17 Securing Copilot

Generative AI is a mysterious and nascent technology. DeepMind and OpenAI created their models for research purposes, ChatGPT hit it big with humanity, and now, we're currently writing history.

The next two chapters cover important aspects of generative AI in the enterprise. To get predictable and protected results, prudent IT executives are working to ensure the data used by Pilots is:

a) Meant for them to see.
b) Accurate and clean.

Securing Copilot starts with the obvious – protecting the data within the tenant. That's where this chapter begins, before touching on more fundamental practices that further an organization's AI readiness.

What is the Risk? How Pervasive is it?

When prompted, **Copilot will search for information in every repository that the *Pilot* has access.** Of course, this includes files in their own OneDrive, their email and chat data, and meetings to which they were invited.

It also means that ***anything* a Pilot has access to in SharePoint – *whether intentionally or inadvertently* – can be found and leveraged by Copilot.**

A true story illustrates one of the worst cases. A new user asked Copilot Chat to "Search for Social Security Number (123-45-6789)." He was shocked to see his number appear in ~50 files that Copilot found on SharePoint!

True SharePoint story number two. My firm has been called multiple times by K-12 school systems whose students "hacked" their way into the grading repository on SharePoint. It turned out that they merely clicked around SharePoint until they found a teacher's folder - whose permissions had been left wide open! They'd have gotten an A if they were in a cybersecurity class!

The risk is real. **The extent of the risk varies depending on how much an organization uses Microsoft 365 and how much work has been done to secure it.** Taking no chances, this chapter starts from the basics. CISOs would be wise to ensure the following checks and balances are in place.

Technical Risks:

Good folks may see too much

Today, people must go out of their way to search for files in SharePoint that they shouldn't see. They either hunt and peck around folders and files, or keep try keyword searches. If they see something valuable, like a patent-pending paper, they may keep it for their own use, or they may sell it to a competitor.

Overexposure of data is more likely with Copilot, since *well-meaning users* may use a simple search and still *may see too much*. If a well-intentioned employee searches for an "example payroll report template" and unknowingly turns up an active salary roster, they may get offended or blow a whistle about PII.

Microsoft makes tools available to minimize oversharing and control nefarious data loss, but none are on by default.

Bad folks may have easier access

Recall the K-12 students who hacked their way into a loosely secured SharePoint to change their grades? If a similarly bad-intended actor hacks the account of a person in the organization, the bad actor will have a GenAI tool to inquire about the data that to which that person has access. Instead of taking time to poke around, the adversary has faster access to potentially confidential information. **Zero trust principles like Multifactor Authentication/Conditional Access, device compliance, and anomalous behavior detection are more important than ever.**[1]

Microsoft can't control these aspects of an individual organization's tenant or security posture. Before getting to the controls that CISOs should implement, the next section briefly covers Microsoft's role.

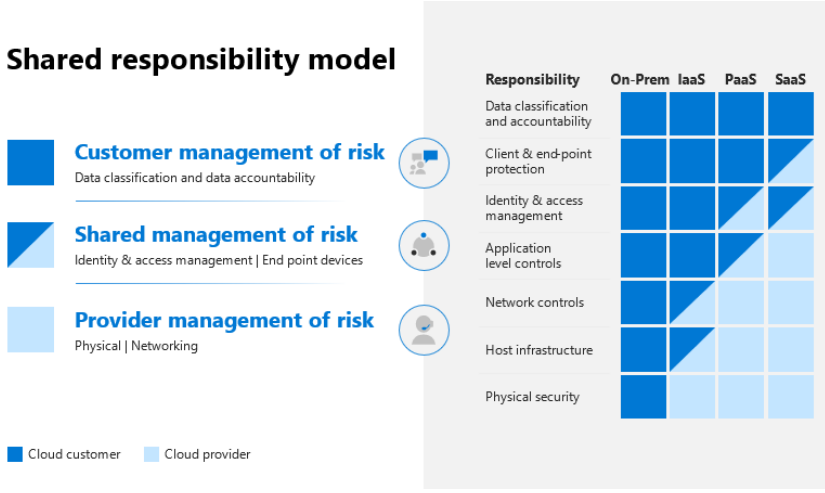
Microsoft's Role in Securing Copilot

Copilot uses the same LLMs as OpenAI and Anthropic, but in a separate enclave dedicated to Microsoft 365 customers. Microsoft further secures Copilot by:

- Ensuring that the inputs and outputs are not used to train the models.
- Deleting the data after processing, ensuring that the prompts and response data are not stored or retained by Microsoft or OpenAI.
- Encrypting the data in transit and at rest.
- Applying the Microsoft Responsible AI principles and practices, such as fairness, reliability, safety, privacy, inclusiveness, and accountability.
- Complying with the Microsoft Trust Center and the Microsoft Service Trust Portal, which provide information and resources on the security, privacy, compliance, and transparency of Microsoft cloud services.[2]
- Monitoring and auditing the generative AI models and data, ensuring that the service is operating as expected and detecting any anomalies.
- Allowing customers to provide feedback and report any issues.

Customers' Role in Securing Copilot

Microsoft **and** their customers share the responsibility for overall security and compliance. In fact, **data security is atop the list of customer responsibilities**, as shown.



To address data classification and protection, Microsoft's own first-party data protection service, Purview Information Protection, can be used. To ensure the correct Identity and Access Management controls are in place, SharePoint and Entra ID are used to manage privileges. **The main risks and methods for addressing them are shown in the table below.**

Risk	Likelihood	Risk Level	Mitigating Controls
Internal Overexposure	High	Low to medium	1. Protect data with Purview 2. Limit SharePoint privileges
Internal Misuse	Medium	Medium to high	1. Protect data with Purview 2. Limit SharePoint privileges 3. Monitor usage
External Misuse	Low	High	1-3 above, *plus:* 4. IAM and other Zero Trust Principles

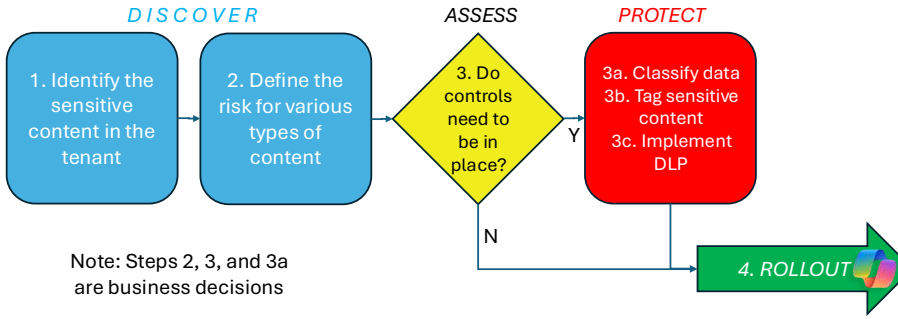
Protect files using Purview Information Protection

First, *at the file level*, use Microsoft Purview Information Protection to identify sensitive information, and then set policies to protect such content. Purview Information Protection helps you discover, classify, label, and protect your sensitive data all across Microsoft 365. It's part of a broader group of Purview data governance solutions from Microsoft. From here on, it'll be shortened to Purview to save space.

Sensitive data is commonly Personally Identifiable (or Health) Information (PII/PHI), but can include intellectual property, R&D, customer files, or legal records.

Purview had primarily been used by firms with regulatory or legal requirements to protect sensitive data. Its original use was to (for instance) keep people from sending confidential emails to outside parties, and to keep confidential files from being uploaded to Box or Dropbox. **In addition, Purview can create and enforce policies that *control how sensitive data is accessed, shared, and used by Copilot.***

A process for protecting data with Purview, and deciding when to activate Copilot in production, is summarized next:



Behind each box in the chart lie some significant steps, outlined next:

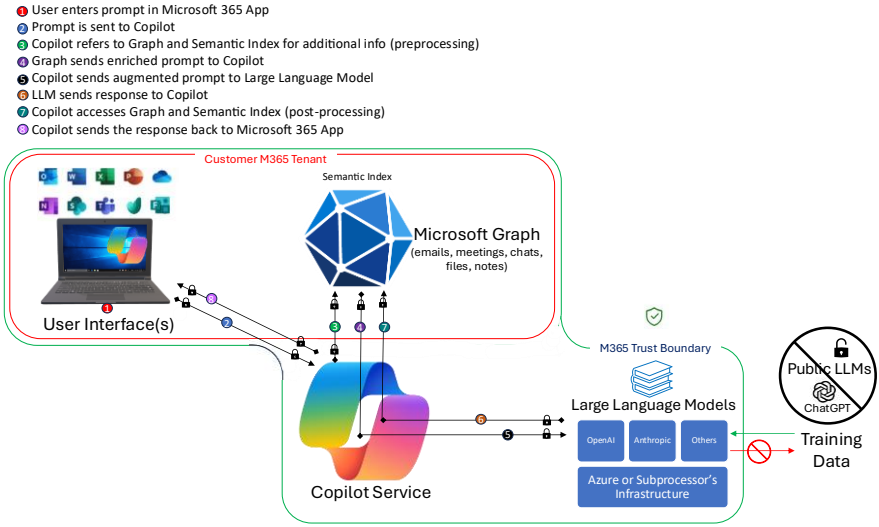
1. Identify Sensitive Information - Purview can discover and classify sensitive information within your tenant. This includes identifying files containing confidential data, personally identifiable information (PII), or other sensitive content based on predefined or custom data patterns.
2. Define the risk for various types of content – This typically includes PII, health data, credit cards, company (trade) secrets, intellectual property, customer data, and corporate records.
3. Do controls need to be put in place to protect that data? This is a tough business decision. While most people would agree to some controls, agreeing on exactly which controls and for which users is a tougher call. For those that agree that protecting data is a prudent prerequisite:
   a. Purview allows files to be **classified and labeled** based on the type and sensitivity of the information they contain. There are built-in classifiers and labels, and organizations can create custom tags.
   b. With a Microsoft 365 E3 license, people need to **manually take action to tag files with a sensitivity label** (such as confidential, personal, or public). **With an E5, the system can automatically tag such files,** in case people forget or ignore the label.

Implementing DLP is easier said than done. The discovery process is simple using Purview Content Explorer. Business input is needed to help IT identify which content (i.e. keywords or sites) are sensitive, and which need to be restricted. **Then, it realistically takes 6-12 months to operationalize a DLP implementation.**

c. Once files have been classified and labeled, Data Loss Prevention (DLP) policies must be configured to dictate how to protect them. For example, sensitive files can be blocked from sharing outside of a small group, or blocked from being printed, or saved. Exceptions can be set, allowing certain Microsoft 365 Security Groups to handle sensitive data, but not others. Each application can have different rules as well, but usually the same DLP rules govern content in SharePoint, OneDrive, Exchange, and Teams. Tool tips, alerts and administrative notifications can flag any violations or incidents that occur.

When Copilot finds confidential data that is legitimately accessible by its Pilot, **Purview will also protect any new files which are created using the sensitive content**.

Copilot checks if the Pilot is a legitimate viewer of the sensitive information before the response is presented to them (Step 7 ("Post-Processing") in the flows and arrows below).



Using Purview Information Protection reduces the chance that Copilot accesses or exposes content that is sensitive while helping to meet relevant regulations and standards.

Set and manage retention policies

All Pilots' prompts, responses, and meeting transcripts are tracked. Some organizations want to manage their retention. Purview can do that too.

All user prompts to Microsoft 365 Copilot and the corresponding responses are stored in a hidden folder in the user's mailbox. It has the RecipientTypeDetails attribute of UserMailbox, which also stores message data for Teams private channels and cloud-based Teams users. This folder is not browsable by users or administrators. Only compliance administrators can search the data with eDiscovery tools.

Messages from Copilot are included in the retention policy location called "Teams chats and Copilot interactions," since they follow the same retention and deletion processes. Once a retention policy is set up for Copilot interactions, and the messages expire, they are transferred to the SubstrateHolds folder, another hidden folder in every user mailbox that temporarily stores "soft-deleted" items before their permanent deletion.

IT admins can use eDiscovery (Premium, an E5 feature) and the Microsoft Graph Explorer to locate and delete user prompts and responses. This helps when it's necessary to identify and remove sensitive or inappropriate content. Note that users are also able to delete their own Copilot history. This function is found in their Microsoft 365 Account Settings and Privacy options.

When a user exits the organization and their Microsoft 365 account is removed, any Copilot messages under retention are stored in an inactive mailbox. These messages still follow the original retention policy and can be accessed via an eDiscovery search.

As for Teams meetings where a Copilot transcript has been taken, the options are more varied. Most of the time, meeting recordings and transcripts are saved to OneDrive. Transcripts are stored separately from recordings in the meeting organizer's Exchange mailbox. Both can have separate retention settings.

A written retention policy should precede any encoded configurations, especially when legal holds are involved. This ensures that the data is retained according to business or regulatory requirements, allowing IT to configure per the organization's data retention and compliance policies.

