security risk of using third-party components** – Having an accurate inventory of third-party components and a plan to respond when new vulnerabilities are discovered will go a long way toward mitigating this risk, but additional validation should be considered, depending on your organization's risk appetite, the type of component used, and potential impact of a security vulnerability.
8. **Use approved tools** – Define and publish a list of approved tools and their associated security checks, such as compiler/linker options and warnings. Engineers should strive to use the latest version of approved tools, such as compiler versions, and to take advantage of new security analysis functionality and protections.
9. **Perform static analysis security testing (SAST)** – Analyzing the source code prior to compilation provides a highly scalable method of security code review and helps ensure that secure coding policies are being followed. SAST is typically integrated into the commit pipeline to identify vulnerabilities each time the software is built or packaged.
10. **Perform dynamic analysis security testing (DAST)** – Performing run-time verification of your fully compiled or packaged software checks functionality that is only apparent when all components are integrated and running. This is typically achieved using a tool or suite of prebuilt attacks or tools that specifically monitor application behavior for memory corruption, user privilege issues, and other critical security problems.
11. **Perform penetration testing –** Penetration tests are often performed in conjunction with automated and manual code reviews to provide a greater level of analysis than would ordinarily be possible.
12. **Establish a standard incident response process -** Preparing an Incident Response Plan is crucial for helping to address new threats that can emerge over time. It should be created in coordination with your organization's dedicated Product Security Incident Response Team (PSIRT).

Updating the Security Development Lifecycle to address AI risk

Microsoft plans to make regular updates to the Security Development Lifecycle (SDL) to keep it aligned with threats as they evolve. The MSRC SDL bug bar has been and will be updated as new risks emerge to better account for AI risk, including generative AI.

We are committed to protecting our customers by providing security updates and guidance that address vulnerabilities when they are reported. Below are some of the vulnerability types found in systems involving AI that

Microsoft seeks to address with the continuously improving Security Development Lifecycle. These AI vulnerabilities have been rated in severity in accordance with the Microsoft Security Response Center advisory rating system and are published publicly with severity ratings. This severity classification helps Microsoft triage and prioritize AI risks identified within the system. Refer to the MSRC Updating our Vulnerability Severity Classification for AI Systems blog post for more guidance.

Pre-release security evaluations and AI red teaming

Prior to release of any update to Copilot for Microsoft 365 or the foundation models on which it depends, the software is evaluated for a range of AI safety and security concerns. Beyond standard security concerns, these evaluations include AI-specific security risks such as prompt injection. Evaluation methods include batteries of standardized measurements and tests as well as detailed testing by AI red teams, which probe for novel risks not yet captured by any standard tests by simulating adversarial behavior.

AI red teaming generally takes place at two levels: at the base model level (e.g., GPT-4) or at the application or service level (e.g., Copilot for Microsoft 365). Both levels bring their own advantages: for instance, red teaming the model helps to identify early in the process how models can be misused, to scope capabilities of the model, and to understand the model's limitations. These insights can be fed into the model development process to improve future model versions but also get a jump-start on which applications it is most suited for. Application-level AI red teaming takes a system view, of which the base model is one part.

When OpenAI provides a new model version to Microsoft, Microsoft's AI Red Team independently reviews it and assesses its performance across all relevant dimensions, including quality and responsible AI, as it performs in the Microsoft production environment with our own serving and security layers on top, before enabling it in the LLM API. Traditional red teaming also occurs to mitigate the risk of model theft and compromise of the service and associated infrastructure.

Red Teaming

Traditional red teaming of the Copilot for Microsoft 365 service
Application-level AI red teaming takes a system view, of which the base model is one part. This helps to identify failures beyond just the model, by including the application specific mitigations and safety system. Red teaming throughout AI product development can surface previously unknown risks, confirm whether potential risks materialize in an application, and inform measurement and risk management. The practice also helps clarify the scope of an AI application's capabilities and limitations, identify potential for misuse, and surface areas to investigate further.

AI red teaming of foundation models
The Azure security red team performed multiple assessments and shared the outcome with key stakeholders to ensure mitigation of found vulnerabilities. These processes are standard best practices applied by Microsoft. Red teaming assessments are conducted periodically and dynamically to identify vulnerabilities and ensure remediation to protect the integrity and confidentiality of OpenAI's and Microsoft's AI IP assets. Traditional red teaming mitigates the risk of model theft and impact to service code confidentiality and integrity.

Code Scanning OpenAI Code

Code Security Vulnerability Management (VM) is a critical component of the OpenAI Security Initiative. It involves implementing cyclical vulnerability identification, assessment, prioritization, and mitigation processes when OpenAI source code is initially ingested into Microsoft. The following guidelines are established to maintain effective security management of OpenAI source code:

Inventory and asset management: Maintain an up-to-date inventory of all OpenAI source code ingested into the Microsoft network, including versions and code sensitivity status. This step facilitates the identification of all potential source code versions containing vulnerabilities that are subject to attacks.

Vulnerability scanning: Initiate secure code scans upon ingestion of OpenAI source code using the CodeQL vulnerability scanning tools. This tool can identify known vulnerabilities and provide insight into potential weaknesses that could be exploited.

Vulnerability assessment and analysis: Collaborate with OpenAI to conduct vulnerability scan results assessments to identify and prioritize vulnerabilities based on the severity of their impact and the likelihood of exploitation. Utilize threat intelligence and vulnerability databases to stay informed about emerging threats and vulnerabilities.

Patch management: Collaborate with the OpenAI Developer team to implement a proactive patch management process, establish vulnerability resolution service level agreements, and an agreed upon reporting cadence. This ensures the timely implementation of code fixes, reduces the risk of unpatched vulnerabilities being exploited, and ensures Microsoft leadership receives vulnerability resolution updates, and updates from the OpenAI team.

Risk assessment: Leverage static code analysis results in conducting risk assessments to understand the potential impact of identified vulnerabilities on OpenAI's intellectual property.

