Microsoft

# Using Generative AI on Structured Data

(to query or modify data)

James Serra

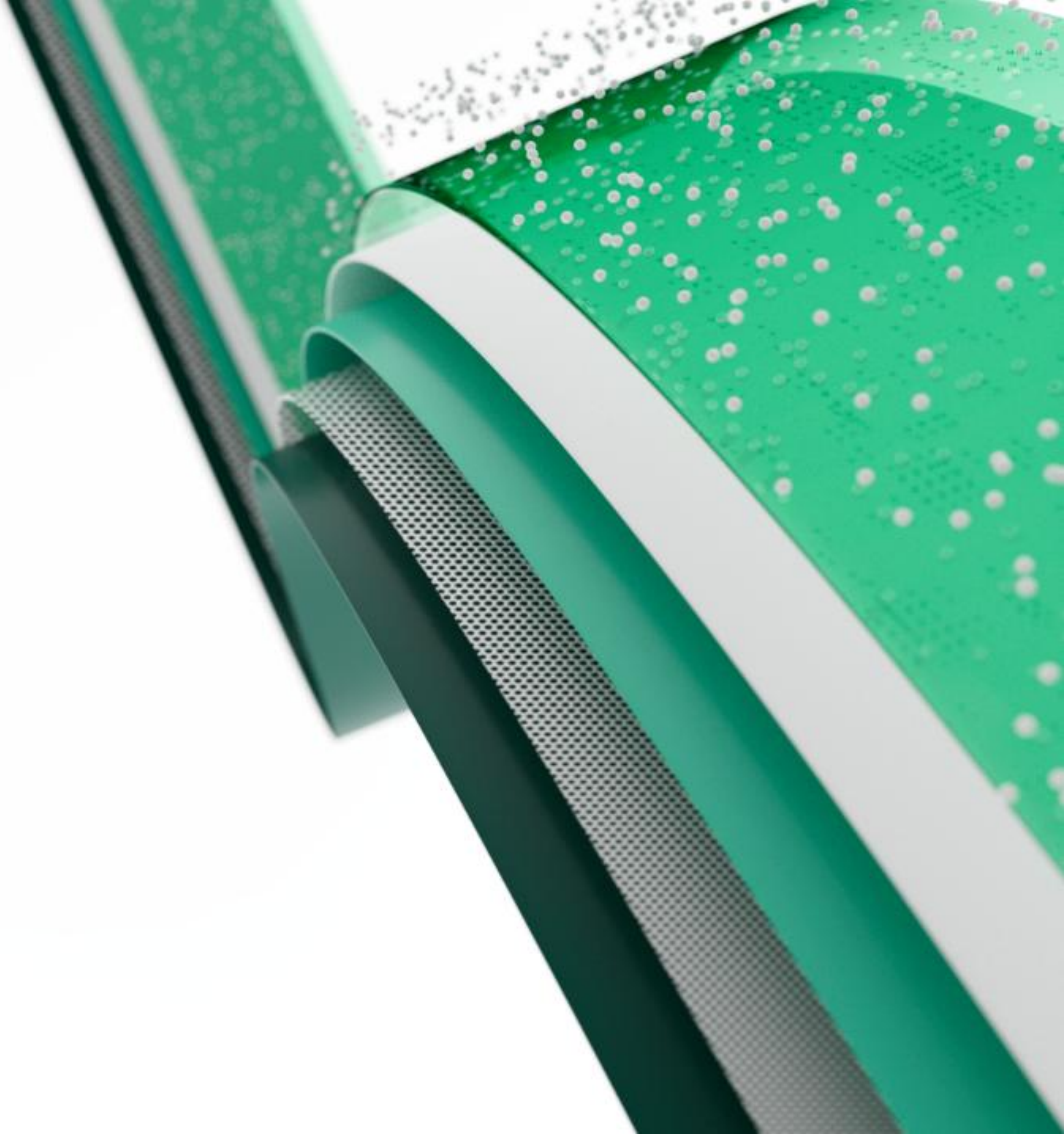
Data & AI Solution Architect

Microsoft, Federal Civilian

jamesserra3@gmail.com

Blog: JamesSerra.com

1/15/25



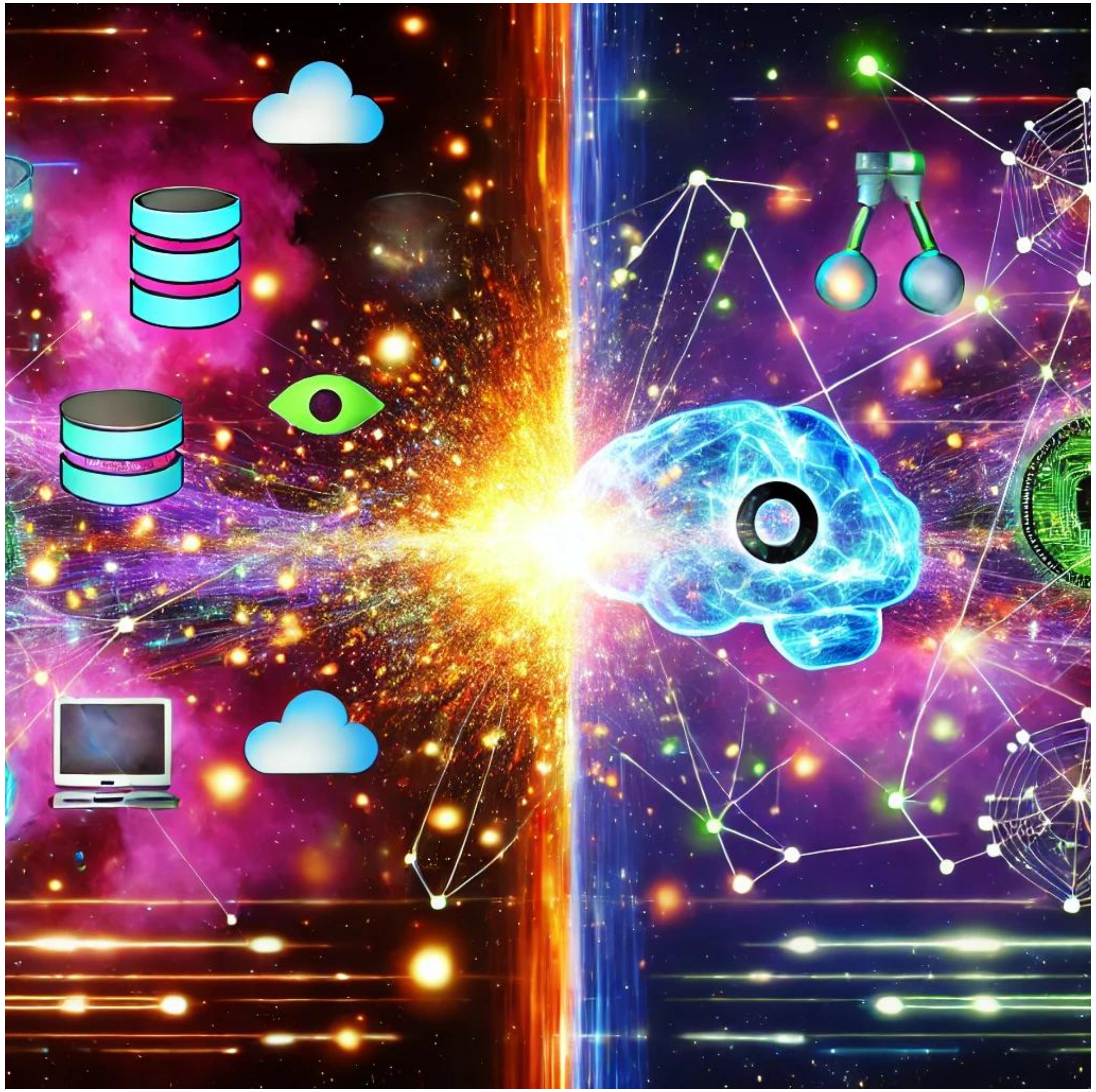
# About Me

- Microsoft, Data & AI Solution Architect in Microsoft Federal Civilian
- At Microsoft for most of the last ten years as a Data & AI Architect, with a brief stop at EY
- In IT for 40 years, worked on many BI and DW projects
- Worked as desktop/web/database developer, DBA, BI and DW architect and developer, MDM architect, PDW/APS developer
- Been perm employee, contractor, consultant, business owner
- Presenter at PASS Summit, SQLBits, Data Summit, SQLDay, Enterprise Data World conference, Big Data Conference Europe, SQL Saturdays, Informatica World (sessionize top 3% most active)
- Blog at JamesSerra.com
- Former SQL Server MVP
- Author of the book “Deciphering Data Architectures: Choosing Between a Modern Data Warehouse, Data Fabric, Data Lakehouse, and Data Mesh”

