

October 30, 2024

11 – 11:45 AM EDT





### Peter Carson

- President, Extranet User Manager and Envision IT
- 14-time Microsoft M365 MVP
- <u>pcarson@envisionit.com</u>
- blog.petercarson.ca
- President, Toronto Microsoft 365User Group



#### About Us

- Established in 2008 in Toronto, Canada area
- Brand of software products owned by Envision IT Inc.
- Leader in the Microsoft 365 space and a Microsoft Partner for 10+ years



#### We Make Microsoft Simple

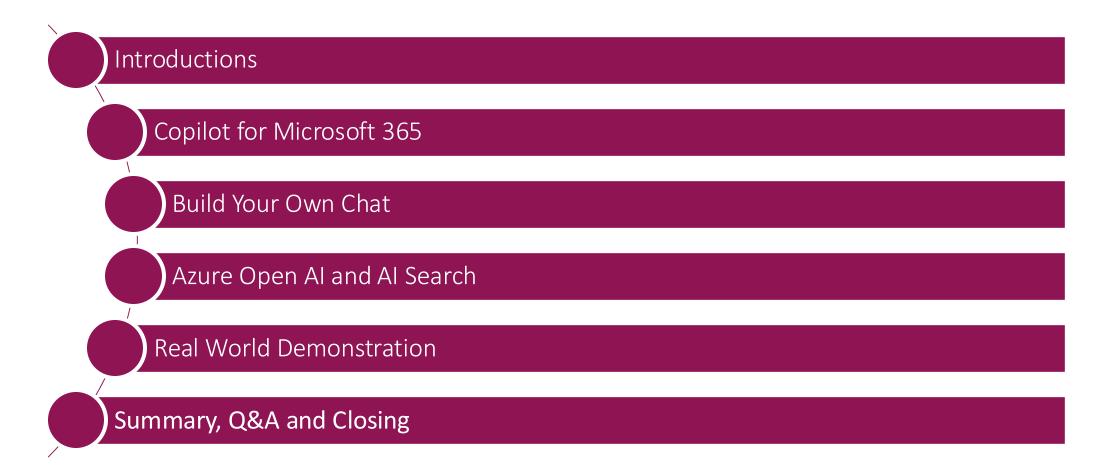
- We make Microsoft simple for organizations to connect with their external stakeholders
- An exceptional user experience lets businesses collaborate with external stakeholders and maintain Microsoft security.
- IT is involved at the beginning, ensuring a seamless integration, while the business can focus on what really matters getting the job done.

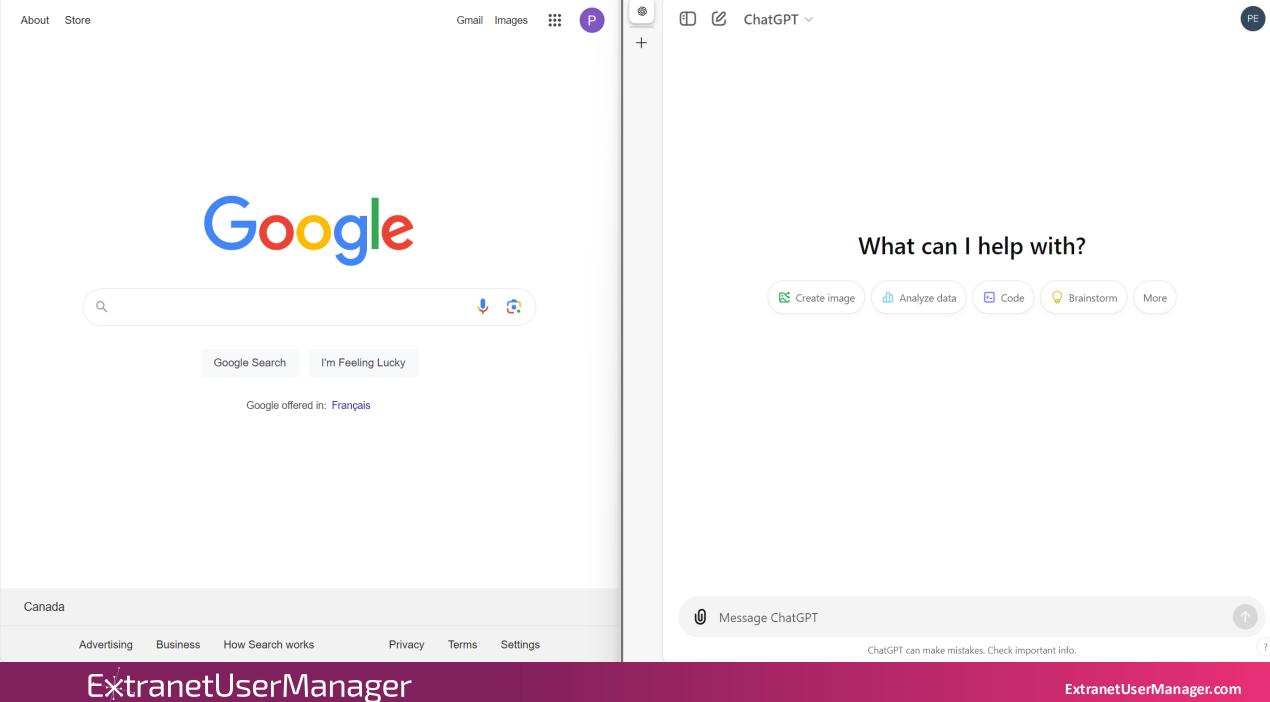
# Imagine this...

You're a hotel guest and you want restaurant recommendations for dinner tonight, so you walk up to the hotel concierge to ask for their restaurant recommendations.