One more note on retention. It's a truism that "Garbage in, garbage out." **Using Purview retention and deletion rules on the files themselves in OneDrive and Sharepoint will cull the data estate, helping Copilot access only up to date, accurate content.**

Minimize Access Privileges in SharePoint

Next, to keep the likes of those K-12 students from *easily* accessing the crown jewels, you should audit and tighten the access to content. This effort falls under the moniker "Data Access Governance."

To secure Copilot in your tenant, you should minimize the privileges of users and groups in SharePoint and Teams, and only grant them the access and capabilities that they need. You should also review and update the permissions and roles regularly, and remove any users or groups that are no longer active or relevant. This way, Copilot won't access or expose content that is not intended for the Pilot, and reduce the risk of unauthorized or accidental access or sharing.

This is an ongoing process, but should begin with the sites and folders that are known to contain sensitive information. How does a CISO get an idea of the sites are over-permissioned? **A recent and rapidly improving tool, SharePoint Advanced Management (SAM), can help.**

To get a big-picture view of the over-permissioning within the organization, SharePoint Admins can navigate to the "Advanced Management" portal and start a "Content Management Assessment."

That report will give a good starting point about the sites the permissions assigned to SharePoint sites and Teams to identify potential over-privileged access. It provides an **executive summary report that can:**

1. Identify sites with no activity in the past 180 days
2. Detect sites without owners or with only one owner
3. Find sites where permission inheritance is broken
4. Locate content shared with all internal users via Everyone Except External Users
5. Discover content shared through overly permissive sharing links

Then, options and instructions for the following techniques are recommended:

1. Reduce exposure by controlling which SharePoint sites are visible to the person using **Restricted Access Control** (RAC) at the site level. This process is ideal for HR, finance, exec, legal, or sites with crown jewels. This completely keeps unauthorized people from navigating, linking to, and using Copilot to find files stored on the site.
2. Going through the process of granularly assigning permissions to all SharePoint sites can be time-consuming. For a quicker compromise,

enable **Restricted Content Discovery** (RCD). This setting can be activated for specific sites. People that don't have direct access to the site won't be able to find the content in their Copilot search. This doesn't completely cut off access to sites like RAC does, but simply hides their content from people using Copilot.

3. Disable or restrict use of company-wide sharing groups and "Anyone" links at the tenant level.
4. Identify inactive and unmanaged SharePoint sites to delete or restrict access to, and retire stale sites using SharePoint lifecycle policies.
5. Limit site sprawl through tighter site creation governance.

Specific to RAC, **each site owner can:**

1. Navigate to the SharePoint site or Team for which you want to review permissions.
2. Select "Settings" and then "Site permissions."
3. Review the list of users and groups with access to the site, their permission levels, and the scope of their access (site or sub-site level).
4. Identify any users or groups with excessive permissions or access to sensitive content they do not require.

This process is more scalable when similar users are aggregated into Microsoft 365 Groups.[3]

By default, if a person is unable to access a file or site, SharePoint provides a button to request access. Site owners can configure the feature to send an email when someone requests access to a site, so that they can choose whether to approve or decline their request.

To configure Access Request Management to control and monitor requests to access sensitive sites or content:

1. Navigate to the SharePoint site or Team for which you want to enable Access Request Management.
2. Select "Settings" and then "Site permissions."
3. Under "Site Permissions," select "Access Request Settings."
4. Configure the appropriate settings, such as who can request access, approval workflows, and notification settings.
5. Assign designated approvers to review and approve or deny access requests.

As you can see, Copilot is not the root cause of the risk of oversharing of data, but **it exacerbates oversights that have quietly existed**. By minimizing privileges and implementing access controls, you can reduce the risk of unauthorized access to sensitive content and limit the potential exposure through Copilot.