# Agenda

- GenAI Definitions
- ChatGPT on semi-structured data
- Microsoft Fabric AI Skill on structured data
- Industry use cases

# Data platform and AI worlds are colliding…



# GenAI Definitions

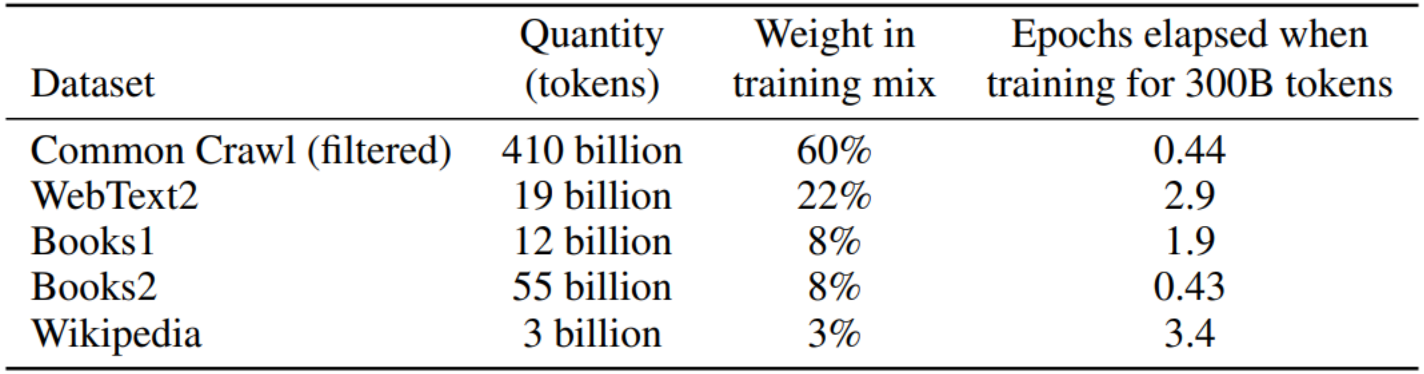
# Is GenAI magic?

# Key definitions

- **AI:** Computational systems and models capable of performing tasks that typically require human intelligence.  GenAI and ML are subsets of AI.
- **Generative AI (GenAI):** AI systems capable of *generating* new content, such as text, images, or audio. It does this by employing neural networks, a type of machine learning process that is loosely inspired by the way the human brain processes, interprets and learns from information over time.
- **Large Language Models (LLMs):** A type of GenAI designed for natural language (text) understanding and generation, trained on diverse datasets. Think of it like a super-smart auto-complete on steroids.  GenAI uses other specialized models for image or video generation.
- **Machine learning (ML):** A broad subset of AI that encompasses algorithms and models capable of learning from data to make *predictions* or decisions. LLMs are a type of deep learning within ML, focused on language understanding and generation, whereas ML includes techniques for diverse tasks like image recognition, data analysis, and predictive modeling.
- **OpenAI**: A leading organization specializing in AI research and development, known for creating Generative AI models such as GPT (used in ChatGPT) and other AI technologies.
- **ChatGPT, Copilot**: Applications ("bots") built on Generative AI models (e.g., GPT). These tools allow users to interact with GenAI via natural language prompts, generating responses tailored to various contexts, such as conversations (ChatGPT) or coding (Copilot).
- **Prompt engineering**: The practice of designing input prompts to guide LLMs in generating accurate and relevant responses, optimizing the model's performance for specific tasks.

# GPT-3 was trained on multiple data sets

- Training Data was taken from the web with some processing
- Some biases inherent in the training data (*as with any model*)
- 530GB of text data, including conversational dialogues



| Dataset | Quantity (tokens) | Weight in training mix | Epochs elapsed when training for 300B tokens |
|---|---|---|---|
| Common Crawl (filtered) | 410 billion | 60% | 0.44 |
| WebText2 | 19 billion | 22% | 2.9 |
| Books1 | 12 billion | 8% | 1.9 |
| Books2 | 55 billion | 8% | 0.43 |
| Wikipedia | 3 billion | 3% | 3.4 |

GTP-4 was trained using both text and image data and is **significantly** larger

- Details of training data are not public

Source: “Language Models are few shot learners”, https://doi.org/10.48550/arXiv.2005.14165 (Section 2.2)

# Data definitions

- **Unstructured data**: emails and documents such as .pdf, docx or .txt files (think text) *-> old school*
- **Semi-structured data**: files and logs in CSV, Parquet, XML, or JSON formats (often in table format, meaning rows and columns); Excel files *-> new school*
- **Structured data**: relational databases (SQL tables with rows and columns) *-> new school*
- **Binary data**: images, audio, video

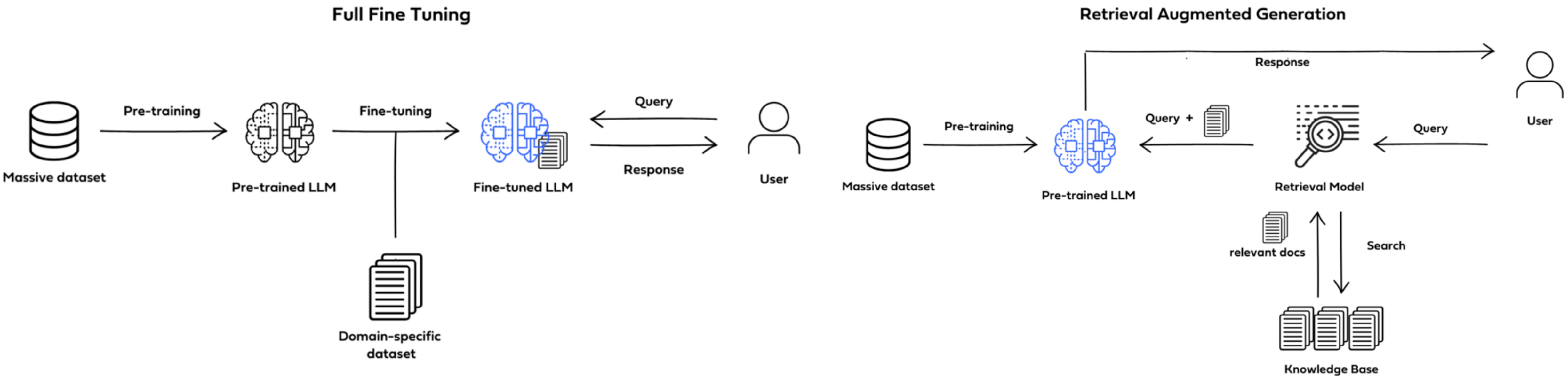
# Model definitions

**Prompt Engineering** is a technique that involves designing prompts for natural language processing models. This process improves accuracy and relevancy in responses, optimizing the performance of the model.

**Retrieval Augmented Generation (RAG)** improves LLM performance by retrieving data from external sources and incorporating it into a prompt . RAG allows businesses to achieve customized solutions while maintaining data relevance and optimizing costs.

**Fine-tuning** adapts an existing LLM using example data, resulting in a new "custom" LLM that has been optimized for the provided examples.