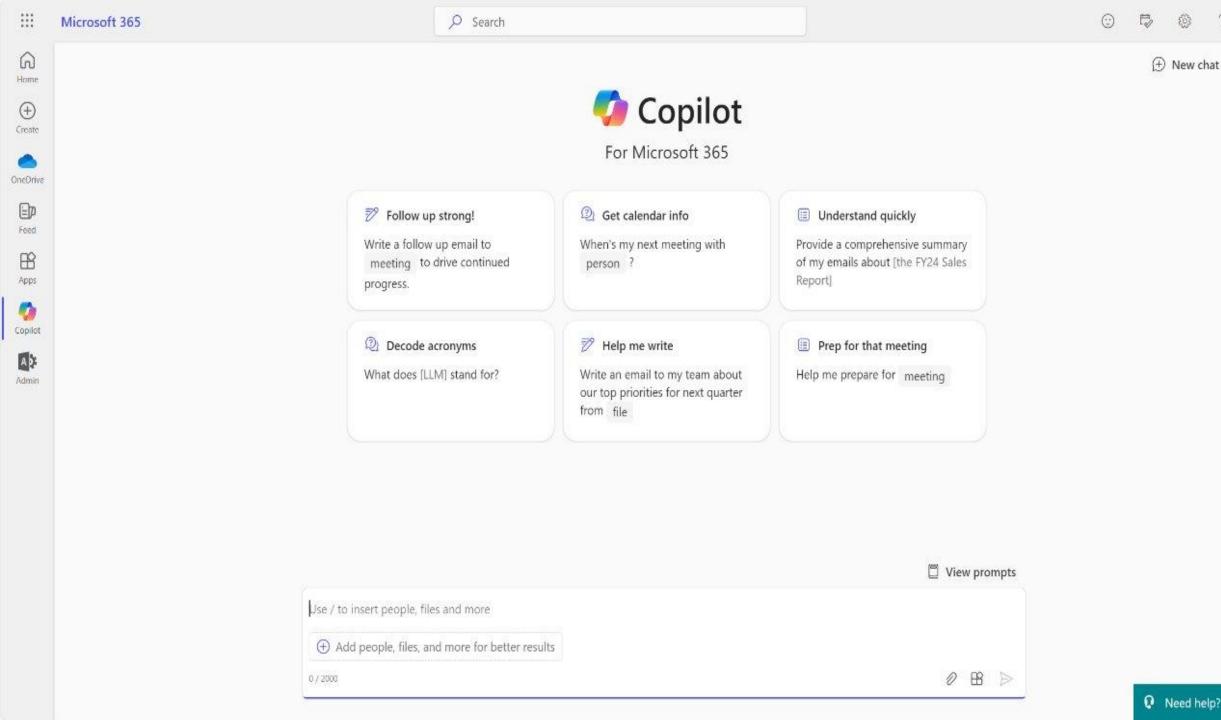


### Agenda



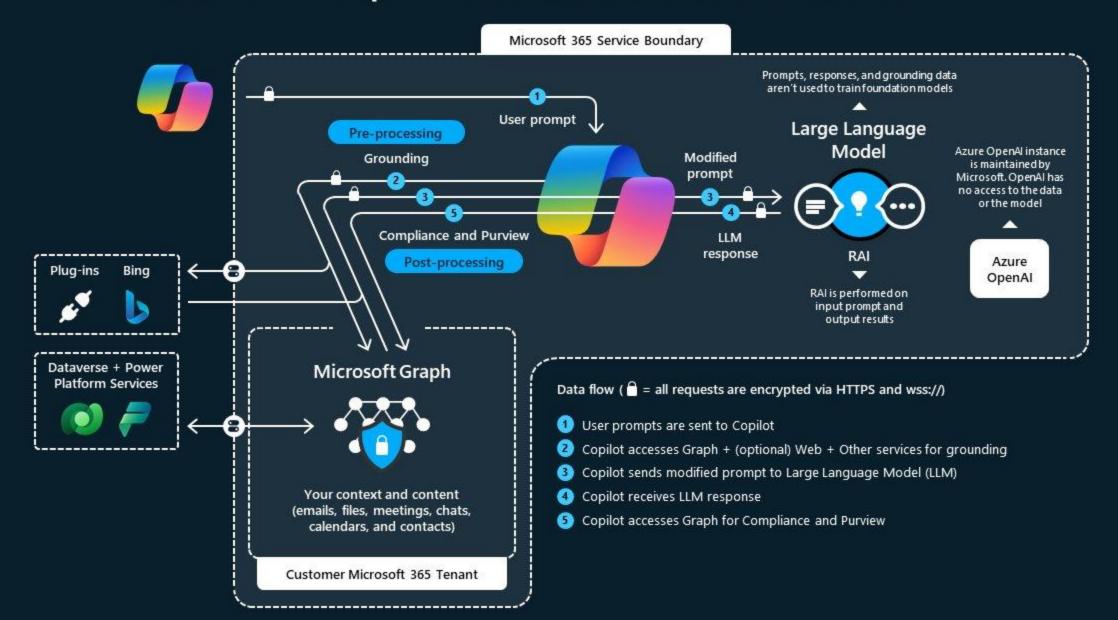


# Copilot





### Microsoft Copilot for Microsoft 365 architecture



### **Azure OpenAl**

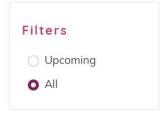


- Azure OpenAl Service Large Language Models for Generative Al (microsoft.com)
- Azure OpenAl service features offers REST API to a wide variety of models
- Different language models co-develop their APIs to help ensure compatibility and easy transitions
- Azure OpenAl ensures the same security standards as Microsoft Azure
- Azure OpenAl offers private networking, regional availability, and responsible Al content filtering

Q

Home > Resources > Events

#### **Events**





#### Microsoft Al Tour in Toronto

Dec 3, 2024 © 7:15 a.m. - 6:15 p.m.

We are looking forward to demonstrating how Al can be used to provide simplified and streamlined access to data for guest users. Stay tuned for more details about our participation.



#### **ESPC 2024**

The Extranet User Manager team is excited to be returning to ESPC 2024 as a Silver Sponsor. Join us for three days of insightful sessions, workshops, and networking with IT professionals and M365 enthusiasts.