Specifically regarding Restricted Content Discovery, which keeps specified SharePoint sites from Copilot's search:

1. Identify the sites: Work with business stakeholders to identify SharePoint sites containing highly sensitive or confidential information that should be excluded from the search index and document the identified sites and their URLs for the next steps.
2. Conduct a pre-test to later validate the results
    a. Put a file with some obvious, unique content into one of the sites that will be excluded in step 3.
    b. Search for content within the file via Enterprise Search and/or Copilot.
    c. Confirm that it's visible.
3. Exclude the site(s) from the Search Index: Use the SharePoint Search Configuration settings to remove confidential sites from the search index.
    a. Browse to the site to exclude (with appropriate administrator permissions).
    b. Select "Settings" then "Site information" from the drop-down menu.
    c. Select "View all site settings" to bring up the Site Settings page.
    d. Select "Search and offline availability" under the Search category and select "No" for "Allow this site to appear in search results." That excludes it from both Microsoft Search and the semantic index search. This can also be performed with PowerShell for multiple sites.
    e. Repeat steps a-d for each confidential site you want to exclude.
4. Verify the exclusions worked: After excluding the confidential sites, verify that their content is no longer accessible through the search index or Copilot.
    a. Perform a search query within Copilot or the Enterprise Search for content that should have been excluded.
    b. Confirm that no results from the confidential sites are returned.

c. If necessary, consult the SharePoint Search logs or Microsoft Support for further troubleshooting.

By removing highly confidential SharePoint sites from the search index, you can ensure that their sensitive content is not inadvertently exposed through Copilot's search capabilities.

Manage Copilot Pages

While editing and sharing Copilot Pages (described in Chapter 4) is beneficial, administrators must understand where the data is stored and how it is managed. Copilot Pages are housed in SharePoint Embedded containers, just like their underlying service, Loop app workspaces. These containers can be found in the SharePoint admin center. Such containers are labeled as "Pages," and the owner's name appears as a container property.

Pages, like their underlying Loop components, can be shared with "People in <the organization>," "People you choose," or "People with existing access." Users should be trained to pick the most restrictive access.

Pages can be tagged with Purview sensitivity labels. If the Copilot Page is built from any document that is tagged with the org's "Internal" or "Confidential" label, so shall the Page. The user can assign a different sensitivity label if appropriate.

What happens to the Page when a user account is deleted? Microsoft states that Copilot Pages content is "lifetime-managed with the user account and is deleted when the user account is removed from the organization. Initially, there's a soft deletion phase where recovery by an IT Admin is possible, followed by final purging." Additionally, there's provision for an "Admin workflow to enable access to these containers before deletion, allowing valuable content to be transferred to new locations." The care of Copilot Pages is now part of the responsibilities associated with preserving the data of personnel leaving an organization.

To Give Feedback (or not)

LLMs, including Copilot, can be wrong. Remember, they're drawing from information created by humans.

When something goes wrong, what can be done? This chapter is about how to give feedback to Copilot itself, and to humans at Microsoft. The chapter title is purposeful, because **giving feedback may unintentionally overshare organizational data.**

Give Feedback Directly to Copilot

A good prompt should be followed by feedback and follow-up questions that can help Copilot improve its response and learn from its mistakes.

Feedback can include positive or negative comments, corrections, suggestions, or requests for clarification that can indicate how well Copilot performed the task and how it can do better next time.

Follow-up questions can include additional or related requests, modifications, revisions, or expansions that can refine, improve, or extend Copilot's response and keep the conversation going.

For example, *unhelpful feedback* about Copilot's response to a task from would be: *"This didn't do what I expected."*

Helpful feedback would be *"This is a decent sales pitch, but it could be more persuasive by using some statistics or testimonials to support the claims."*

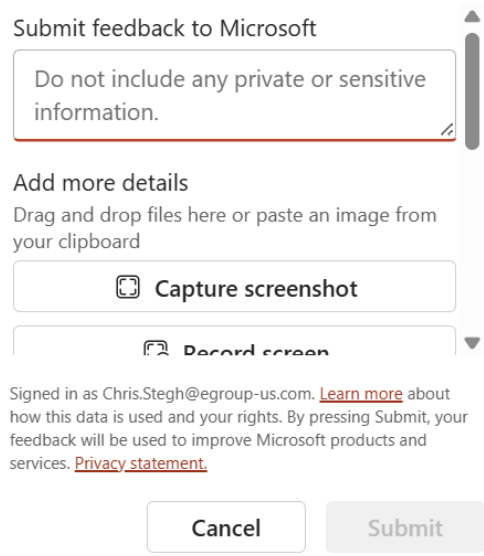
An example of a good follow up question is: *"What are some ways I can give you feedback and ask follow-up questions to get better results? Give an example or two."*

Have a conversation to coach Copilot, and rest assured, none of these interactions are shared with humans at Microsoft.

Give Feedback to People at Microsoft?