	## Third-Party Vulnerability Assessment of Copilot for Microsoft 365 Microsoft engaged a third-party assessor to perform penetration testing of nine Copilot implementations across the M365 product suite with a focus on identifying LLM-related vulnerabilities, as well as traditional security vulnerabilities in supporting application infrastructure. This report provides information and insights into the findings of that assessment, which was performed between January and May of 2024. The report of the assessment can be found here: Service Trust Portal (microsoft.com).
Model evaluation – Generative AI Models are subject to fallibility, bias, and other harmful behavior.	Microsoft implements a layered approach to addressing AI risks: governing, mapping, measuring, and managing risks of harm and misuse as AI is developed and deployed across the technology architecture, including at the model, API service, and application layers. As products evolve or we learn more, we also continue to invest throughout the product lifecycle, which includes the evaluation of the entire stack, including the internal code of Copilot for Microsoft 365, metaprompts, and models that are used by the service. After mapping risks, we use systematic measurement to evaluate application and mitigation performance against defined metrics. For example, we can measure the likelihood of our applications to generate identified content risks, the prevalence of those risks, and the efficacy of our mitigations in preventing those risks. We regularly broaden our measurement capabilities. Examples of established metrics include: • Groundedness, to measure how well an application's generated answers align with information from input sources. • Relevance, to measure how directly pertinent a generated answer is to input prompts. • Similarity, to measure the equivalence between information from input sources and a sentence generated by an application • Content risks, multiple metrics through which we measure an application's likelihood to produce hateful and unfair, violent, sexual, and self-harm related content • Jailbreak success rate, to measure an application's resiliency against direct and indirect prompt injection attacks that may lead to jailbreaks. Refer to the 2024 Responsible AI Transparency Report for more information on this measurement process. It is important to note that the AI model red teaming process, described above in this document, is a key component of model evaluation.

Intellectual property – Publishing content that has been generated by AI that may infringe upon third party intellectual property rights.	In accordance with the Product Terms, Microsoft offers to defend customers from IP infringement claims arising from the customer's use and distribution of the output content generated by Microsoft's Copilot services, including Copilot for Microsoft 365. Specifically, should a third party sue a commercial customer for copyright infringement for using a Microsoft Copilot service or the output they generate, we will defend the customer and pay the amount of any adverse judgements or settlements that result from the lawsuit, as long as the customer used the guardrails and content filters we have built into our products. To learn more, please refer to Introducing the Microsoft Copilot Copyright Commitment.
Explainability - Since AI relies on extrapolated logic rather than hard-coded rules, it can sometimes function as a black box, where users do not understand how the system's outputs were derived from its inputs.	The Microsoft Responsible AI Standard outlines goals for responsible AI aligned to six principles: fairness, reliability & safety, privacy & security, inclusiveness, transparency, and accountability. From the Responsible AI Standard Transparency Goal T2.2: Publish documentation for the system so that stakeholders can understand the system. Include: 1) capabilities, 2) intended uses, 3) uses that require extra care or guidance, 4) operational factors and settings that allow for effective and responsible system use, 5) limitations, including uses for which the system was not designed or evaluated, and 6) evidence of system accuracy and performance as well as a description of the extent to which these results are generalizable across use cases that were not part of the evaluation. When the system is a platform service made available to external customers or partners, a Transparency Note is required. Please reference the following description on how Copilot for Microsoft 365 works: Microsoft Copilot for Microsoft 365 overview. Customers are also encouraged to take advantage of the audit logging capabilities offered to them. Please reference the following documentation: Audit log activities CopilotInteraction

Microsoft References

- Microsoft Responsible AI Standard v2 General Requirements
- How Microsoft 365 Delivers Trustworthy AI
- Data, Privacy, and Security for Microsoft Copilot for Microsoft 365
- Microsoft's AI Safety Policies
- Protecting the data of our commercial and public sector customers in the AI era
- Human review for automation with a prompt
- Microsoft AI Red Team building future of safer AI
- Introducing the Microsoft Copilot Copyright Commitment
- Staying ahead of threat actors in the age of AI
- How Microsoft discovers and mitigates evolving attacks against AI guardrails
- Announcing Microsoft's open automation framework to red team generative AI Systems

Sample Risk Assessment: Questions & Answers

This section presents a sample set of questions and answers that can be used to assess the features, functionalities, and the surrounding security and compliance posture of Copilot for Microsoft 365, as well as the broader implications and impacts of using the service. The questions are derived from customer inquiries and the answers are based on information from various Microsoft internal teams and sources, including engineering, legal, compliance, privacy, trust, customer experience, the Office of the CTO, Azure Security, Azure OpenAI, and many more teams. Some responses also include direct attestation from OpenAI, the Microsoft strategic partner from which we source our Copilot for Microsoft 365 foundation models.

The risk assessment questions are grouped by ID with the key as follows:

P = Privacy

S = Security

SR = Supplier Relationship

MD = Model Developer

From OpenAI = Official statement obtained directly from OpenAI

ID	Question	Microsoft Response
P1	Are the AI models trained on personal data? If not, then how does the AI solution provider ensure that the model is not being trained with personal data e.g. anonymization techniques?	**From OpenAI:** OpenAI maintains high standards for data anonymization prior to training models. OpenAI continues to improve these systems iteratively and maintain quality control through human data review to remove false negatives. OpenAI employs extensive PII scrubbing, PII filtering, and/or human review on all training data. Refer to OpenAI for additional information.
P2	If the AI models are trained on personal data, what measures are in place to protect the privacy of individuals whose data is used to train generative AI?	Refer to the GPT-4 system card, published by OpenAI, for how OpenAI models are trained. **From OpenAI:** OpenAI maintains high standards for data anonymization prior to training models. OpenAI continues to improve these systems iteratively and maintain quality control through human data review to remove false negatives.
P3	What types of personal data does the AI solution collect?	The Microsoft Data Protection Addendum (DPA) outlines the processing of personal data in section, "Processing of Personal Data: GDPR". The DPA covers data processing for all data across our enterprise services offerings. The offerings in scope of the DPA are covered in the Product Terms and include Copilot for Microsoft 365.
P4	Do existing privacy commitments extend to your AI solutions?	Yes, refer to the Julie Brill's (Microsoft Chief Privacy Officer) blog post Protecting the data of our commercial and public sector customers in the AI era which states, "Microsoft's privacy commitments apply to AI." See also, "FAQ: Protecting the Data

