Part II

For IT Pros, CIOs, and CISOs

14 How Copilot Works

This chapter covers how Copilot works in detail. It's meant for the more technical reader or anyone who's curious about the technology, its capabilities, and constraints.

The Components

Microsoft 365 Copilot is built on natural language and large language models. It uses data from the Microsoft Graph, and optionally, the Internet. People access the capabilities via the Microsoft 365 Apps, or through a browser at **www.microsoft365.com**.

Let's break down these components to preface deeper topics that follow.

The Microsoft 365 Apps

You use Copilot within the Microsoft 365 apps such as Word, Excel, Teams, PowerPoint, Outlook, and OneNote.

In each, Copilot interacts with your files and content to perform various tasks, such as formatting, editing, creating content, summarizing, and extrapolating insights.

PRO TIP: For the most consistent experience, use the M365 apps through the browser. Log into **www.microsoft365.com** and open your app(s) of choice. You'll find the same Copilot button as in the desktop apps. The desktop apps are less consistent and less current.

The Copilot Service

Copilot operates as a multitenant service. It's partitioned off from other customer tenants in the same way Exchange Online and Teams are separated for each organization. In the same way one Exchange admin cannot see the emails in another tenant, one organization's Copilot content back and forth to the LLM can't be seen by others.

Large Language Models

Recall that large language models are digital libraries made from deep neural networks that are trained on massive amounts of text, code, and images to learn the patterns and structures of natural language and multimedia. LLMs then generate coherent and fluent text, videos, or images for various purposes.

Which LLM is used?
Copilot is multi-model. Both ChatGPT, and recently, Claude models are available to use at **www.microsoft365.com** and within some apps, like Excel.

Microsoft defaults to "Auto" mode, to let the system decide which model to use, but users who have a preference can choose.

Microsoft Graph

Think of Microsoft Graph as a card catalog, or index, of your work-related library.

A library catalog allows visitors to search and then find physical books on different shelves. You could say that the Graph categorizes and retrieves data from digital "books" – files created in Microsoft 365 services like Outlook, OneDrive, SharePoint, and Teams.

The Graph is the data platform underneath all Microsoft 365 apps and services. It lets Copilot access and analyze your documents, emails, calendars, contacts, tasks, and meetings, with personalized and contextual assistance.

The Graph has been around well before Copilot. It is the fundamental technology behind SharePoint / Enterprise search. It became better known because its data also fueled Delve and Viva Insights. But don't worry, no one can see all of another person's Graph data. People can't even go to a page to see all of their own Graph data.

Semantic Index

With all this data in the Graph, how does Copilot know which data to send to the LLM? Keywords won't do. Keywords are like the old-school catalog system indexing a library's shelves.

It's easier to speak with an all-knowing librarian, in English! That's the role of the *Semantic Index*, covered in detail in Chapter 15.

The Semantic Index is a behind the scenes technical component of Microsoft 365 search.

A semantic search is a search method that tries to understand the meaning and context of search queries, not just the keywords. It considers the person's intention, the word associations, and the meaning and relationships in the data. It's the semantic index that allows fast and accurate augmentation of the original prompt, searching for contextual information in the Microsoft Graph.

The Architecture

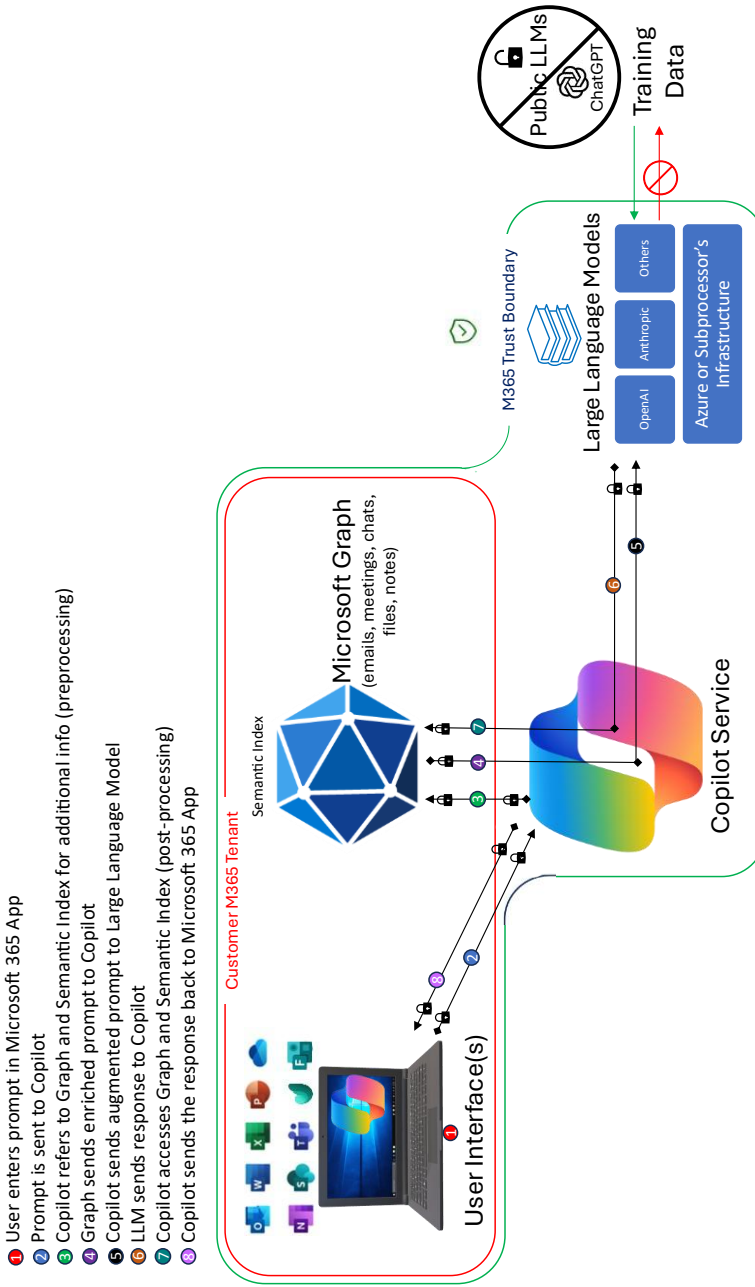
The basic architecture of Microsoft 365 Copilot involves data flow between people using their Microsoft 365 apps, Copilot, the Graph (and its Semantic Index), and the Large Language Model. Note, customer data is not used to train the LLM.