If you're the kind of person that likes to help improve software, you can provide feedback to people working at Microsoft. It's one of the ways they legitimately fine tune their algorithms, **but it has risks.**

After receiving a response, you can click on the Thumbs Up or Thumbs Down button. If you don't see the buttons, your IT department has turned off that option. **IT Pros and CISOs can configure specific feedback controls**.[4]



Copilot asks if it's OK to "Share relevant content samples, and additional log files?" if it's OK to "Share prompt, generated response, relevant content samples, and additional log files?" A "contextData.json" link appears which, if clicked, shows the exact prompt and response that will be sent (as shown at right). It's OK to click that hyperlink and actually quite interesting to see how your prompt is augmented before it's sent to the LLM. But……

If you click "Yes" here, the prompt and any document or spreadsheet that was referenced in the prompt will be sent to Microsoft.[5] There's a chance a person will look at that data. So, to be safe, ask your management or IT leaders if you should include the prompt and content samples. **Unless your company allows it, you should not click "Yes" to share this data with Microsoft, especially if the .doc or .xls you referenced has organizational information in it.**

See Something, Say Something?

When Copilot's not accurate or appropriate, you'll have a decision to make. Whether you provide feedback to relieve stress or to improve the product, use your best judgment. **Don't share the prompt and content if they contain organizational information.**

For business and information security leaders, decide in advance if and who within your organization can send feedback to Microsoft. Set up policies to allow/disallow specific content to be shared.[6] Leaving the decision up to users may overexpose your prompts and samples, which could contain sensitive organizational information.

Applying other Zero Trust Tactics

Zero Trust is a security model that helps you protect your data and resources from cyberattacks, by assuming that no user, device, or network is trustworthy by default. You can apply other techniques of Zero Trust to secure Microsoft 365 Copilot, such as:

- Explicitly verifying the identity and device of your users
  - Using passwordless and multifactor authentication
  - Using the combination of Intune, which can confirm the device is trusted, and Entra ID, which can use multiple factors of authentication, ensures that a login is from a trusted person on a trusted device. If an adversary were to phish a user or convince them into accepting an MFA request, the device that

the bad actor would attempt to login from would be untrusted, and their request to login blocked.

- Enforcing the principle of least privilege
  - Microsoft 365 administrative access should be limited to highly authenticated and trusted personnel on specific devices that are managed by Intune and verified by Entra ID's Conditional Access rules.
- Monitoring and auditing activities and events
  - Using Microsoft 365's audit logs, and potentially rolling the logs into a Security Information Event Management (SIEM) system like Microsoft Sentinel can help flag unauthorized behavior and interrupt a potential breach
- When securing Plugins or Connectors (See Chapter 22)
  - Segmenting your data into secure networks with least privileged partitions is advisable, as opposed to allowing unfettered access to a network or server housing the data.
  - Encrypting your data and communications reduces the opportunity for hijacking. Consider using your own encryption key, using Microsoft Azure Key Vault.
  - Build a software bill of materials, with details about the open source and Git components that were used, and track their vulnerabilities.

Finally, purging old information so that it's not unnecessarily exposed can reduce risk (and increase data accuracy). That's such an important topic that it's covered in the next chapter.

Just Keep Securing

Securing data is an ongoing process that requires regular reviews, updates, and collaboration with business stakeholders to align security measures with the organization's needs. No matter the risk profile, every organization should start by tightly protecting the crown jewels.

Once the data is protected, how can people trust that it's accurate and up to date? That's the focus of the next chapter.

Resources

[1] [2404.01833] Great, Now Write an Article About That: The Crescendo Multi-Turn LLM Jailbreak Attack (arxiv.org)
[2] https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/innovate/best-practices/trusted-ai
[3] https://support.microsoft.com/en-us/office/learn-about-microsoft-365-groups-b565caa1-5c40-40ef-9915-60fdb2d97fa2
[4] https://learn.microsoft.com/en-us/microsoft-365/admin/manage/manage-feedback-ms-org?view=o365-worldwide#specific-policies-you-can-configure
[5] https://support.microsoft.com/en-us/topic/providing-feedback-about-microsoft-copilot-with-microsoft-365-apps-c481c26a-e01a-4be3-bdd0-aee0b0b2a423
[6] https://learn.microsoft.com/en-us/microsoft-365/admin/manage/manage-feedback-ms-org?view=o365-worldwide#specific-policies-you-can-configure