		of our Commercial and Public Sector Customers in the AI Era" linked from this blog.
P5	Does the AI model perform sentiment analysis in the workplace, which is banned by the EU AI Act?	We are committed to compliance with the EU AI Act. Our multi-year effort to define, evolve, and implement our Responsible AI Standard and internal governance has strengthened our readiness. At Microsoft, we recognize the importance of regulatory compliance as a cornerstone of trust and reliability in AI technologies. We are committed to creating responsible AI by design. Our goal is to develop and deploy AI that will have a beneficial impact on and earn trust from society. Our work is guided by a core set of principles: fairness, reliability and safety, privacy and security, inclusiveness, transparency, and accountability. Microsoft's Responsible AI Standard takes these 6 principles and breaks them down into goals and requirements for the AI we make available. Our Responsible AI Standard takes into account regulatory proposals and their evolution, including the initial proposal for the EU AI Act. We developed our most recent products and services in the AI space such as Microsoft Copilot and Microsoft Azure OpenAI Service in alignment with our Responsible AI Standard. As final requirements under the EU AI Act are defined in more detail, we look forward to working with policymakers to ensure feasible implementation and application of the rules, to demonstrating our compliance, and to engaging with our customers and other stakeholders to support compliance across the ecosystem.
P6	If the AI models **are** trained on personal data, what measures are in place to protect the privacy of individuals whose data is used to train generative AI?	Refer to the GPT-4 system card, published by OpenAI, for how OpenAI models are trained. **From OpenAI:** OpenAI maintains high standards for data anonymization prior to training models. We continue to improve these systems iteratively and maintain quality control through human data review to remove false negatives.
P7	If the AI models **are not** trained on personal data, which anonymization techniques are used to ensure the models are not trained on personal data e.g. differential privacy?	Refer to the GPT-4 system card, published by OpenAI, for how OpenAI models are trained. **From OpenAI:** OpenAI maintains high standards for data anonymization prior to training models. We continue to improve these systems iteratively and maintain quality control through human data review to remove false negatives. OpenAI employs extensive PII scrubbing, PII filtering, and/or human review on all training data. Refer to OpenAI for more information.

P8	Did the AI/ML provider or their third party (the entity training the models in question) have the necessary lawful basis to use personal data to train AI models for each particular purpose of use?	Refer to the GPT-4 system card, published by OpenAI, for how OpenAI models are trained.
P9	Are AI foundation models trained on data that was collected in compliance with GDPR?	Microsoft does not train Copilot for Microsoft 365 foundation models. Please refer to the Microsoft Product Terms, which states "Microsoft Generative AI Services do not use Input or Output Content to train, retrain, or improve Azure OpenAI Service foundation models." Moreover, "Microsoft's AI products and solutions are compliant with applicable data protection and privacy laws today." Refer to the FAQ: Protecting the Data of our Commercial and Public Sector Customers in the AI Era.
P10	How are AI-based biometrics being used to identify individuals in the AI/ML solutions?	N/A. Copilot for Microsoft 365 leverages customer managed user identification systems to validate authentication and authorization to information.
P11	Does your organization have a privacy impact assessment (PIA) for the service(s) in question?	Microsoft performs privacy reviews on product features including Copilot for Microsoft 365 features.
P12	What data elements associated with the service are captured by the AI/ML provider in the form of telemetry or diagnostic data?	Data collected for Copilot for Microsoft 365 meets Microsoft 365 data collection and data handling standards. Microsoft publishes public documentation describing diagnostic data elements captured by the service: Required diagnostic data for Office and Optional diagnostic data for Office .
P13	Can customers opt out of AI telemetry or diagnostic data collection?	Microsoft provides controls for managing diagnostic data and connected experiences as described in public documentation Use policy settings to manage privacy controls for Microsoft 365 Apps for enterprise.
P14	Are there categories of telemetry or diagnostic data that are required for the service to function?	Yes, please see Required diagnostic data for Office - Deploy Office | Microsoft Learn.
P15	Are there specific disclaimers within the AI/ML service that users or admins see when they are operating the service? Please describe them.	Copilot for Microsoft 365 interactions are clearly labeled within Microsoft 365 apps. Microsoft teams follow the Microsoft responsible AI standard and one of the goals of the standard is that Microsoft AI systems are designed to inform people that they are interacting with an AI system.

		Reference: Responsible AI Standard v2 Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf
P16	How is the AI model output (generated content) screened or filtered to ensure it does not contain sensitive personal data?	**From OpenAI:** OpenAI maintains high standards for data anonymization prior to training models. OpenAI continues to improve these systems iteratively and maintain quality control through human data review to remove false negatives. OpenAI employs extensive PII scrubbing, PII filtering, and/or human review on all training data. Refer to OpenAI for additional information.
P17	Is there a mechanism to allow individual users to opt-out of having their prompts and AI generated output used in generative AI training?	*Microsoft does not use customer content to train foundational models*. Refer to the Julie Brill's blog post Protecting the data of our commercial and public sector customers in the AI era which states, "Microsoft's privacy commitments apply to AI." See also, FAQ: Protecting the Data of our Commercial and Public Sector Customers in the AI Era and Data, Privacy, and Security for Microsoft Copilot for Microsoft 365.
P18	What steps are taken to prevent reidentification of individuals based on content generated by AI (model output)?	**From OpenAI:** OpenAI maintains high standards for data anonymization prior to training models. We continue to improve these systems iteratively and maintain quality control through human data review to remove false negatives. OpenAI employs extensive PII scrubbing, PII filtering, and/or human review on all training data. Refer to OpenAI for additional information.
P19	Are user generated prompts (input) stored or retained on the backend even in cases where chat history is disabled? If so, what precautions are taken to protect the user's privacy?	Please refer to our public documentation: Data, Privacy, and Security for Microsoft Copilot for Microsoft 365. This document describes how user generated prompts are retained, and also describes controls that admins and users can use in relation to these user generated prompts.
P20	Are there mechanisms to address the potential risks of unauthorized reproduction or impersonation through generative AI?	All responses are generated within the user context, so this is not applicable.
P21	Is there transparency regarding the data sources used to train AI models and does your organization, in attempt to avoid privacy violations, restrict using publicly available data? Refer to the class action lawsuit against OpenAI regarding how it used people's data to train the algorithm.	**From OpenAI:** OpenAI utilizes public internet data; proprietary and licensed data; and various anonymized user data for users that have opted into training.
P22	Are there instances where the AI/ML provider is the controller of the data processed by the AI/ML solution (as opposed to the role of processor which applies to majority of scenarios)?	From the Data Protection Addendum (DPA): "Customer and Microsoft agree that Customer is the controller of Personal Data and Microsoft is the processor of such data, except (a) when Customer acts as a processor of Personal Data, in which case Microsoft is a subprocessor; or (b) as stated otherwise in the