Let's walk through those components with a flow and arrows.

1. A person enters a prompt into a Microsoft 365 App.
   Using a computer or browser for Word, Excel, PowerPoint, or their smartphone for Teams (/Chat) and Outlook, people make requests to Microsoft 365 Copilot by typing or speaking.

2. The prompt is sent to Copilot.
   The prompt and all subsequent transactions are encrypted by HTTPS.

3. Copilot scans the Graph / Semantic Index for additional info.
   Copilot improves the input prompt by using a technique called grounding, which makes the prompt more specific. Grounding helps get answers that are better suited and more useful for the particular task. *This step is also called "pre-processing."*


1 User enters prompt in Microsoft 365 App
2 Prompt is sent to Copilot
3 Copilot refers to Graph and Semantic Index for additional info (preprocessing)
4 Graph sends enriched prompt to Copilot
5 Copilot sends augmented prompt to Large Language Model
6 LLM sends response to Copilot
7 Copilot accesses Graph and Semantic Index (post-processing)
8 Copilot sends the response back to Microsoft 365 App
Customer M365 Tenant
Semantic Index
Microsoft Graph
(emails, meetings, chats, files, notes)
User Interface(s)
M365 Trust Boundary
Large Language Models
OpenAI
Anthropic
Others
Azure or Subprocessor's Infrastructure
Copilot Service
Public LLMs
ChatGPT
Training Data

The Graph accesses relevant information about files, messages, meeting transcripts, and other relevant data. The semantic index creates vectorized indices that enable conceptual understanding, helping Copilot identify and access relevant organizational content.
Now, the enriched prompt contains text from input files or other content that Copilot finds, and is ready to go to the LLM for processing. Copilot only uses data that an individual user can already access, based on existing Microsoft 365 permissions. See Chapter 17 to secure data and permissions.

4. The Graph sends its enriched prompt to Copilot.
   Much more is sent without being seen (i.e. if the prompt included a "/" reference to a file, the entire file is sent as well).
   It's about at this point that something may go wrong, like if the document that was referenced was of an unsupported type or had unsupported styles. The Troubleshooting sections in relevant chapters in Part I highlighted some such circumstances and error messages that could occur.

5. Copilot sends the augmented prompt to Large Language Model.
   When your data is transferred to the LLM, it is encrypted in a secure channel passing through the Copilot shared service. The process of sending this data through a shared (multitenant) Copilot service is like the way Exchange Online works. Millions of customers send emails securely through Microsoft's multitenant Exchange Online engine, but they have no access, visibility, or awareness of one another's traffic.

6. The LLM sends its response to Copilot.
   After processing the modified prompt and using its extensive training data and language capabilities, the LLM generates and returns its response to the Copilot service.

7. Copilot accesses the Graph and Semantic Index (post-processing)
   Copilot interacts with the Graph and Semantic Index several times during post-processing. Some key actions occur:
   a. Grounding Calls: Copilot makes grounding calls to the Microsoft Graph to ensure the response is contextually relevant to the user's data and organizational content.
   b. Responsible AI Checks: The system performs checks to align with Microsoft's standards for responsible AI.
   c. Security and Compliance: Security, compliance, and privacy reviews are conducted to adhere to the organization's policy, typically

instantiated through Purview Information Protection. See Chapter 17.

d. Command Generation: When applicable, Copilot generates commands that can be executed within Microsoft 365 apps (to "Insert a slide" or "Create a table," as examples).

This process ensures that the responses provided by Copilot are not only relevant and personalized but also secure and compliant with organizational policies and privacy standards.

8. Copilot sends the response back to the Microsoft 365 App
   There, the user sees the response to their request, and are (in some apps) given options to "Keep it" or "Regenerate," and give feedback to improve the results. Subsequent prompts remember the prior prompt(s).

All of that can happen rather quickly, depending on the device's connection speed and the size of the prompt and any files uploaded or referenced with "/".

Following that core explanation, the next chapters get into other aspects of technology and controls that provide for secure, safe use of Copilot.