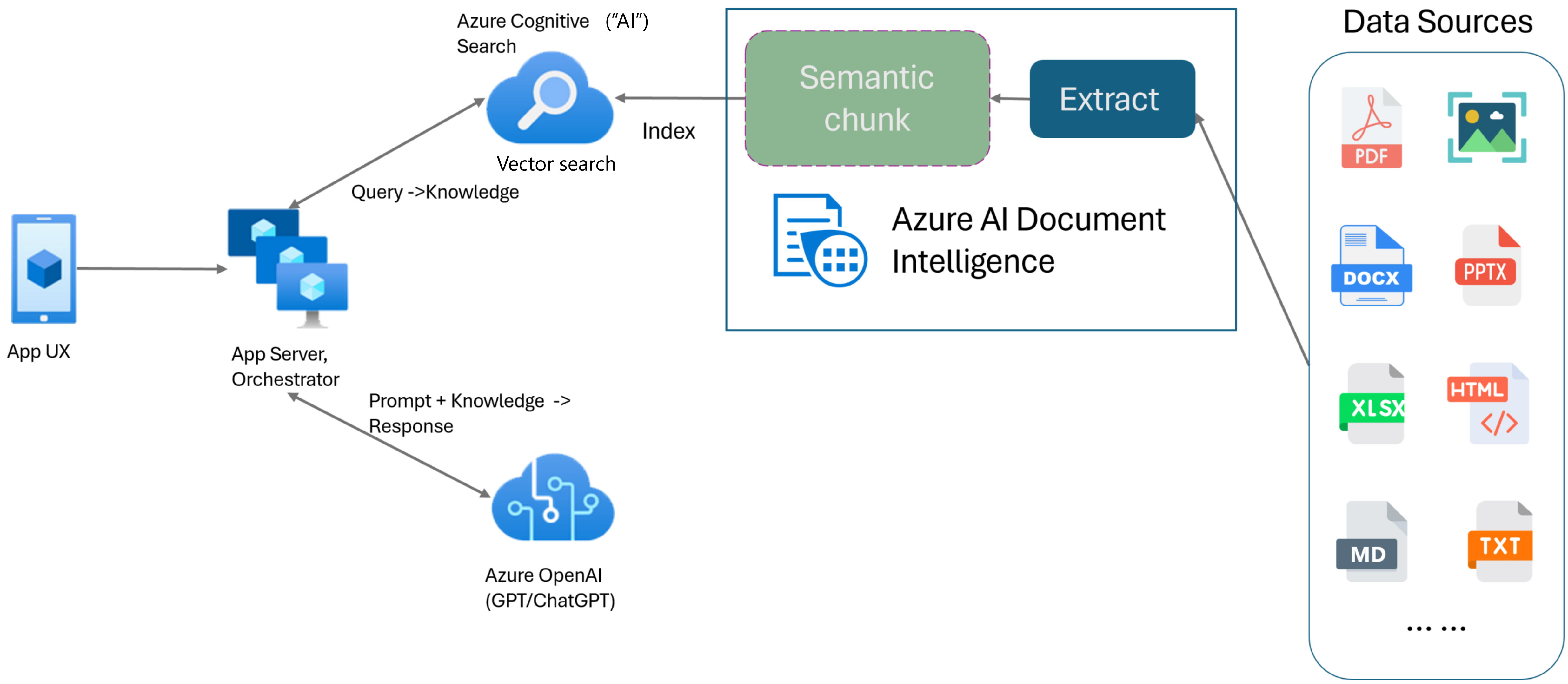
# Fine-tuning vs RAG: A comparison of two techniques for enhancing LLMs



## When to use Azure OpenAI fine-tuning

- Fine-Tuning creates a specialized model with fast, accurate responses for specific tasks but is costly in time and resources. It requires re-training to stay updated, making it static and harder to maintain over time.
- RAG provides real-time, dynamic information, ensuring more accurate answers for constantly changing data. It's cost-effective for scalability, but response times can be slower, and accuracy depends on the quality of external data sources.

# Retrieval-Augmented Generation (RAG) with Azure AI Document Intelligence



# GenAI use cases on semi-structured and structured data

- **Conversational querying**
- **Data enrichment and cleaning**
- Sample data creation
- Data summarization
- Trend identification
- Forecasting and predictions
- What-if analysis
- Anomaly detection and correction
- Product/service recommendations
- Mapping fields, MDM, creating semantic models…

# ChatGPT on semi-structured data

# ChatGPT definition

The “GPT” in ChatGPT stands for generative pre-trained transformer. The acronym covers the key aspects of the magic behind the bot, as follows:

**Generative**: The model creates new content instead of merely classifying or labeling data.

**Pre-trained**: Before engaging in conversation or writing code, the model learns from massive datasets to understand language, grammar, and structure. ChatGPT is a master of pattern recognition and repetition, and it turns out that this is a nearly miraculous capability.

**Transformer**: Introduced in Google’s 2017 paper Attention Is All You Need, this architecture excels at recognizing relationships between words, code, and data, making it adept at complex problem-solving.

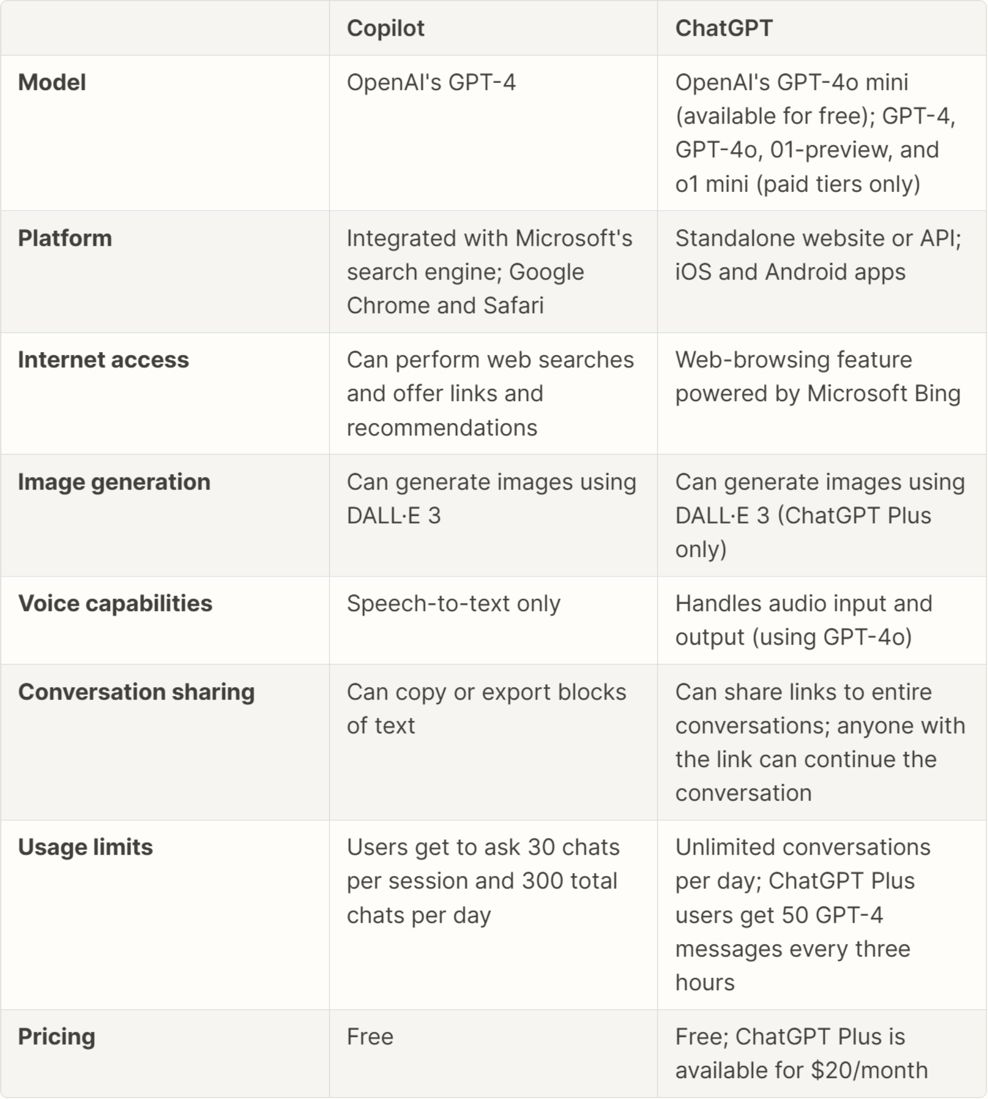
# Bing Copilot (formerly Bing Chat) vs. ChatGPT

Copilot - here is a detailed list of the file types you can upload:

- **Documents**: DOCX, PDF, TXT, MD
- **Images**: JPG, JPEG, PNG, GIF
- **Spreadsheets**: XLSX, CSV
- **Presentations**: PPTX
- **Code files**: PY, JS, HTML, CSS, JAVA, C, CPP, SQL, XML, JSON, YAML, YML
- **Compressed files**: ZIP, RAR, TAR, GZ
- **Audio files**: MP3, WAV
- **Video files**: MP4, AVI, MOV

Max file upload size: 50MB

*Note: only the enterprise version of Copilot supports file uploads*



| | Copilot | ChatGPT |
|---|---|---|
| **Model** | OpenAI's GPT-4 | OpenAI's GPT-4o mini (available for free); GPT-4, GPT-4o, 01-preview, and o1 mini (paid tiers only) |
| **Platform** | Integrated with Microsoft's search engine; Google Chrome and Safari | Standalone website or API; iOS and Android apps |
| **Internet access** | Can perform web searches and offer links and recommendations | Web-browsing feature powered by Microsoft Bing |
| **Image generation** | Can generate images using DALL·E 3 | Can generate images using DALL·E 3 (ChatGPT Plus only) |
| **Voice capabilities** | Speech-to-text only | Handles audio input and output (using GPT-4o) |
| **Conversation sharing** | Can copy or export blocks of text | Can share links to entire conversations; anyone with the link can continue the conversation |
| **Usage limits** | Users get to ask 30 chats per session and 300 total chats per day | Unlimited conversations per day; ChatGPT Plus users get 50 GPT-4 messages every three hours |
| **Pricing** | Free | Free; ChatGPT Plus is available for $20/month |

ChatGPT - You can upload a variety of file types, including:

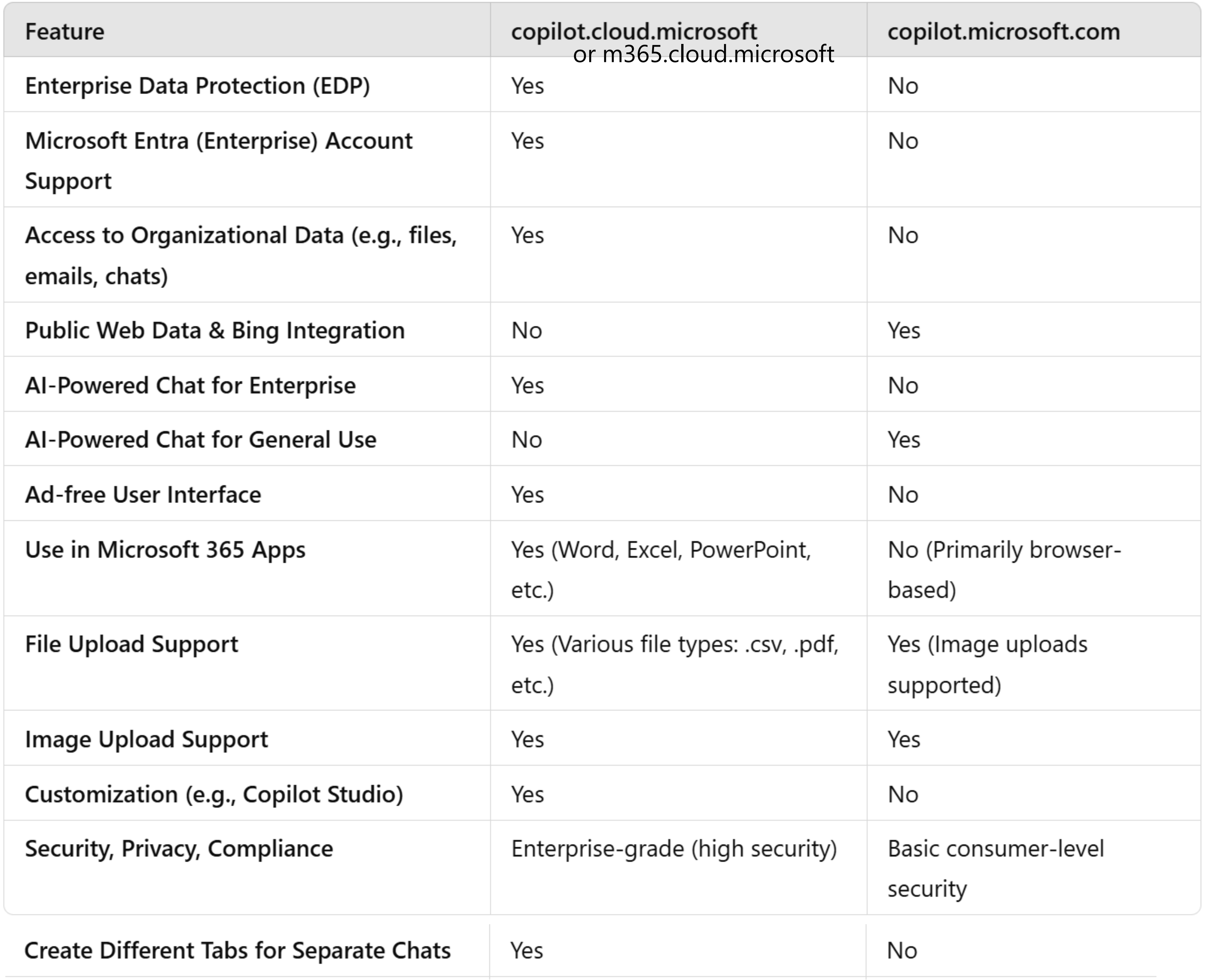
- **Text and Document Files**: PDF, DOC, DOCX, TXT, RTF
- **Spreadsheet Files**: XLS, XLSX, CSV
- **Image Files**: PNG, JPG, JPEG, BMP, GIF, and TIFF
- **Data Files**: JSON, XML, SQL

Max file upload size: 50MB

*Also: Google Gemini, Claude by Anthropic, Perplexity AI, Meta AI*

# ChatGPT versions

| **Plan** | **Cost** | **Features** |
|---|---|---|
| Free | $0/month | - Access to GPT-4o mini<br>- Standard voice mode<br>- Limited access to GPT-4o<br>- *Limited access to file uploads, advanced data analysis, web browsing, and image generation*<br>- Use custom GPTs |
| Plus | $20/month | - Everything in Free<br>- Extended limits on messaging, file uploads, advanced data analysis, and image generation<br>- Standard and advanced voice mode<br>- Limited access to o1 and o1-mini<br>- Opportunities to test new features<br>- Create and use custom GPTs |
| Pro | $200/month | - Everything in Plus<br>- Unlimited* access to GPT-4o and o1<br>- Unlimited* access to advanced voice<br>- Access to o1 pro mode, which uses more compute for the best answers to the hardest questions |
| Team | $25/user/month (annual)<br>or $30/user/month (monthly) | - Higher message limits than Plus on GPT-4, GPT-4o, and tools like DALL·E, web browsing, data analysis, and more<br>- Limited access to o1 and o1-mini<br>- Standard and advanced voice mode<br>- Create and share GPTs with your workspace<br>- Admin console for workspace management<br>- Team data excluded from training by default |
| Enterprise | Custom Pricing | - Everything in Team<br>- High-speed access to GPT-4, GPT-4o, GPT-4o mini, and tools like DALL·E, web browsing, data analysis, and more<br>- Expanded context window for longer inputs<br>- Enterprise data excluded from training by default & custom data retention windows<br>- Admin controls, domain verification, and analytics<br>- Enhanced support & ongoing account management |



| | Microsoft 365 Copilot | Microsoft Copilot |
|---|---|---|
| **Feature** | **copilot.cloud.microsoft**<br>or m365.cloud.microsoft | **copilot.microsoft.com** |
| **Enterprise Data Protection (EDP)** | Yes | No |
| **Microsoft Entra (Enterprise) Account Support** | Yes | No |
| **Access to Organizational Data (e.g., files, emails, chats)** | Yes | No |
| **Public Web Data & Bing Integration** | No | Yes |
| **AI-Powered Chat for Enterprise** | Yes | No |
| **AI-Powered Chat for General Use** | No | Yes |
| **Ad-free User Interface** | Yes | No |
| **Use in Microsoft 365 Apps** | Yes (Word, Excel, PowerPoint, etc.) | No (Primarily browser-based) |
| **File Upload Support** | Yes (Various file types: .csv, .pdf, etc.) | Yes (Image uploads supported) |
| **Image Upload Support** | Yes | Yes |
| **Customization (e.g., Copilot Studio)** | Yes | No |
| **Security, Privacy, Compliance** | Enterprise-grade (high security) | Basic consumer-level security |
| **Create Different Tabs for Separate Chats** | Yes | No |