		Product-specific terms or this DPA." Refer to the Product Terms and the DPA for additional information relevant to the question.
S1	What is the propensity of your models for hallucination and bias, how are you protecting against this?	Microsoft performs extensive testing and AI model red teaming to mitigate the risk of hallucination. Furthermore, Microsoft measures model performance, including groundedness of responses, which mitigates the risk of hallucination.
S2	Have you instituted monitoring across the data and model lifecycle? If so, please comprehensively describe the metrics being monitored.	Please refer to controls in the SOC 2 report and the Office 365 FedRAMP SSP. **From the Office 365 FedRAMP SSP Control SI-4:** "Service teams have deployed Near Real Time monitoring solutions that generate all real-time alerts and audit logs from both SCOM and the repository. All the service team-specific monitoring requirements beyond the base set defined in Office 365 Security Auditing SOP for Office 365 are integrated into a SCOM pack. In addition, service teams upload their logs to a repository service, where they are aggregated and processed. The uploader service generates reports using automated security alerting tools. These automated security alerting tools assist in identifying normal usage of the system and deviations from that normal range. Additionally, they examine records to confirm that the system is functioning in an optimal, resilient, and secure state. Unusual activity is flagged and alerted in Near Real Time. The repository also aggregates logs for further review. Any log event that indicates a potential security violation must be immediately brought to the attention of Office 365 Security. Teams using Pilotfish use K9, a tool that collects security logs and performs real-time analysis, as well as archiving them to a repository service for forensic analysis. Unusual activity will generate alerts and the Pilotfish team coordinates with the appropriate service team for any required follow-ups or investigations. Service teams will work with the Office 365 SIR team to address all security incidents. Local connections are disallowed by policy within Office. No personnel have local access. Azure performs network monitoring and detection of unauthorized connections in accordance with their security policy. Monitoring is also inherited for services using inherited compute from Azure. Remote authentication failures are logged and stored within a repository service. For further information, please see AC-17 and AC-7. Audit logs are uploaded to repository and reports are generated using automated security alerting tools. These automated security alerting tools assist in identifying normal usage of the system and deviations from that normal range. The automated security alerting tools use heuristics to identify unauthorized use of the operating system. Unusual activity is flagged and alerted

		in near real time and a repository aggregates logs for further review. Any log event that indicates a potential security violation must be immediately brought to the attention of Office 365 Security. All servers act as monitoring devices and are configured to log all security-relevant events. Office 365 monitors and alerts in Near Real Time for all hosts in the environment. Suspicious events generate alarms and notifications to service team staff and appropriate contingent staff. Logs are aggregated in a repository service and reports are generated using automated security alerting tools. Office 365 Security notifies service teams if a change in the level of monitoring is necessary due to indications of increased risk, and service teams adjust monitoring accordingly. Servers are configured to increase logging parameters in response to an indication of increased risk, and the automated security alerting tools heuristics are tailored to look for specific threats, in conjunction with any alerts received via NRT Security Monitoring, based on the nature of the risk to organizational operations and assets. Office 365 Security, in consultation with Corporate External Legal Affairs (CELA), has defined a set of log events and alerts that meet federal regulatory requirements for incident management and investigation. This structure is intended to support identification of known suspicious activity and to support the investigation of misuse and abuse of Office 365 services. To fully comply with applicable regulations, the service teams follow defined requirements for event collection and notification processes. These requirements are contained in the Office 365 Security and Pilotfish/K-9 onboarding document. All servers upload logs to a repository service for aggregation and analysis. Reports are generated from this data using NRT Security Monitoring and the automated security alerting tools as described in AU-6 and AU-7. These reports are available daily and as needed."
S3	Please provide an architectural diagram and an end-to-end data flow diagram	Microsoft publishes a Copilot for Microsoft 365 architecture diagram: Microsoft Copilot for Microsoft 365 overview | Microsoft Learn.
S4	How do you manage access control to the model code during development and in production?	Please refer to controls in the Office SOC 2 report and FedRAMP SSP. Same way as we provide code control through standard DevOps tooling (i.e. Azure DevOps, etc.). Only authorized engineers have access to model assets and the source-code of the model, by identifying authorized personnel through ADO project memberships, and through corporate account management as

		described in the AC control family of the Office 365 FedRAMP SSP.
S5	Which AI/ML algorithms are in use e.g. generative AI + model name + model version?	OpenAI GPT-4 is the current foundation model in use within the Copilot for Microsoft 365 system as of May 15, 2024.
S6	Is the model designed to continuously learn? If so, how does it continuously learn e.g. reinforcement learning from human feedback (RLHF)?	Copilot for Microsoft 365 foundation models are not designed to continuously learn.
S6.1	If the model is designed for continuous learning, what is the data source from which the model learns?	Copilot for Microsoft 365 foundation models are not designed to continuously learn.
S6.2	If the model is designed for continuous learning, what is the frequency of learning input data fed to the model?	Copilot for Microsoft 365 foundation models are not designed to continuously learn.
S6.3	If the model is designed for continuous learning, what is the supervision level of the learning e.g. supervised, semi-supervised, unsupervised?	Copilot for Microsoft 365 foundation models are not designed to continuously learn.
S6.4	If the model is designed for continuous learning, what is the frequency at which the model is refreshed?	Copilot for Microsoft 365 foundation models are not designed to continuously learn.
S6.5	Which organizations played a role in model development? Do you have any sub processors? If there are sub processors, please explain their role in model design and operation.	OpenAI, which is a third-party strategic partner to Microsoft, developed the foundation model utilized by Copilot for Microsoft 365. Note: OpenAI is not a subprocessor.
S7	How is the model deployed securely into production?	Microsoft has a deployment system adhering to existing SOC 2 controls; Azure OpenAI is included in the SOC 2 scope. Please refer to the Azure SOC 2 report for further information.
S8	Please describe the classification of data elements flowing through the system e.g. personal data (as defined by GDPR), financial data (PCI DSS), confidential, public. Please share your enterprise data taxonomy to illustrate internal and external classifications.	Personal Data, Customer Content as defined in the Data Protection Addendum available at https://aka.ms/dpa.
S9	How are you monitoring for abnormal behavior and are you exposing these monitoring capabilities to customers? Does the model provide any direct output metadata such as confidence scores, citations, or other performance or explainability metrics?	Microsoft collects user feedback on responses, which can be used to identify anomalous behavior. Microsoft Copilot for Microsoft 365 also use Azure Content Safety Service or filtering similar in functionality to filter out hate, sex, violence, and self-harm prompts and responses. If an anomalous number of prompts are being blocked, Microsoft can investigate the Copilot for Microsoft 365 service to understand the root cause of the issue. Microsoft offers customers the capability of utilizing Azure Content Safety Service when building their own GenAI systems. Customers utilizing Copilot for Microsoft 365 can