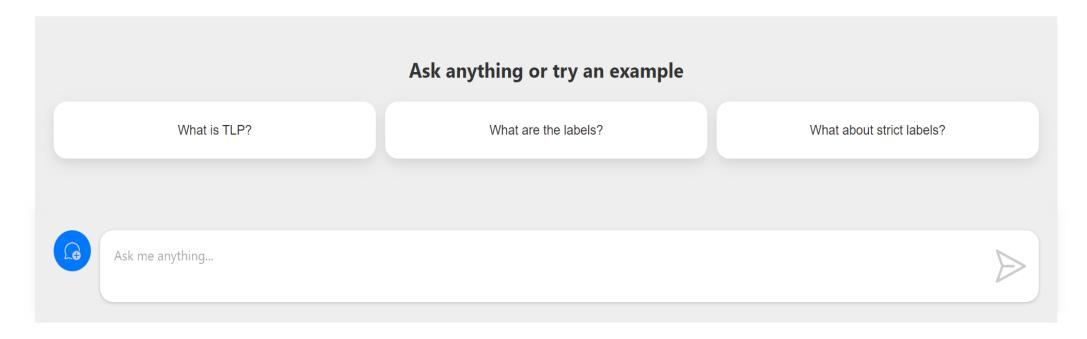
#### TechCon 365

Nov 11, 2024 - Nov 15, 2024 © 9:00 a.m. - 5:00 p.m.



Home > Demo > Chat

#### Chat



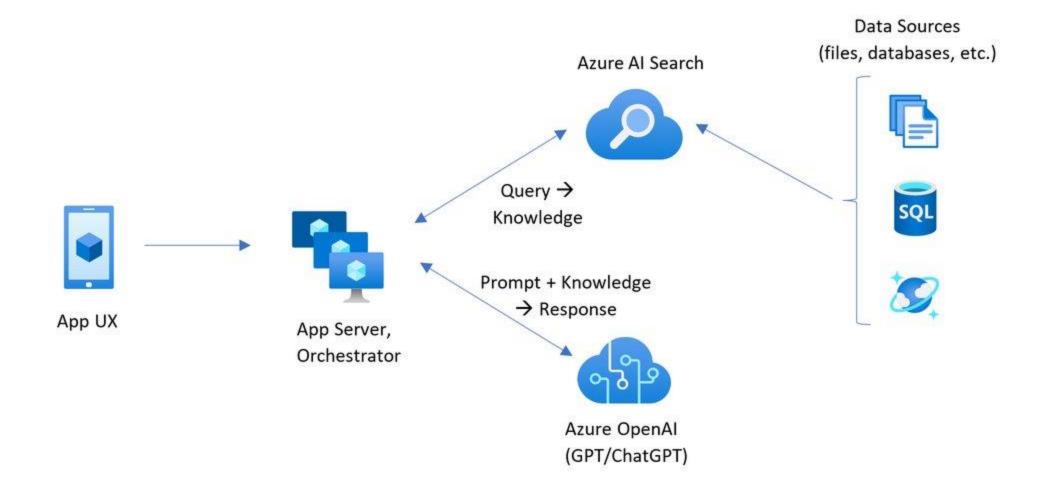
© Extranet Portal 2024 · Privacy · Terms of Use

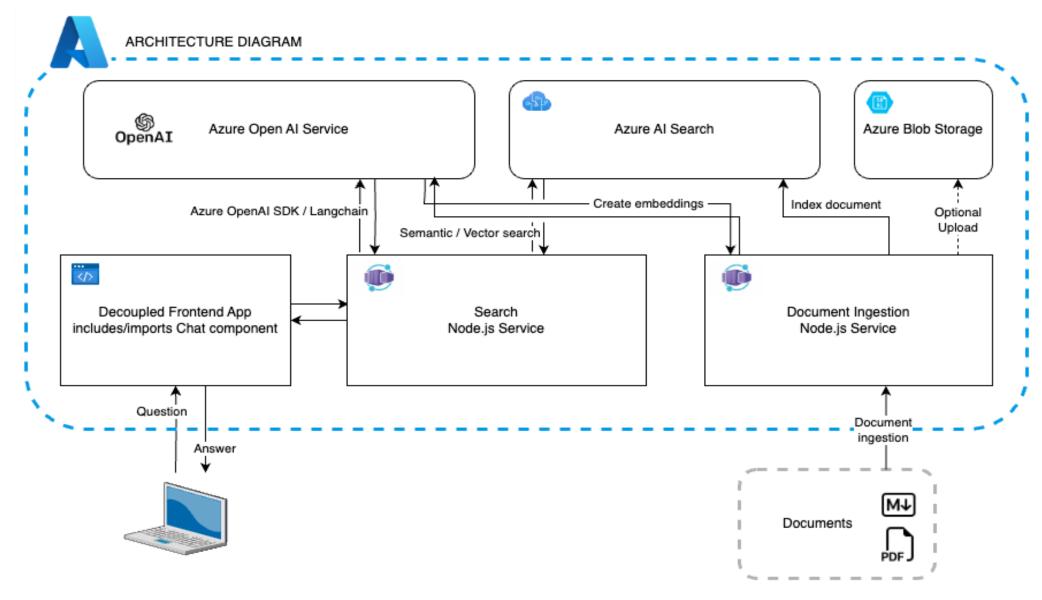
### **RAG Retrieval-Augmented Generation**



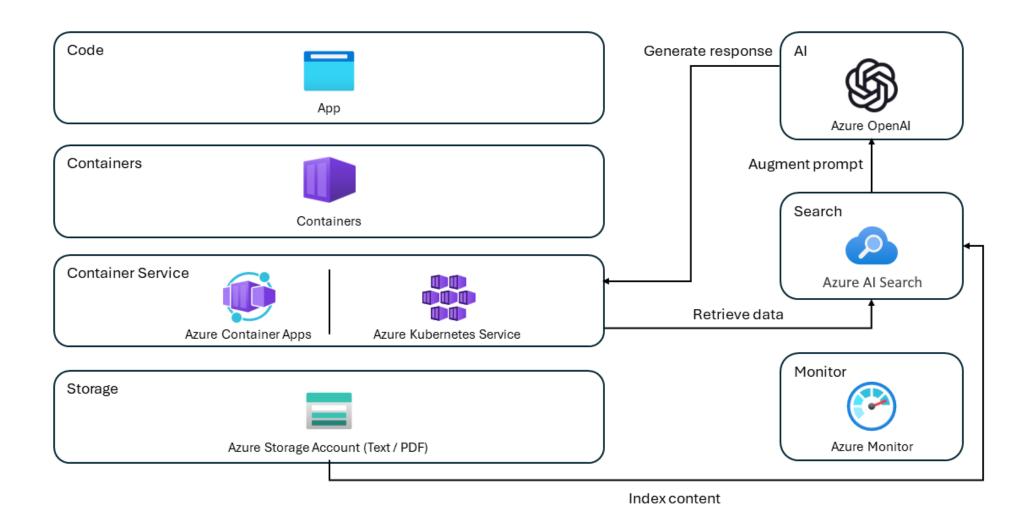


### **Architectural Overview**





Get started with the chat using your own data sample for JavaScript - JavaScript on Azure | Microsoft Learn