**Microsoft Copilot**
Free version

**Microsoft Copilot Pro Subscription**
$20.00 user/month

**Microsoft 365 Copilot Subscription**
$30.00 user/month

# ChatGPT demo on semi-structured data

*Also: askondata.com*

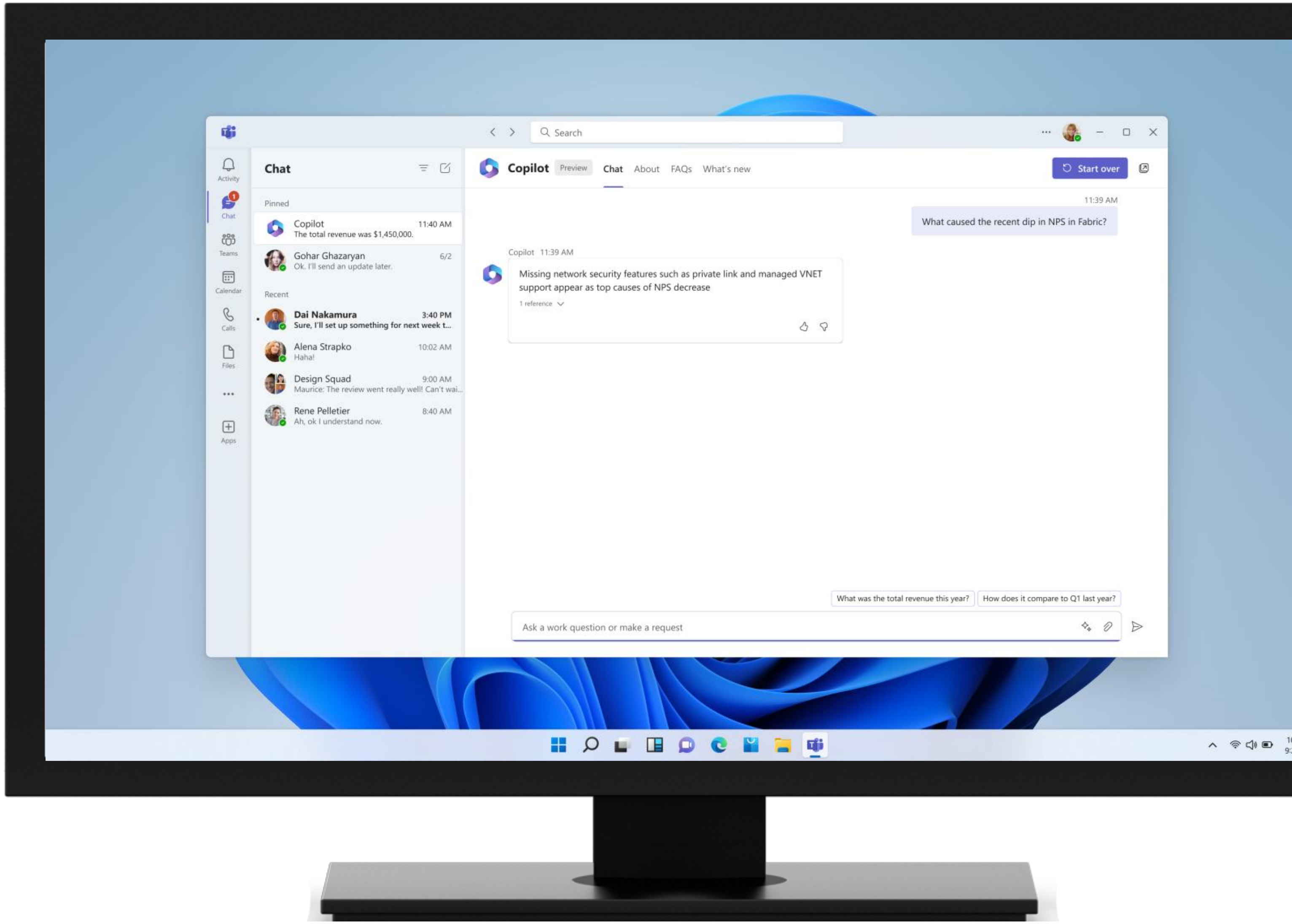
# Microsoft Fabric AI Skill on structured data

# Generative AI on your data

Public Preview

## Deliver custom generative AI experiences for *your* data

- Enable custom Q&A on your data in Fabric
- Define custom business semantics and grounding **unique to your org**
- Scale the custom experiences to **M365 Chat**, **Copilot Studio**, and **Azure AI Studio**



Chat

Pinned

Copilot 11:40 AM
The total revenue was $1,450,000.

Gohar Ghazaryan 6/2
Ok. I'll send an update later.

Recent

**Dai Nakamura** 3:40 PM
**Sure, I'll set up something for next week t...**

Alena Strapko 10:02 AM
Haha!

Design Squad 9:00 AM
Maurice: The review went really well! Can't wai...

Rene Pelletier 8:40 AM
Ah, ok I understand now.

11:39 AM
What caused the recent dip in NPS in Fabric?

Copilot 11:39 AM
Missing network security features such as private link and managed VNET support appear as top causes of NPS decrease
1 reference

What was the total revenue this year?

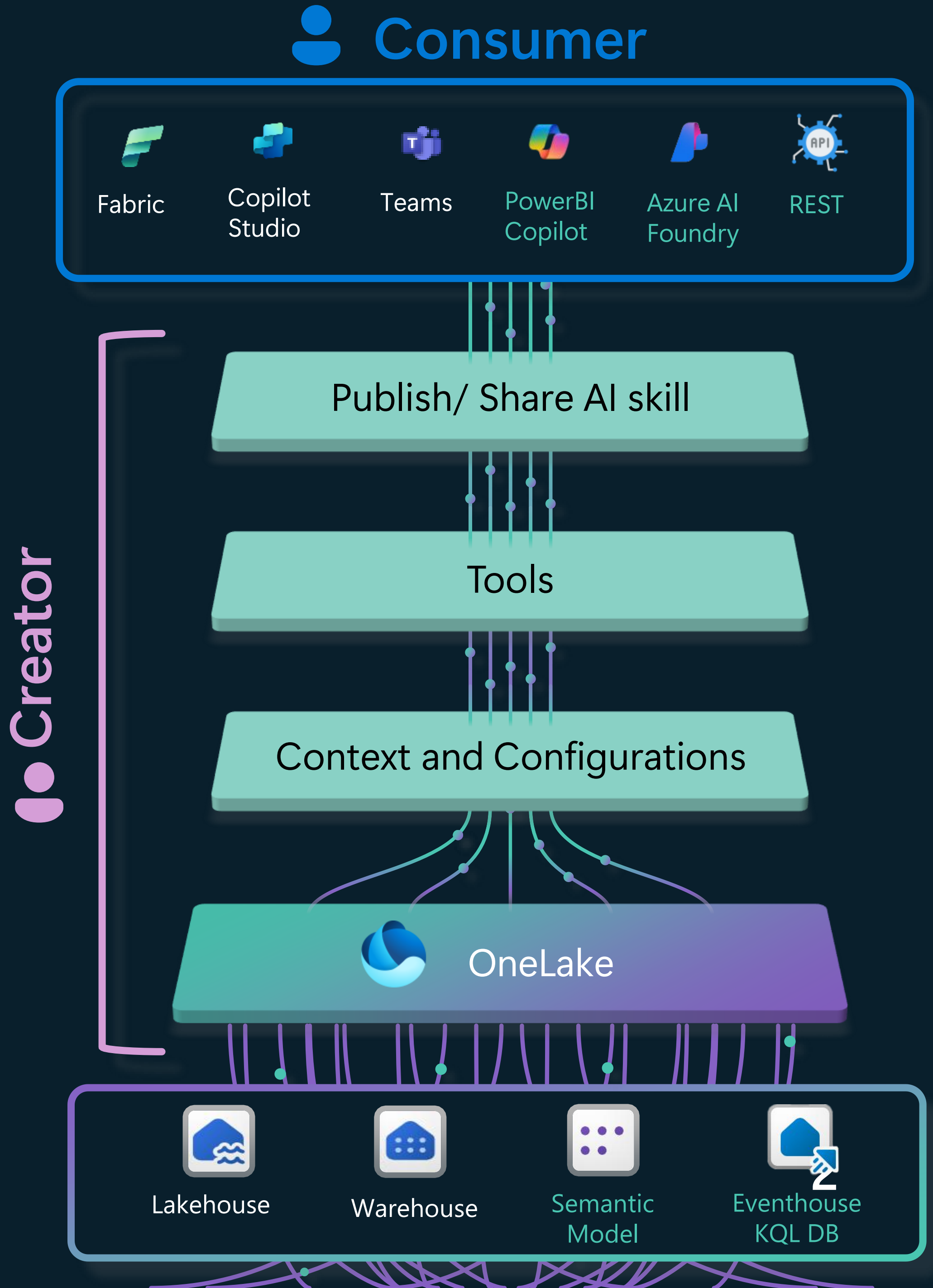
How does it compare to Q1 last year?

Ask a work question or make a request

# AI Skills are evolving

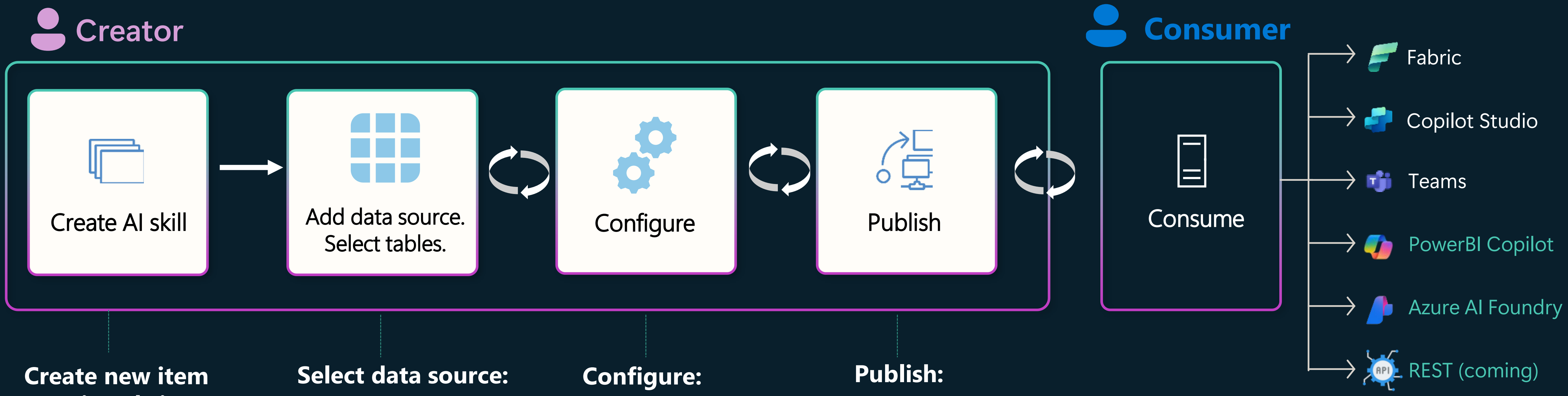
## A **Data Agent** for analyzing your data

- The conversational AI skill allows users to interact with data naturally, enhancing accessibility and usability.
- Improved chat canvas for creators with new debugging capabilities, making it easier to understand and refine responses.
- Seamlessly reason over multiple data sources, including semantic models and Eventhouse KQL databases, to create a powerful Data Expert tailored to your data domain.
- Your Data Agent can be consumed inside and outside of Fabric. Stay tuned for upcoming integrations with Copilot Studio, Teams, Azure AI Foundry and your own custom applications.