		monitor prompts and responses using Microsoft Purview Communication Compliance, Content Search, or eDiscovery.
S10	With what privileges can the model execute actions? Does it use its own credentials or the user's?	The model only processes prompts and emits responses, it does not take actions outside of these defined, limited capabilities. Information fed into the model can only be accessed within the querying user's access scope. Copilot does not provide additional access beyond that to which the user has access based on their permissioning model.
S11	How do you prevent model compromise from cascading into a compromise of other systems or accounts e.g. compromise of the larger service in which the model is running?	Fault isolation and blast radius reduction within our Microsoft 365 services prevents the compromise of a single service or service component from affecting other services or service components.
S12	How do you sanitize and validate model inputs and outputs?	Microsoft Copilot for Microsoft 365 service teams perform ingress and egress filtering, mitigating the risk of bad prompts reaching the LLM and mitigating the risk of bad responses reaching the end user.
S13	If the model begins behaving anomalously, what is the process for performing root cause analysis?	In alignment with all features of the service, Microsoft would investigate erroneous behavior of the system per our standard incident response process. Examples of post incident reports can be found at: Azure status history | Microsoft Azure. Microsoft 365 provides public post-incident reports to customers. These reports are typically published within **72 hours** of resolving a cloud service incident and offer a detailed analysis of the incident, including its root cause, impact, and the steps taken to prevent similar incidents in the future. Customers can access these Post Incident Reports (PIRs) through the Service Health Dashboard in the Microsoft 365 admin center or the Azure portal. To view the PIRs, customers need to sign in with an admin account, navigate to the Service Health Dashboard, locate the incident, and then click on the incident to find the link to the PIR.
S14	How can model users (customers) validate the performance of the model with controls available to them?	Customers can perform eDiscovery against prompts and responses. Customers can also use Purview Communication Compliance to identify Copilot for Microsoft 365 abuse per customer-configured parameters.
S15	If the model was developed by a third party, describe the due diligence or supplier reviews performed and at what cadence they will be performed ongoing?	Microsoft does not share any information about the due diligence performed between Microsoft and OpenAI - this is a confidential matter between the involved parties.
S16	Describe how you mitigate the risk of data poisoning via the supply chain e.g. third-party training data and/or	**From OpenAI:** OpenAI captures numerous model backups throughout the training and deployment process and could revert to an earlier snapshot of the model or training data, as necessary, mitigating the risk of data poisoning.

	code modification prior to obtaining the model?	Microsoft mitigates the risk of model weights and other code modification by securely transferring the OpenAI model into the Microsoft environment. The model transfer happens between OpenAI and the Microsoft storage tenant. The model transfer occurs on the backend network over Azure via the cross-tenant object replication feature that is set between Microsoft and OpenAI storage accounts. The latest models are always transferred with encryption of the model weights, and the master encryption key (AES256) for each model drop are wrapped using industry standard asymmetric encryption key pairs, where the private key is stored in a Microsoft managed HSM (using the Azure managed HSM feature).
S17	Describe how you sanitize the training data to ensure it is not adversarial or biased.	**From OpenAI:** Safety filtering, copyright removal, and PII scrubbing.
S18	If you do train on personal data, do you anonymize the data prior to training? If so, please describe the anonymization process(es) e.g. differential privacy with X Epsilon value.	**From OpenAI:** OpenAI maintains high standards for data anonymization prior to training models. We continue to improve these systems iteratively and maintain quality control through human data review to remove false negatives. OpenAI employs extensive PII scrubbing, PII filtering, and/or human review on all training data. Refer to OpenAI for additional information.
S19	Please describe the model integrity validation mechanisms in place i.e. code integrity and supporting file integrity as applicable both during initial creation and during any updates to the model.	Microsoft does not further train the model once it enters the Microsoft environment. Model weights and code are protected in the same way other Microsoft 365 service code is protected. From Control SI-7 in Office 365 FedRAMP SSP, "Software and Information Integrity": "Strong configuration management controls and processes ensuring that only reviewed and approved changes are made to the system. • Standardized deployment processes that ensure that server roles are configured reliably and consistently across the service. There is no manual configuration and thus no risk of accidental deviations. • Office 365 frequently redeploys the service using these standardized deployment processes. This provides defense-in-depth against change controls being circumvented because manual changes will be quickly overwritten. • Integrity verification of core Windows files is implemented using System File Checker (SFC) and Windows Resource Protection (WRP). • Office 365 has the auditing capability to reliably identify who made a change. • Office 365 performs vulnerability scanning daily, which can detect vulnerable software and provide defense-in-depth if change controls are somehow circumvented. • Office 365 performs a manual reconciliation of its software inventory against software identified by scanning and discrepancies are investigated."

S20	Has the AI/ML provider updated their security development lifecycle (SDL) or established a secure data development lifecycle process (SDDLC) that addresses the AI specific threats and the AI specific data lifecycle stages identified in ENISA Artificial Intelligence Cybersecurity Challenges, ENISA Securing Machine Learning Algorithms (SMLA, page 13 figure 3)?	Refer to this public article confirming that the Microsoft Security Response Center updated the Microsoft bug bar to include and effectively prioritize AI risks by severity.
S21	Does the AI/ML provider ensure that effective security controls are implemented to protect models against model theft type of attacks including oracle and inference attacks? Such controls include but are not limited to: (i) data leakage detection and prevention controls, (ii) preventing direct exposure of the model and/or direct interactions with the model by users e.g. exposing the model directly in the user device/browser/untrusted/uncontrolled environment without DRM-like protection, (iii) limiting the number and type of interactions users can perform with the model, (iv) limiting the output of the model, (v) monitoring abnormal H2M/M2M interactions with the model. For example, such controls include but not limited to:	Refer to the last section of the following public documentation on the Service Trust Portal describing defense-in-depth controls mitigating the risk of model compromise and theft: https://aka.ms/trustworthyAI.
S21.1	i. Data leakage detection and prevention controls	**Monitoring Description**: Security logs from Office 365 are sent to a centralized repository for monitoring. Office 365 uses Vanquish as a Near Real Time monitoring tool, which leverages the Geneva infrastructure to provide logging and alerting upon detection of breaches or attempts to breach Office 365 platform trust boundaries. Vanquish collects security-relevant system information and some enrichment data, using these inputs to generate real-time alerts that service teams can use to correct vulnerabilities and improve any weaknesses found. Findings are reported and escalated using standard security incident management channels: ticketing tools are used for tracking, and the Office 365 Security Incident Response Team is engaged when appropriate. Geneva is an extensible collection of libraries, tools and services that enable services to do Monitoring, Diagnostics and Analytics at scale. **Control SI-4 Information System Monitoring (Office 365 FedRAMP SSP)**: Service teams have deployed Near Real Time monitoring solutions that generate all real-time alerts and audit