Get started with the chat using your own data sample for .NET - .NET | Microsoft Learn

# Search Q

# Transitioning from Keyword to Semantic and Vector Search



**Keyword Search:** Direct matching of query terms to terms in documents or databases.

**Use Cases:** Early search engines, basic database queries.

Advantages: Simple, quick, easy to implement.

#### **Limitations:**

- Requires exact term matches
- Can yield irrelevant results if terms are out of context
- Can miss relevant documents if they don't use the exact terms

#### **Semantic Search**

Understands the meaning and intent behind user queries, rather than just matching keywords.

Use Cases: modern search engines, CMS, ecommerce recommendations.

#### Advantages:

- Relevant results
- Can answer queries instead of just returning files
- User-friendly with natural language queries

#### **Limitations:**

- Requires more sophisticated algorithms
- Potential for ambiguity in interpretation

#### **Vector Search**

Finds items in a dataset based on their similarity in meaning or context, rather than exact keyword matches.

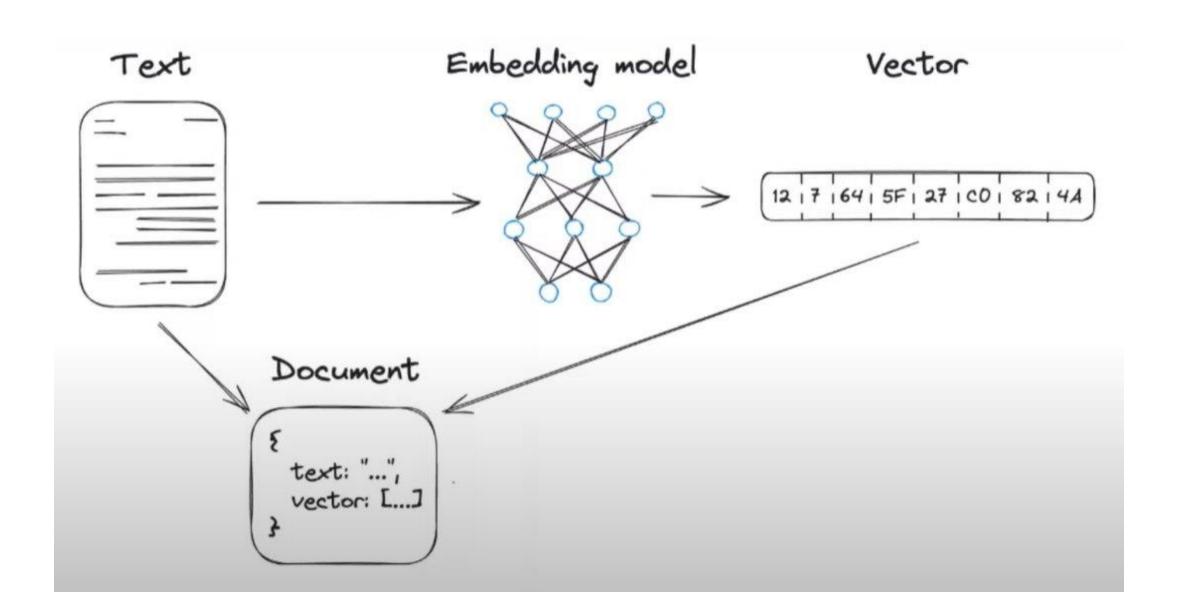
Use Cases: advanced data retrieval systems, document clustering, and topic modelling.

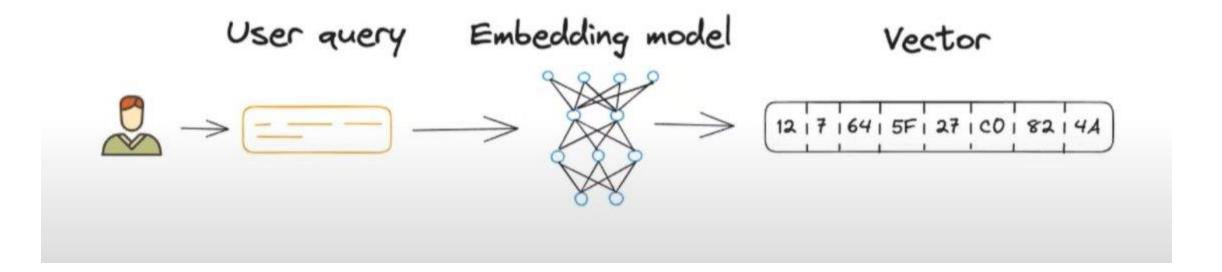
#### Advantages:

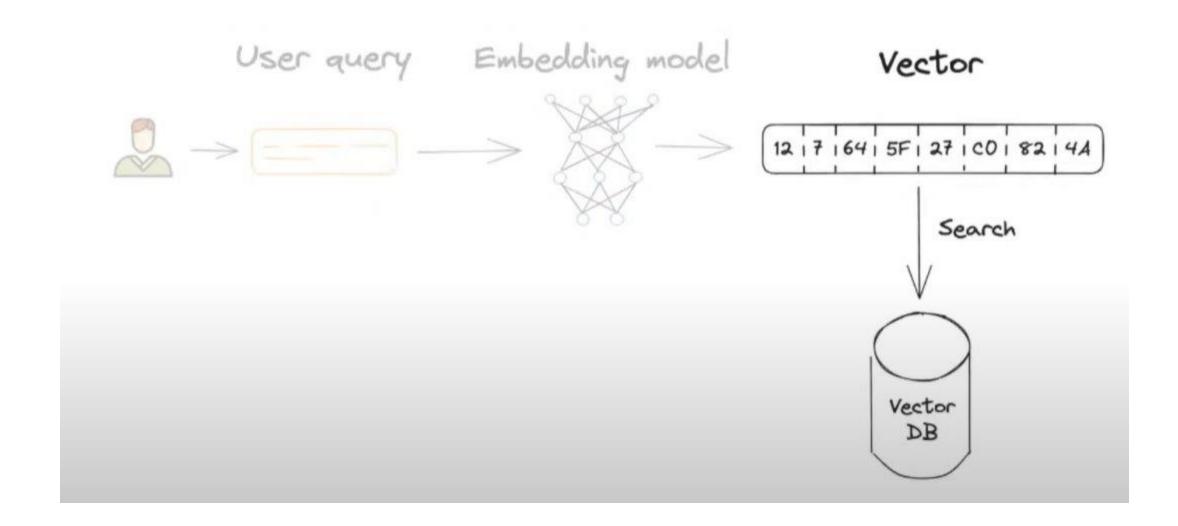
- Relevant results
- Understands context, relationships between words
- Supports languages and terms with sparse data

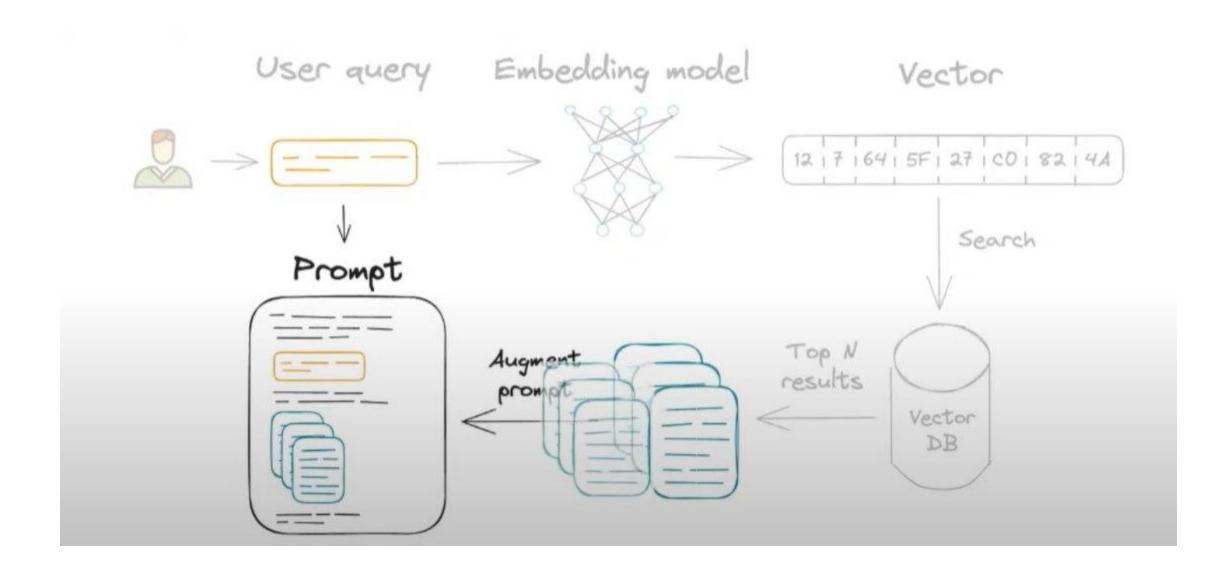
#### **Limitations:**

- Requires substantial computational power
- Vector models need training and refining









### Azure Al Search

- Cloud service providing a full-text search engine, offering semantic and vector search capabilities
- Intended for developers needing application/based search experiences
- Benefits from vast data sources and Al-enhanced indexing



### **Azure OpenAl Pricing**

#### Azure OpenAl Service - Pricing | Microsoft Azure

#### Priced and scaled by tokens and context

- 1 token ~= 4 chars in English
- 1 token ~= 34 words
- 100 tokens ~= 75 words

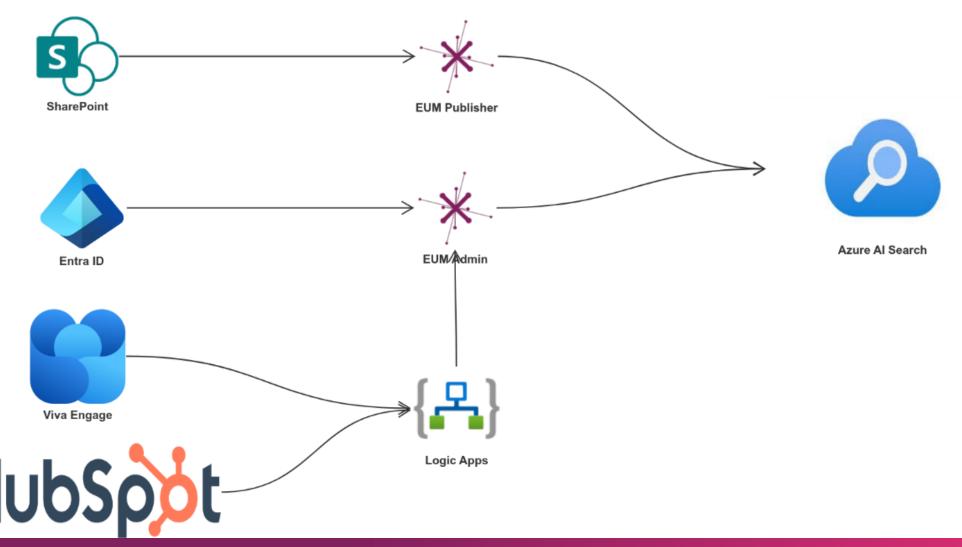
| Models                 | Context | Input (Per 1,000 tokens) | Output (Per 1,000 tokens) |
|------------------------|---------|--------------------------|---------------------------|
| GPT-3.5-Turbo-0125     | 16K     | \$0.0007                 | \$0.0021                  |
| GPT-3.5-Turbo-Instruct | 4K      | \$0.0021                 | \$0.003                   |
| GPT-4                  | 8K      | \$0.042                  | \$0.083                   |
| GPT-4                  | 32K     | \$0.083                  | \$0.165                   |