# AI skill end-to-end flow

End-to-end scenario



**Create new item type in Fabric:**

Track lineage

Manage permissions

Share

**Select data source:**

Lakehouse

Warehouse

Semantic model (Coming)

Eventhouse KQL DB (Coming)

Table selection

**Configure:**

Instructions for AI

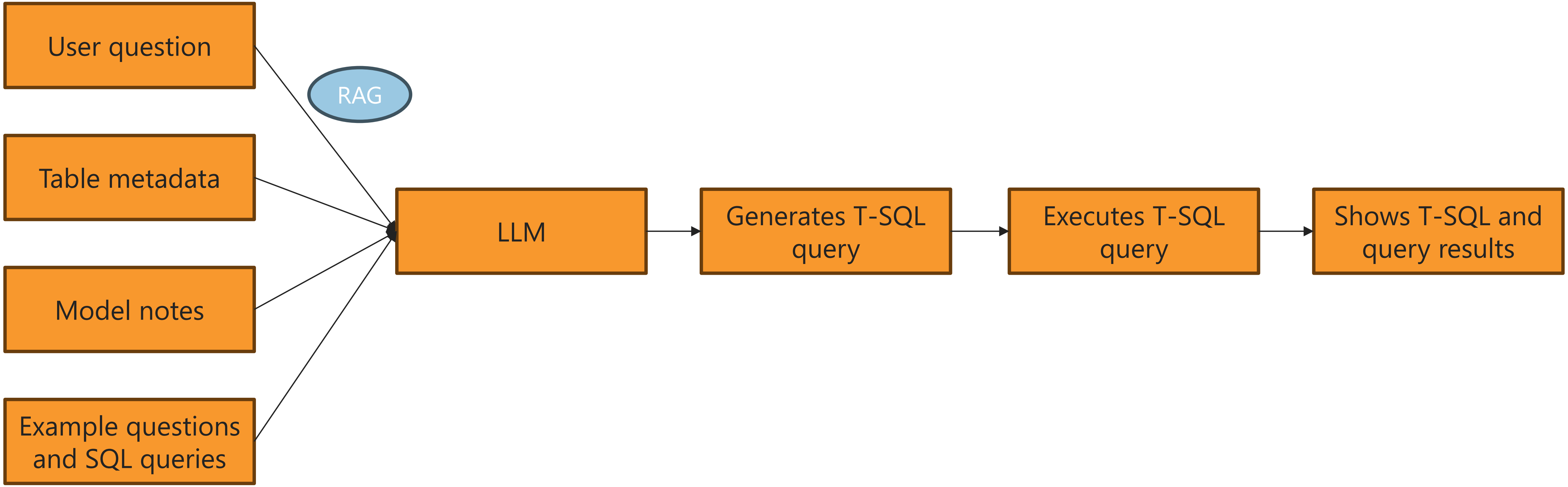
Few shot examples

**Publish:**

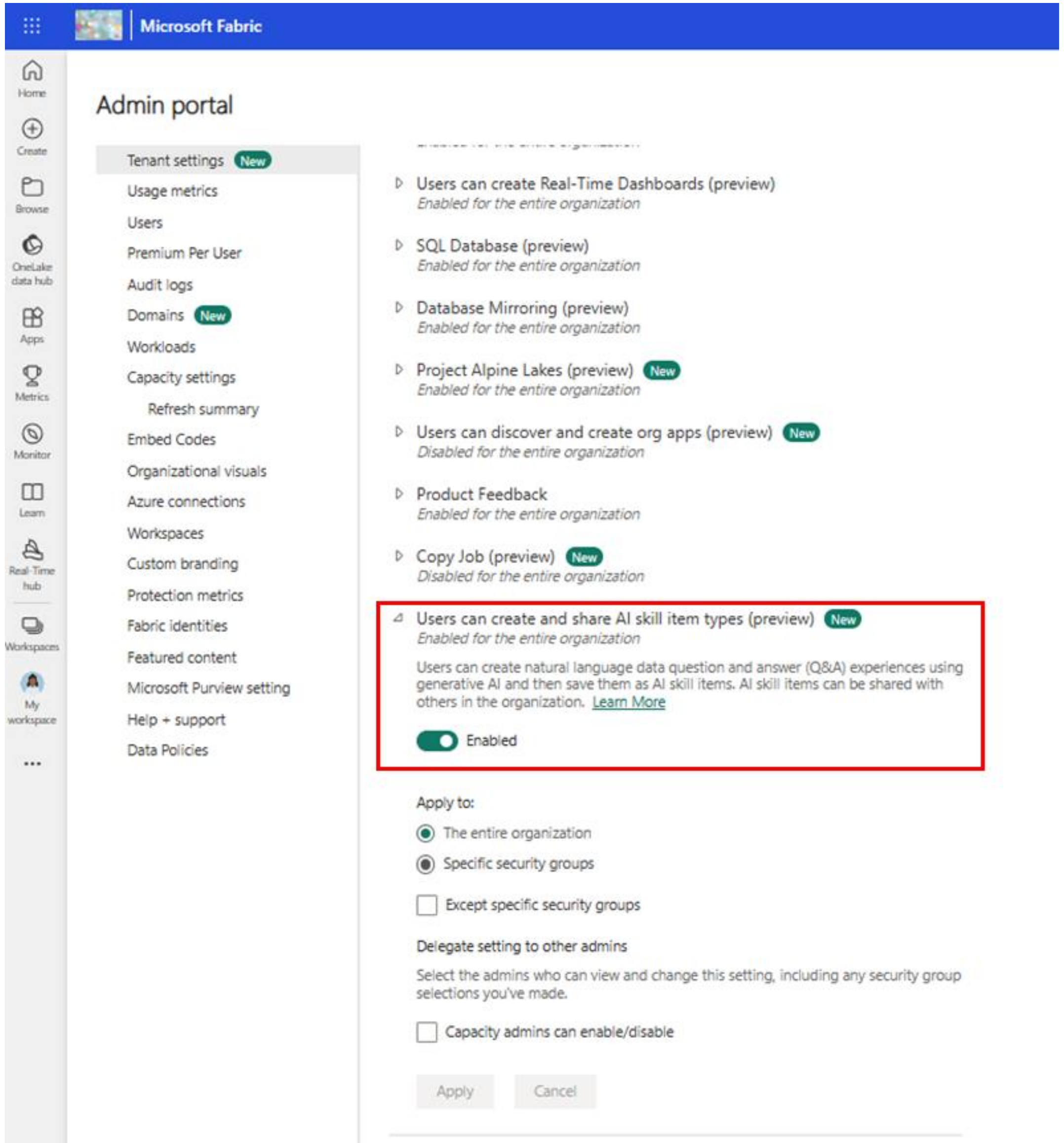
Creates a published version

Can be shared with consumers

# Microsoft Fabric AI Skill – model flow



# AI skill tenant setting



## Admin portal

- Users can create Real-Time Dashboards (preview)
  *Enabled for the entire organization*
- SQL Database (preview)
  *Enabled for the entire organization*
- Database Mirroring (preview)
  *Enabled for the entire organization*
- Project Alpine Lakes (preview) New
  *Enabled for the entire organization*
- Users can discover and create org apps (preview) New
  *Disabled for the entire organization*
- Product Feedback
  *Enabled for the entire organization*
- Copy Job (preview) New
  *Disabled for the entire organization*
- Users can create and share AI skill item types (preview) New
  *Enabled for the entire organization*

Users can create natural language data question and answer (Q&A) experiences using generative AI and then save them as AI skill items. AI skill items can be shared with others in the organization. Learn More

● Enabled

Apply to:

● The entire organization

● Specific security groups

□ Except specific security groups

Delegate setting to other admins

Select the admins who can view and change this setting, including any security group selections you've made.