		logs from both SCOM and the repository. All the service team-specific monitoring requirements beyond the base set defined in Office 365 Security Auditing SOP for Office 365 are integrated into a SCOM pack. In addition, service teams upload their logs to a repository service, where they are aggregated and processed. The uploader service generates reports using automated security alerting tools. These automated security alerting tools assist in identifying normal usage of the system and deviations from that normal range. Additionally, they examine records to confirm that the system is functioning in an optimal, resilient, and secure state. Unusual activity is flagged and alerted in Near Real Time. The repository also aggregates logs for further review. Any log event that indicates a potential security violation must be immediately brought to the attention of Office 365 Security. Audit logs are uploaded to repository and reports are generated using automated security alerting tools. These automated security alerting tools assist in identifying normal usage of the system and deviations from that normal range. The automated security alerting tools use heuristics to identify unauthorized use of the operating system. Unusual activity is flagged and alerted in near real time and a repository aggregates logs for further review. Any log event that indicates a potential security violation must be immediately brought to the attention of Office 365 Security.
S21.2	ii. Preventing direct exposure of the model and/or direct interactions with the model by users e.g. exposing the model directly in the user device/browser/untrusted/uncontrolled environment without DRM-like protection.	Customers purchase Copilot for Microsoft 365 and they interact with the model through defined, restricted, front-end user interfaces. There is no direct access to the model. Direct access is prevented by the controls described in the Office 365 SOC 2 Report.
S21.3	iii. Limit the number and type of interactions users can perform with the model	Copilot integrations are distinct, and the user does not have direct input to the model. Prompts are processed/validated as part of the functioning of Copilot for Microsoft 365 services.
S21.4	iv. Limit the output of the model	Copilot for Microsoft 365 performs egress filtering to ensure harmful output does not reach end users.
S21.5	v. Monitor abnormal H2M/M2M interactions with the model	Please refer to monitoring controls in the Office SOC 2 report and Office 365 FedRAMP SSP.

S22	Does the AI/ML provider ensure that effective security controls are implemented to protect the model against resiliency and or availability attacks which target the model and/or may use the model for availability attacks against other systems (DoS/DDoS) specifically causing the model to be unresponsive (DoS), causing the model to make other internal systems unavailable (DoS), or causing the model to attack other customers or external systems e.g. amplification DDoS attacks? Does the AI/ML provider implement controls that would protect against attacks that would target the AI model causing:	**From Office 365 FedRAMP SSP Control SI-4(4)**: "Azure monitors for unusual traffic patterns using OneDDOS. In addition, service teams monitor for denial-of-service attacks by monitoring the following key health metrics: CPU usage, network connections, disk input/output operations per second (IOPS), and disk space usage. Office 365 service teams also monitor and review web server (e.g. IIS) logs and other application logs (as applicable) for unusual or unauthorized activities or conditions. Any unapproved connections detected through auditing or alerting will be triaged using security incident response processes."
S22.1	i. The model to be unresponsive (DoS)	From Office 365 FedRAMP SSP Control SI-4(4): "Azure monitors for unusual traffic patterns using OneDDOS. In addition, service teams monitor for denial-of-service attacks by monitoring the following key health metrics: CPU usage, network connections, disk input/output operations per second (IOPS), and disk space usage. Office 365 service teams also monitor and review web server (e.g. IIS) logs and other application logs (as applicable) for unusual or unauthorized activities or conditions. Any unapproved connections detected through auditing or alerting will be triaged using security incident response processes."
S22.2	ii. The model to make other internal systems unavailable (DoS)	From Office 365 FedRAMP SSP Control SI-4(4): "Azure monitors for unusual traffic patterns using OneDDOS. In addition, service teams monitor for denial-of-service attacks by monitoring the following key health metrics: CPU usage, network connections, disk input/output operations per second (IOPS), and disk space usage. Office 365 service teams also monitor and review web server (e.g. IIS) logs and other application logs (as applicable) for unusual or unauthorized activities or conditions. Any unapproved connections detected through auditing or alerting will be triaged using security incident response processes."
S22.3	iii. The model to attack other customers or external systems (e.g. amplification DDoS attacks)	Office 365 was designed using the principles of defense in depth. Cross tenant protections are implemented at the application layer to ensure that customers cannot compromise Office 365 applications to gain unauthorized access to the information of other tenants. Protections are also implemented at the network layer to prevent interception of network traffic and resource starvation attacks. Protections are additionally implemented at the operating system layer to prevent side channel attacks. Details regarding the protections implemented to prevent cross-tenant attacks are documented in Microsoft 365 isolation controls - Microsoft Service Assurance | Microsoft Docs

S23	When model is trained on customer data and/or uses customer data as input for training or inference, does the AI/ML provider implement effective controls to prevent and detect information leakage and information disclosure by the model? Example: model A is trained on customer's A data (explicitly and/or implicitly), customer B has buys from the AI/ML provider access to the model and queries the model for information, customer B receives customer A's information which is either stored within the model and/or can be accessed by the model.	Please refer to Product Terms, which states, "Microsoft Generative AI Services do not use Input or Output Content to train, retrain, or improve Azure OpenAI Service foundation models."
S24	Does the AI/ML provider log and monitor interactions (machine or human) with the model, including the inputs and outputs of the model? Is the information protected and isolated against unauthorized access by unauthorized parties, including malicious insiders or 3rd parties?	From the Data protection Addendum on Data Access: "Microsoft employs least privilege access mechanisms to control access to Customer Data and Professional Services Data (including any Personal Data therein). Role-based access controls are employed to ensure that access to Customer Data and Professional Services Data required for service operations is for an appropriate purpose and approved with management oversight. For Core Online Services and Professional Services, Microsoft maintains Access Control mechanisms described in the table entitled "Security Measures" in Appendix A; and there is no standing access by Microsoft personnel to Customer Data, and any required access is for a limited time." See also, Identity and access management overview - Microsoft Service Assurance | Microsoft Learn.
SR1	Has the AI/ML provider risk assessment for outsourcing and third-party suppliers been extended to include AI specific cybersecurity (industrial control systems - digital devices used in industrial processes) risks and controls extending, specifically those not covered by existing attestation frameworks such as SOC, ISO 27001, FedRAMP, etc. for example threats identified by: MITRE Atlas OWASP TOP 10 LLM ENISA Artificial Intelligence Cybersecurity Challenges ENISA Securing Machine Learning Algorithms	Refer to the OpenAI Trust Portal to review OpenAI SOC 2 report and the third-party external penetration testing report available at https://trust.openai.com/. OpenAI is a strategic partner and Microsoft does not disclose details about strategic partnerships or the associated reviews and agreements therein. OpenAI is not a supplier, they are a strategic partner, and their models are an acquired capability over which Microsoft has performed due diligence.