### **Azure Al Search Pricing**

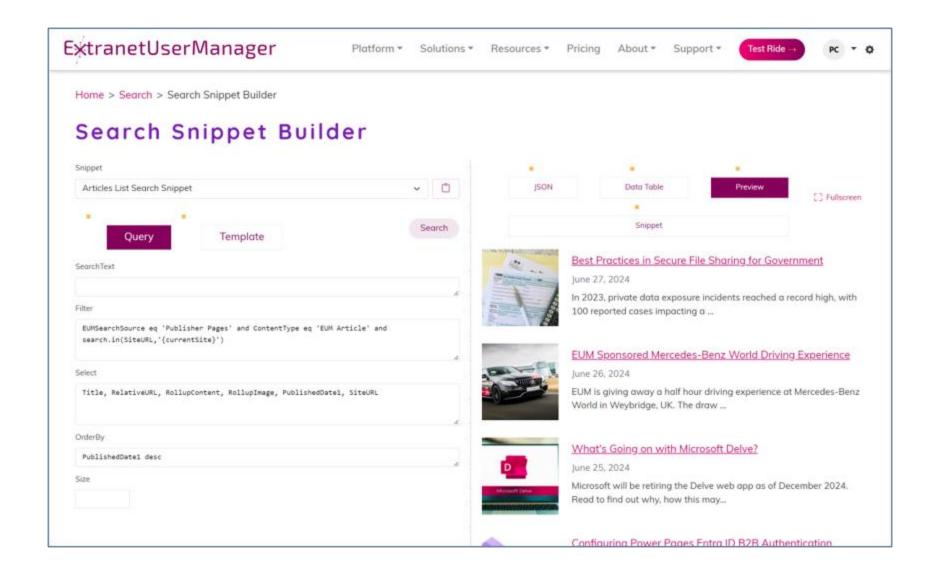
|                            | Free             | Basic   | Standard S1   | Standard S2   | Standard S3   | Storage Optimised<br>L1   | Storage Optimised<br>L2   |
|----------------------------|------------------|---|---|---|---|---|---|
| Storage <sup>1</sup>       | 50 MB            | 15 GB<br>(max 45 GB per<br>service)   | 160 GB<br>(max 1.9 TB per<br>service)                                   | 512 GB<br>(max 6 TB per<br>service)                                     | 1 TB<br>(max 12 TB per<br>service)  | 2 TB<br>(max 24 TB per<br>service)                                      | 4 TB<br>(max 48 TB per<br>service)                                      |
| Max indexes per<br>service | 3                | 15  | 50  | 200   | 200 or<br>1,000/partition in<br>high density mode   | 10  | 10  |
| Scale out limits           | N/A              | Up to 9 units per<br>service <sup>2</sup><br>(max 3 partition;<br>max 3 replicas) | Up to 36 units per<br>service<br>(max 12 partition;<br>max 12 replicas) | Up to 36 units per<br>service<br>(max 12 partition;<br>max 12 replicas) | Up to 36 units per<br>service<br>(max 12 partition;<br>max 12 replicas)<br>up to 3 partitions<br>in high density <sup>3</sup><br>mode | Up to 36 units per<br>service<br>(max 12 partition;<br>max 12 replicas) | Up to 36 units per<br>service<br>(max 12 partition;<br>max 12 replicas) |
| Price per SU (Scale        | <b>\$0</b> /hour | <b>\$0.14</b> /hour   | <b>\$0.47</b> /hour   | <b>\$1.85</b> /hour   | \$3.69/hour   | <b>\$5.26</b> /hour   | \$10.52/hour  |

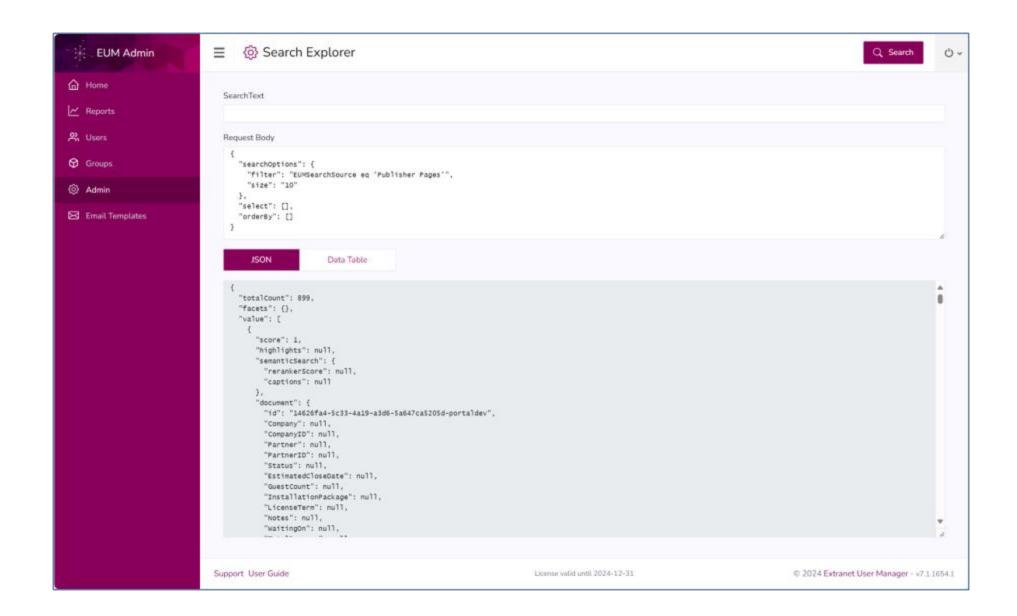
Unit)