□ Capacity admins can enable/disable

Apply Cancel

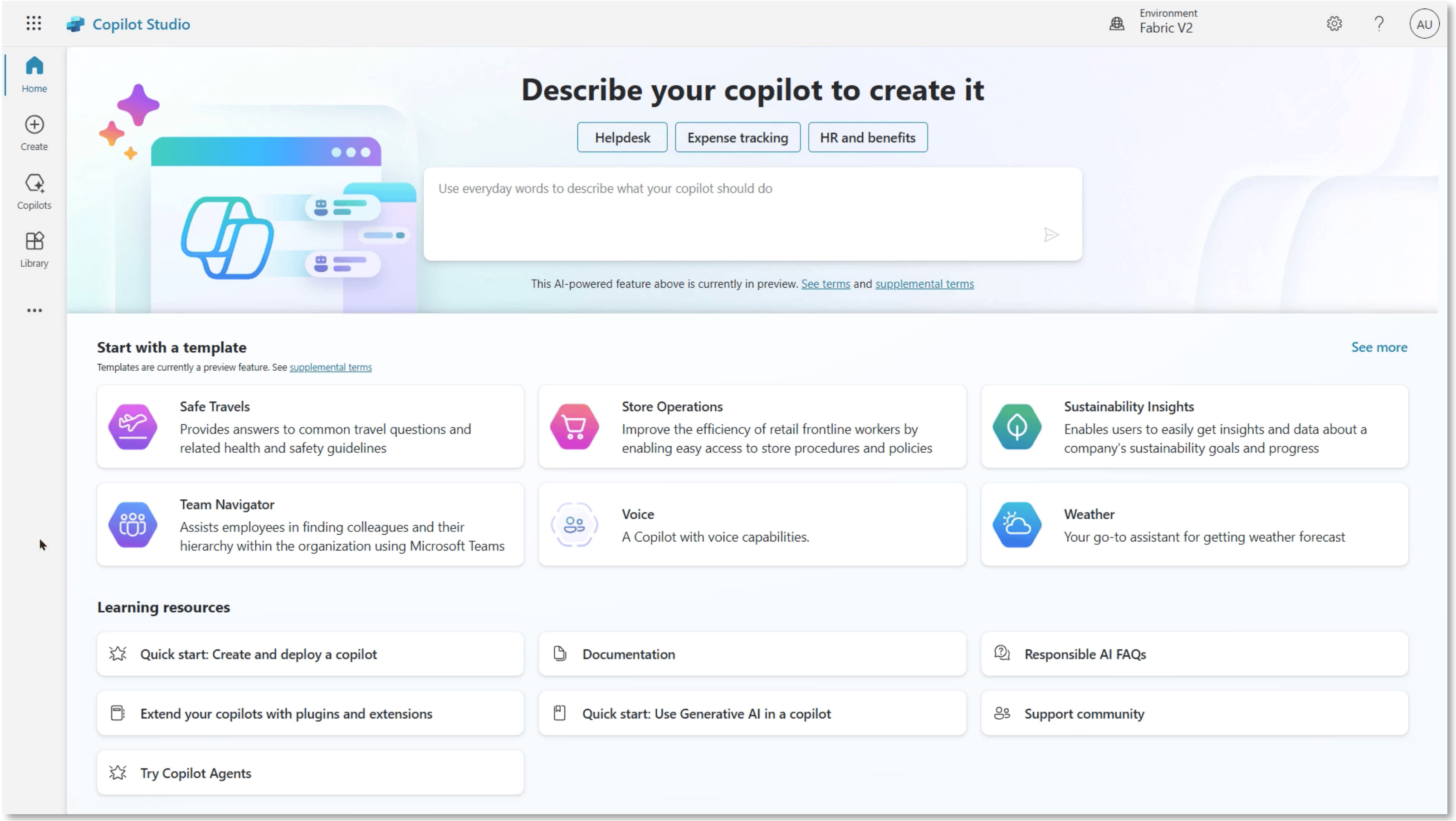
# Microsoft Fabric AI Skill demo on structured data

https://learn.microsoft.com/en-us/fabric/data-science/ai-skill-scenario

Private Preview

# Integration with Microsoft Copilot Studio

- Create custom agent in MCS with AI Skill as Knowledge Source
- Enhance your custom agent with different Knowledge Sources and Actions
- Publish your custom agent to Microsoft Teams



## Describe your copilot to create it

Helpdesk | Expense tracking | HR and benefits

Use everyday words to describe what your copilot should do

This AI-powered feature above is currently in preview. See terms and supplemental terms

**Start with a template** — See more

Templates are currently a preview feature. See supplemental terms

- Safe Travels — Provides answers to common travel questions and related health and safety guidelines
- Store Operations — Improve the efficiency of retail frontline workers by enabling easy access to store procedures and policies
- Sustainability Insights — Enables users to easily get insights and data about a company's sustainability goals and progress
- Team Navigator — Assists employees in finding colleagues and their hierarchy within the organization using Microsoft Teams
- Voice — A Copilot with voice capabilities.
- Weather — Your go-to assistant for getting weather forecast

**Learning resources**

- Quick start: Create and deploy a copilot
- Documentation
- Responsible AI FAQs
- Extend your copilots with plugins and extensions
- Quick start: Use Generative AI in a copilot
- Support community
- Try Copilot Agents

# What is the difference between AI Skill and Copilot?

The technology behind the AI skill and Fabric Copilot is similar. They both use Generative AI to reason over data. However, they have some key differences:

- **Configuration:** With an AI skill, you can configure the AI to behave the way you need. You can provide it with instructions and examples that tune it to your specific use case. A Fabric Copilot doesn’t offer this configuration flexibility.
- **Use Case**: A Copilot can help you do your work in Fabric. It can help you generate Notebook code or Data Warehouse queries. In contrast, the AI skill operates independently, and you can eventually connect it to Microsoft Teams and other areas outside of Fabric.
- **Purpose**: Copilot on Power BI works against a report/model and updates the report, while AI skill is for ad-hoc queries that work against a lakehouse or warehouse and returns T-SQL and query results.

# Industry use cases

for using both structured and unstructured data

# Healthcare: Patient Treatment Optimization

- **Scenario**: A healthcare organization wants to improve patient outcomes by combining unstructured patient records, clinical notes, and structured data such as diagnostic codes and treatment history.

- **Approach**: Generative AI can analyze doctor notes, lab reports (unstructured data), and link them to patient demographics, medications, and treatments stored in relational databases (structured data) to suggest the most effective treatments.

- **Business Decision**: Optimizes treatment plans, improves patient outcomes, and reduces the time needed to identify suitable care paths.

# Retail: Personalized Marketing

- **Scenario**: A retail company wants to enhance its customer targeting for upcoming campaigns.

- **Approach**: The company combines historical purchase data (structured data) from SQL databases with customer feedback, product reviews, and social media sentiment (unstructured data) to generate personalized product recommendations.

- **Business Decision**: Increases customer engagement and drives revenue through targeted promotions.

# Finance: Fraud Detection

- **Scenario**: A bank wants to improve its ability to detect fraudulent transactions by combining transactional data with suspicious customer communications.
- **Approach**: Generative AI processes emails, call transcripts, and legal documents (unstructured) alongside transactional records, financial logs, and account details (structured data) to identify patterns of potential fraud.
- **Business Decision**: Enhances fraud detection accuracy, reduces false positives, and minimizes financial risk.

# NFL Football: Game Strategy and Injury Risk Assessment

- **Scenario**: An NFL coaching staff wants to improve game strategy and reduce injury risks by analyzing both game data and medical reports.

- **Approach**: Structured data could include player stats (e.g., passing yards, rushing attempts, completion percentages), play-by-play game data, and historical win/loss records stored in relational databases. Unstructured data would include medical reports, injury rehabilitation progress notes, player interviews, and social media activity that may indicate player morale or readiness to play. Generative AI can combine both data types to generate recommendations for game-day decisions, such as which players to rest, which plays to run based on the opponent's weaknesses, and which players are at risk of re-injury.

- **Business Decision**: AI-generated insights can suggest optimal game strategies by evaluating player conditions, predicting fatigue or injury risk, and highlighting how certain players perform against specific types of defenses. This helps the coaching staff make more precise decisions regarding lineups, in-game play calling, and player usage, potentially leading to better team performance and fewer injuries.

# NFL Football: Play Calling and Defensive Adjustments

- **Scenario**: An NFL team wants to improve play-calling decisions during the game by dynamically adjusting based on the opposing team’s formation and tendencies.