SR2	Does the AI/ML provider perform periodic and ad-hoc risk assessment for outsourcing and third-party suppliers which include evaluation of their suppliers / outsourcing party AI cybersecurity risks and cybersecurity maturity specifically existence of frameworks addressing AI cybersecurity risks, AI cybersecurity policies, uplift of cybersecurity processes to address AI risks, and uplift of cybersecurity controls to address cybersecurity AI specific risks.	Refer to the OpenAI Trust Portal to review OpenAI SOC 2 report and the third-party external penetration testing report available at https://trust.openai.com/. OpenAI is a strategic partner and Microsoft does not disclose details about strategic partnerships or the associated reviews and agreements therein. OpenAI is not a supplier, they are a strategic partner, and their models are an acquired capability over which Microsoft has performed due diligence.
SR3	Do the AI/ML provider supply chain risk assessment reports and scoring (risk assessments) include sections detailing accounting of the AI cybersecurity maturity of their suppliers and/or outsourcing partners and its impact on the AI/ML provider's risk rating and AI/ML provider selection criteria.	Refer to the OpenAI Trust Portal to review OpenAI SOC 2 report and the third-party external penetration testing report available at https://trust.openai.com/. OpenAI is a strategic partner and Microsoft does not disclose details about strategic partnerships or the associated reviews and agreements therein. OpenAI is not a supplier, they are a strategic partner, and their models are an acquired capability over which Microsoft has performed due diligence.
SR4	Do the AI/ML provider audits of third-party suppliers and/or outsourcing partners include reviews of their AI cybersecurity framework, AI cybersecurity policies, AI cybersecurity controls and their effectiveness against known AI threats specifically those specified in MITRE Atlas, OWASP TOP 10 LLM, ENISA Artificial Intelligence Cybersecurity Challenges, ENISA Securing Machine Learning Algorithms, etc.	Refer to the OpenAI Trust Portal to review OpenAI SOC 2 report and the third-party external penetration testing report available at https://trust.openai.com/. OpenAI is a strategic partner and Microsoft does not disclose details about strategic partnerships or the associated reviews and agreements therein. OpenAI is not a supplier, they are a strategic partner, and their models are an acquired capability over which Microsoft has performed due diligence.
SR5	Does the AI/ML provider perform and/or maintain a list of acceptable use cases where AI is allowed as well as use cases where AI is forbidden e.g. as defined in the EU Commission proposal for a Regulation on artificial intelligence and enforces it on its third-party suppliers and outsourcing partners e.g. via contractual clauses.	Microsoft develops documentation that describes how to use our generative AI solutions. The Microsoft Office of Responsible AI defines restricted (prohibited) uses of AI within Microsoft.
MD1	When sourcing data, does the model developer perform threat modelling and risk assessment for AI specific threats, specifically those specified in MITRE Atlas, OWASP TOP 10 LLM, ENISA Artificial Intelligence Cybersecurity Challenges, and ENISA Securing Machine Learning	**From OpenAI:** OpenAI has a robust threat modeling function owned by the Security team. This threat model is based on OpenAI's Model Development and Deployment framework and takes steps to mitigate risks at every layer of training and deployment, as identified by our threat model. This includes many of the AI specific risks under the listed frameworks, but also risks identified specific to our company and deployment model. This includes, but is in no way limited to, the

	Algorithms? These risks include but are not limited to identification of data poisoning attacks (data and/or labels), malicious content within the data, adversarial perturbations, or any other data that may create a cybersecurity risk?	identification of data poisoning attacks, malicious content in the data, adversarial perturbations, etc. Beyond this, OpenAI deploys numerous mitigating controls, applies strict access controls to training data, and rigorously tests the models at every stage of training to identify possible issues with the data.
MD2	When sourcing data, does the model developer implement AI specific controls to detect/prevent/respond and recover from AI specific threats, specifically those specified in MITRE Atlas, OWASP TOP 10 LLM, ENISA Artificial Intelligence Cybersecurity Challenges, ENISA Securing Machine Learning Algorithms, including but not limited to AI specific tools to detect and prevent data poisoning attacks (such as adversarial perturbations, embedded malware, malicious AI specific content, label switching), human inspection of sourced data, data sanitization, etc.?	**From OpenAI:** Yes, OpenAI utilizes numerous public and proprietary controls specific to AI to detect, prevent, respond, and recover from threats specific to data quality.
MD3	When sourcing data, does the model developer implement all of the following (i) change management controls for the data and model (ii) data integrity controls (iii) identity and access management (IAM) controls (iv) logging, monitoring and accounting controls (v) backup controls that would detect, prevent, and recover from modification of the data throughout its lifecycle, specifically during the following phases outlined in ENISA Securing Machine Learning Algorithms (page 13 figure 3): data collection, data cleaning, data pre-processing, model training, model testing, optimization, and model evaluation phases. Example: data is stored in a central repository with version control and integrity protection, with tamper proof / resistant logs and role based access control implementing least privilege access to the data and logging any changes to the data and the model is tracked including the relationship between the model and the relevant	Microsoft has implemented these controls as described in the existing SOC 2 report; Azure OpenAI is included in the SOC 2 scope. Please refer to the Azure SOC 2 report for further information.

	data sets used in the specific stages of the model development.	
MD4	Does the model developer ensure that effective security controls are implemented to protect the model throughout the model development lifecycle considering each lifecycle stage of the ENISA Securing Machine Learning Algorithms (SMLA page 13 figure 3)? Do the controls include:	See below
MD5.1	i. cybersecurity and AI specific threat modelling and risk assessment is performed to identify potential cybersecurity threats to the developed model and system, specifically those listed in MITRE Atlas, OWASP TOP 10 LLM, ENISA Artificial Intelligence Cybersecurity Challenges, ENISA Securing Machine Learning Algorithms and appropriate controls are deployed to manage those risks.	**From OpenAI:** OpenAI has a robust threat modeling function owned by the Security team. This threat model is based on OpenAI's Model Development and Deployment framework and takes steps to mitigate risks at every layer of training and deployment, as identified by our threat model. This includes many of the AI specific risks under the listed frameworks, but also risks identified specific to our company and deployment model. This includes, but is in no way limited to, the identification of data poisoning attacks, malicious content in the data, adversarial perturbations, etc. Beyond this, OpenAI deploys numerous mitigating controls, applies strict access controls to training data, and rigorously tests the models at every stage of training to identify possible issues with the data.
MD5.2	ii. training & testing against adversarial and/or malicious content / interactions	**From OpenAI:** Requires role assignment to access; data moves from external sources to storage accounts controlled by OpenAI over an encrypted channel (PrivateLink or Internal Network).
MD5.3	iii. testing for fail safe and graceful shutdown of the model	Microsoft hosts the OpenAI model in the Azure OpenAI environment and performs service error handling as part of service uptime commitments. Refer to the Microsoft 365 uptime SLAs.
MD5.4	iv. testing for model availability against classic and model specific AI DoS attacks (e.g. long inference activities)	Microsoft hosts the model and manages availability to the model. Microsoft mitigates the risk of DDoS attacks using a mechanism called OneDDoS. From the Office 365 FedRAMP SSP Control SC-5, "Denial of Service Protection, "Azure implements OneDDOS for the Office 365 service teams as defense against single point and distributed network flooding denial of service attacks. Individual service teams request OneDDOS deployment from Azure. OneDDOS is a solution for network-wide, non-intrusive reporting, anomaly detection and intelligent mitigation. Using flow data, SNMP and BGP updates, OneDDOS learns normal traffic and routing behavior across hundreds of routers and thousands of interfaces and correlates the traffic patterns with the topology data to build logical data models. This information enables network and security operations staff to detect and mitigate threats to availability, improve network/service