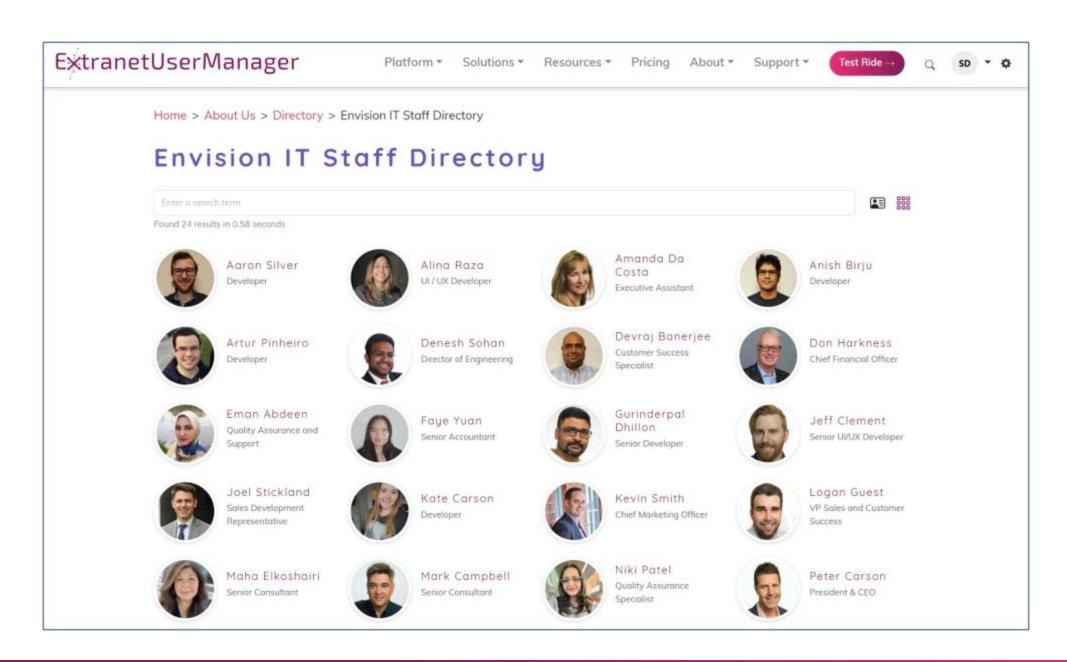
### **EUM AI Search Architecture**

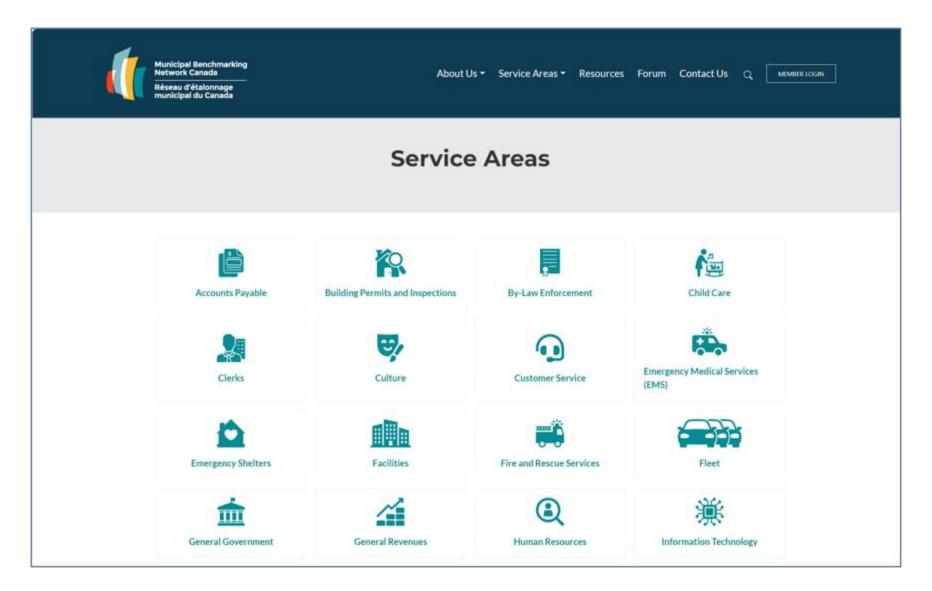


## **Azure Al Search Demo**

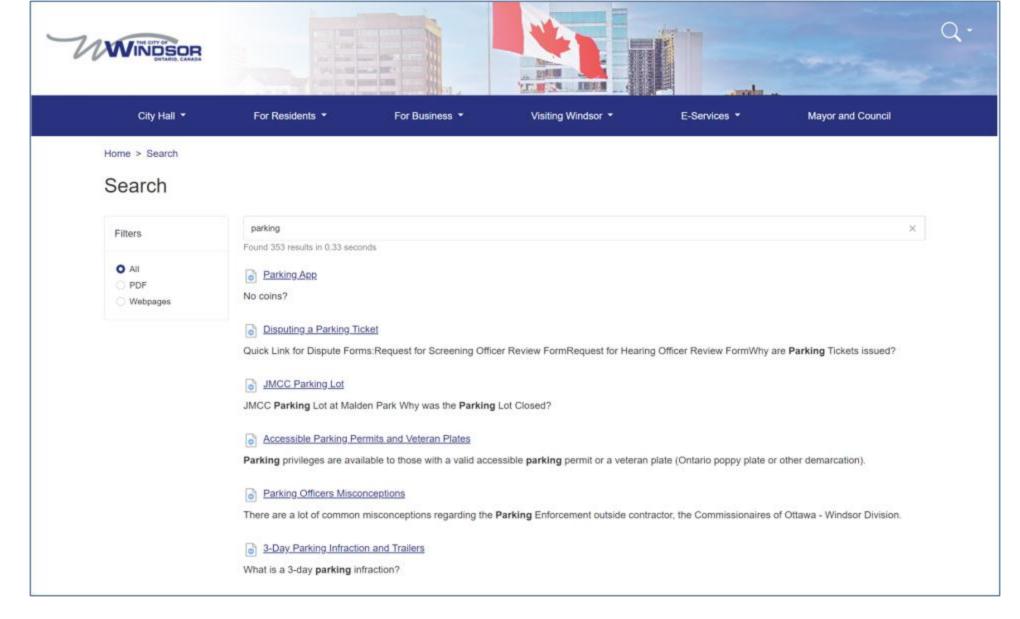




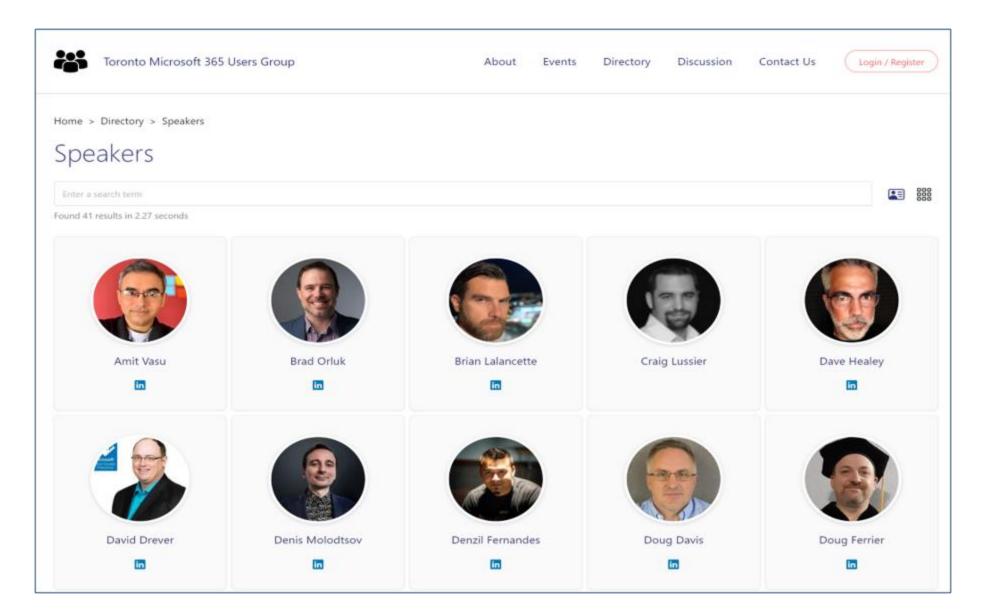




https://mbncanada.ca



https://www.citywindsor.ca



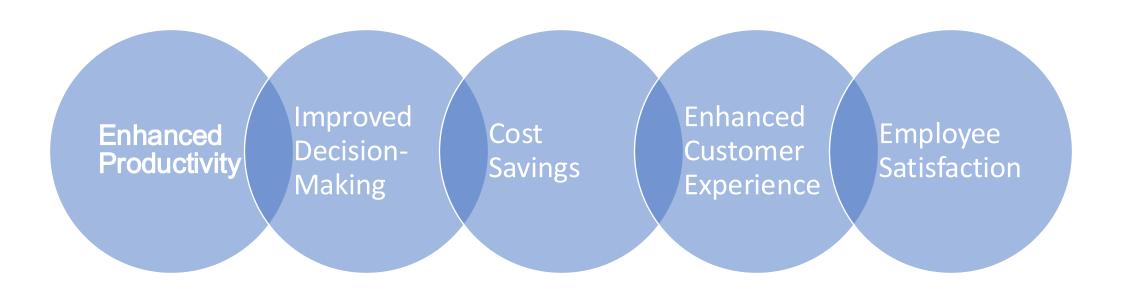
https://tm365.org

### **Key Points**

- Search is a key component of a chat experience
- Think about grounding and security
- There are many ways to do this within Microsoft
  - Copilot
  - Copilot Studio
  - Custom dev and Azure
- Azure Al Search is incredibly powerful

### **Summary of Key Benefits**





### Resources

- Al Community Conference Vancouver 2024 | Extranet User Manager
- Azure Al Search and Open Al Webinar | Extranet User Manager
- Get started with the chat using your own data sample for JavaScript JavaScript on Azure |
   Microsoft Learn
- Azure Al Search documentation | Microsoft Learn
- Simple query syntax Azure Al Search | Microsoft Learn