- **Approach**: Structured data includes real-time stats (down, distance, time left, score), formation data, historical play success rates, and player matchups, which are typically stored in databases.  Unstructured data includes video feeds, playbook annotations, defensive alignments, and real-time comments from coaches and analysts about player performance or situational cues.  Generative AI combines real-time game data with historical trends to suggest optimal play calls based on the current game scenario. For example, it might recommend a specific blitz when the opponent is in a known pass-heavy formation on third down, or suggest a specific offensive play based on the defensive alignment.

- **Business Decision**: The AI can help the coaching staff decide in real-time whether to run, pass, or kick, and what kind of formation or play to use based on factors like down and distance, current score, and opposing defensive tendencies.  It can also suggest defensive adjustments, such as double-teaming a star receiver or adjusting coverage based on how an opposing quarterback is performing under pressure that day.

# Data query/analysis architecture approaches

- Traditional: Pulling numbers and text from documents (unstructured data) and putting it into a database (structured data) that includes other structured data and using SQL to query the data
  - Best when know what questions to ask
  - Returns consistent and accurate results
  - Use Azure AI Document Intelligence to pull numbers and text from documents and then use LLMs to convert from JSON to CSV to more easily add to a database
  - Can use LLMs to take English questions and convert to SQL queries on data (Microsoft Fabric AI Skill)
  - Cons: not using LLM's to improve answers to questions, just SQL queries; not accurately pulling data from documents
  - Ideal for financial reporting or operational dashboards where accuracy and consistency are critical
  - Document → Database → SQL Query → Answer
  - Summarize: Best for accuracy and predefined queries
- GenAI: Feeding all types of data (structured, semi-structured, and unstructured) into an LLM and using a bot to ask questions on the data
  - Best when not sure what questions to ask
  - Can return inconsistent and inaccurate results
  - Bot can allow unstructured data (documents) and semi-structured data (csv files) to be uploaded
  - Structured data can be exported to a csv and uploaded, or bot can be made to pull structured data from database automatically
  - Cons: need to make sure csv data has a lot of metadata; not accurately pulling data from documents if using RAG
  - Useful for customer support bots or summarizing and analyzing large document repositories
  - Document/CSV/Database → LLM → Bot Query → Answer
  - Summarize: Best for exploration and leveraging many types of data

# Helpful blogs

Microsoft Fabric AI Skill | James Serra's Blog

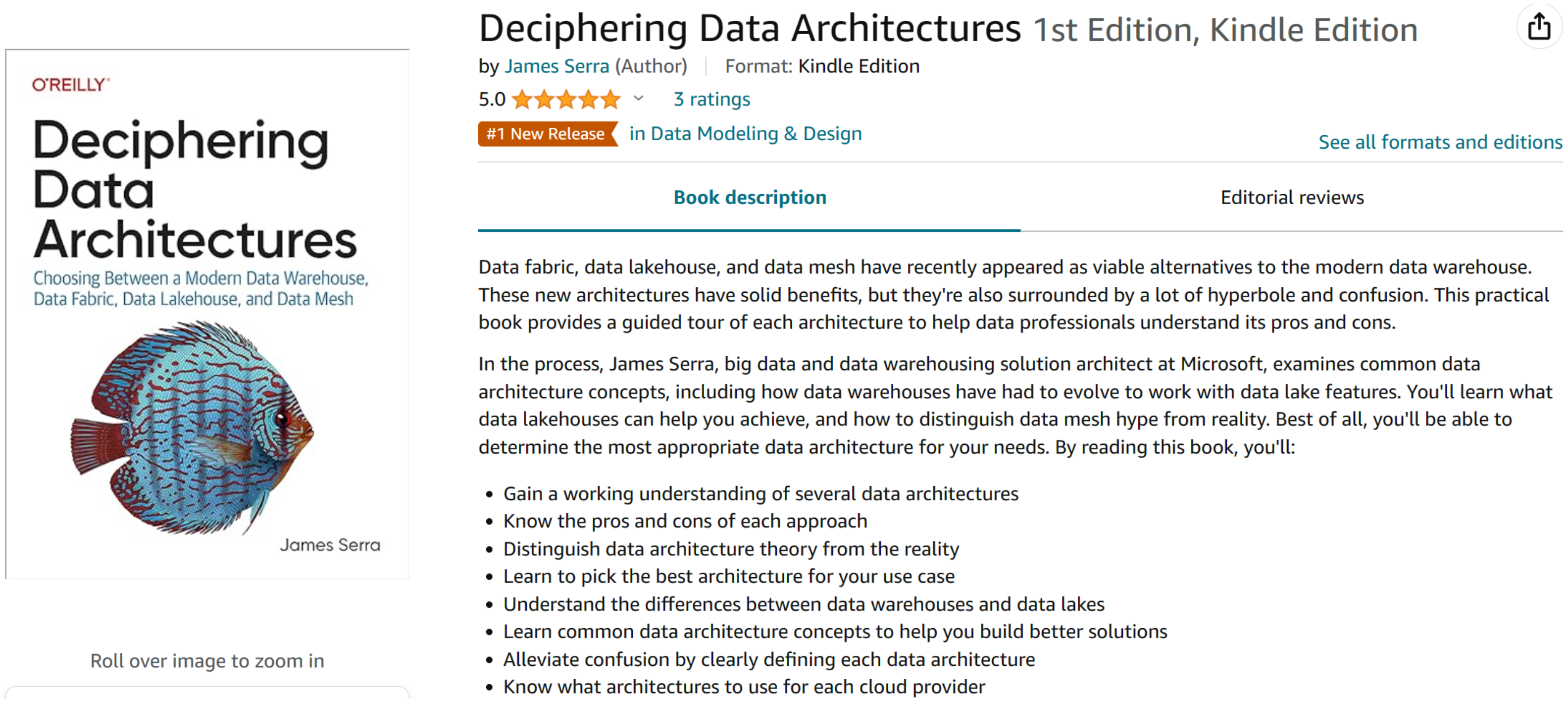
Introduction to OpenAI and LLMs | James Serra's Blog

Introduction to OpenAI and LLMs – Part 2 | James Serra's Blog

Introduction to OpenAI and LLMs – Part 3 | James Serra's Blog

Copilot in Microsoft Fabric | James Serra's Blog

# My book


O'REILLY
Deciphering Data Architectures
Choosing Between a Modern Data Warehouse, Data Fabric, Data Lakehouse, and Data Mesh
James Serra
Roll over image to zoom in

**Deciphering Data Architectures** 1st Edition, Kindle Edition

by James Serra (Author) | Format: Kindle Edition

5.0 ★★★★★ 3 ratings

#1 New Release in Data Modeling & Design

See all formats and editions

**Book description** | Editorial reviews

Data fabric, data lakehouse, and data mesh have recently appeared as viable alternatives to the modern data warehouse. These new architectures have solid benefits, but they're also surrounded by a lot of hyperbole and confusion. This practical book provides a guided tour of each architecture to help data professionals understand its pros and cons.

In the process, James Serra, big data and data warehousing solution architect at Microsoft, examines common data architecture concepts, including how data warehouses have had to evolve to work with data lake features. You'll learn what data lakehouses can help you achieve, and how to distinguish data mesh hype from reality. Best of all, you'll be able to determine the most appropriate data architecture for your needs. By reading this book, you'll:

- Gain a working understanding of several data architectures
- Know the pros and cons of each approach
- Distinguish data architecture theory from the reality
- Learn to pick the best architecture for your use case
- Understand the differences between data warehouses and data lakes
- Learn common data architecture concepts to help you build better solutions
- Alleviate confusion by clearly defining each data architecture
- Know what architectures to use for each cloud provider

Read entire book now with an O'Reilly subscription:
Deciphering Data Architectures (O'Reilly.com)

- Foundation
  - 1. Big Data
  - 2. Types of Data Architectures
  - 3. The Architecture Design Session
- Common Data Architecture Concepts
  - 4. The Relational Data Warehouse
  - 5. Data Lake
  - 6. Data Storage Solutions and Processes
  - 7. Approaches to Design
  - 8. Approaches to Data Modeling
  - 9. Approaches to Data Ingestion
- Data Architectures
  - 10. The Modern Data Warehouse
  - 11. Data Fabric
  - 12. Data Lakehouse
  - 13. Data Mesh Foundation
  - 14. Should You Adopt Data Mesh? Myths, Concerns, And The Future
- People, Process, and Technology
  - 15. People And Processes
  - 16. Technologies

More details

jamesserra3@gmail.com

Order on Amazon!

# Q & A

James Serra, Microsoft, Data & AI Solution Architect
Email me at: jamesserra3@gmail.com
Follow me at: @JamesSerra
Link to me at: www.linkedin.com/in/JamesSerra
Visit my blog at: JamesSerra.com

***(I'll put this deck in the chat or via email)***