		performance and make better investment decisions concerning capacity planning, service offerings and traffic management. In addition, service teams use redundant server implementation within each data center as well as mirrored active/active data centers to enhance availability of services. The Microsoft Security Development Lifecycle requires planning for software/logic-based denial of service attacks and minimizing their potential effects."
MD5.5	v. change and version management and control of the model and its data throughout its lifecycle	Refer to the GPT-4 System Card
MD5.6	vi. integrity protection of the model throughout its development lifecycle	**From OpenAI:** All training data is stored in secure Azure Storage Accounts, and communication with these storage accounts happen over Private Links. Access to this data is restricted and only available to those that have a business need.
MD5.7	vii. logging, monitoring, and accounting of all changes to the model	Refer to the GPT-4 System Card
MD5.8	viii. protection against insider threats such as maker – checker or similar consensus / quorum-based controls	Refer to the GPT-4 System Card
MD5.9	ix. IAM role (RBAC) or attribute-based (ABAC) access controls that enforce least privilege access	**From OpenAI:** All training data is stored in secure Azure Storage Accounts, and communication with these storage accounts happen over Private Links. Access to this data is restricted and only available to those that have a business need. [Data and Model Code] Requires role assignment to access; data moves from external sources to storage accounts controlled by OpenAI over an encrypted channel (PrivateLink or Internal Network).
MD5.10	x. Input sanitation and input bounding (when appropriate / feasible)	Microsoft has internal Microsoft 365 security requirements, which require both input and output sanitization. These requirements prescribe untrusted data, including inputs, must be inspected and validated before processing. As part of Microsoft's RAI processes, services are assessed against additional internal policies before they are released to production. These policies require necessary features and processes to monitor and address problematic inputs (prompts) and responses as they are discovered and as close to real-time as possible. Microsoft employs several layers of mitigation, including the safety system (filters), meta prompting and grounding and UX level. Note the Office 365 FedRAMP SSP Control SI-10 Information Input Validation: "...Office 365 follows system development methodology and security guidelines outlined in the Office 365 Information Security Policy. The process addresses requirements around input data validation within applications. Office 365 has implemented information validation through checking of data inputs.

		Thorough code reviews and testing are completed during the Verification Phase prior to software being put into a production environment. The code reviews and testing check among others for cases of SQL injection, format string vulnerabilities, XSS, integer arithmetic, command injection, and buffer overflow vulnerabilities..."
MD5.11	xi. Output sanitation and Output bounding (when appropriate / feasible)	Microsoft has internal Microsoft 365 security requirements, which require both input and output sanitization. These requirements prescribe untrusted data, including inputs, must be inspected and validated before processing. As part of Microsoft's RAI processes, services are assessed against additional internal policies before they are released to production. These policies require necessary features and processes to monitor and address problematic inputs (prompts) and responses as they are discovered and as close to real-time as possible. Microsoft employs several layers of mitigation, including the safety system (filters), meta prompting and grounding and UX level. Note the Office 365 FedRAMP SSP Control SI-10 Information Input Validation: "...Office 365 follows system development methodology and security guidelines outlined in the Office 365 Information Security Policy. The process addresses requirements around input data validation within applications. Office 365 has implemented information validation through checking of data inputs. Thorough code reviews and testing are completed during the Verification Phase prior to software being put into a production environment. The code reviews and testing check among others for cases of SQL injection, format string vulnerabilities, XSS, integer arithmetic, command injection, and buffer overflow vulnerabilities..."
MD5.12	xii. Output and model behavior monitoring for abnormal activities / behavior	Microsoft service teams monitor the model for abnormal activities with egress filtering to ensure that malformed, malicious, or inappropriate responses do to reach the end user.

Additional Resources

Transparency Notes

Responsible AI Transparency Report 2024

The following are the Copilot Transparency Pages for the various Microsoft 365 offerings:

- Word: Frequently asked questions about Copilot in Word - Microsoft Support
- PowerPoint: Welcome to Copilot in PowerPoint - Microsoft Support
- Loop: Frequently asked questions about Copilot in Loop - Microsoft Support
- Outlook: Frequently asked questions about Copilot in Outlook - Microsoft Support
- Teams: Frequently asked questions about Copilot in Microsoft Teams - Microsoft Support
- Excel: Get started with Copilot in Excel - Microsoft Support
- OneNote: Summarize your OneNote notes with Copilot for Microsoft 365 - Microsoft Support

Responsible AI and NIST AI Risk Management Framework

NIST AI Risk Management Framework (RMF): Microsoft participated in the development of the NIST AI RMF and its internal Responsible AI Standard is closely aligned with it. We've committed to moving forward to ensure implementation of the AI RMF across Microsoft. To learn more, please visit the Compliance offerings for Microsoft 365, Azure, and other Microsoft services.

Copilot Frequently Asked Questions

Frequently asked questions for Copilot for Microsoft 365 to help address a number of the transparency questions – please see further: Frequently asked questions about Microsoft 365 Copilot. Additionally, the Microsoft Copilot for Microsoft 365 overview | Microsoft Learn and Data, Privacy, and Security for Microsoft Copilot for Microsoft 365 | Microsoft Learn documentation also includes several answers to questions most frequently asked by our customers, including:

- How does Microsoft Copilot for Microsoft 365 use your proprietary organizational data?
- How does Microsoft Copilot for Microsoft 365 protect organizational information and data?
- What data is stored about user interactions with Microsoft Copilot for Microsoft 365?
- What data residency commitments does Microsoft Copilot make?
- Can Microsoft Copilot for Microsoft 365 use web content in its responses?
- What extensibility options are available for Microsoft Copilot for Microsoft 365
- How does Microsoft Copilot for Microsoft 365 meet regulatory compliance requirements?
- Do controls for connected experiences in Microsoft 365 Apps apply to Microsoft Copilot for Microsoft 365?
- Can I trust the content that Microsoft Copilot for Microsoft 365 creates? Who owns that content?
- What are Microsoft's commitments to using AI responsibly?

Industry Resources

- MITRE Atlas (link)
- OWASP TOP 10 LLM (link)
- ENISA Artificial Intelligence Cybersecurity Challenges (link)
- ENISA Securing Machine Learning Algorithms (link)
- IBM AI Risk Atlas (link)