- Pricing Azure Al Search | Microsoft Azure
- Azure OpenAl Service Pricing | Microsoft Azure
- GitHub Azure-Samples/serverless-chat-langchainjs: Build your own serverless AI Chat with Retrieval-Augmented-Generation using LangChain.js, TypeScript and Azure
- Building RAG apps with LangchainJS YouTube

### **Upcoming Events**



TechCon 365 2024

Dallas, TX Nov 11-15



**ESPC 2024** 

Stockholm, Sweden
Dec 2-5



Announcing our Self-Serve Freemium M365 Tenant Dashboard

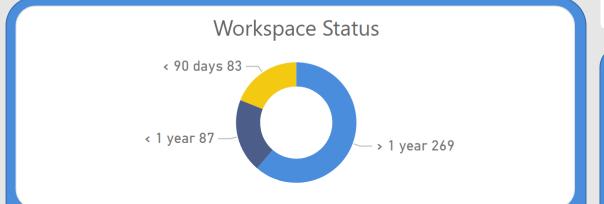
Online
November – Stay tuned!

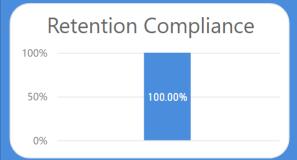
**Events | Extranet User Manager** 

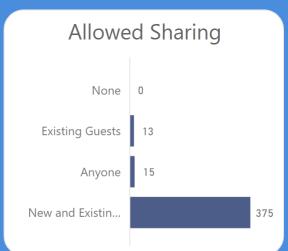
184
Teams

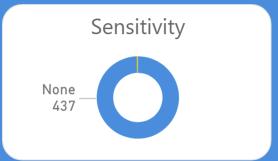


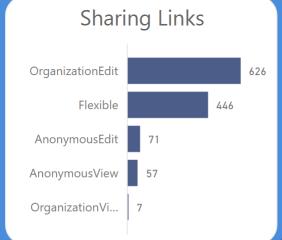












Go to drill-through

Reset filters

117,311

Count of Files

777.57

File Size (GB)

**753** 

Active storage used (GB)

387

Inactive storage used (GB)

# Thank you!

Any Questions?