



Search smarter in Generative AI apps

Global AI Bootcamp
2024

Abby Shen



Agenda



Retrieval-Augmented Generation (RAG)



Vectors and vector databases



Vector search



Data processing



Retrieval

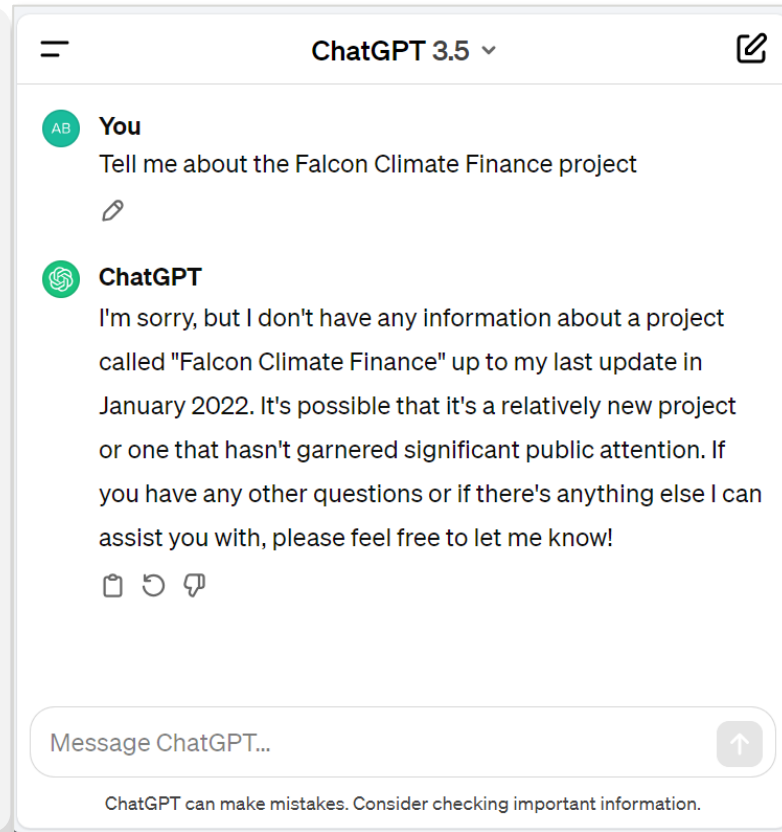
Retrieval-Augmented Generation (RAG)

What's the difference?



ChatGPT (free version)

Interacts directly with the OpenAI models



M365 Copilot

Grounded in your M365 data



Copilot

Tell me about the Falcon Clima...

+ New chat



February 23, 2024 at 03:46 PM

Tell me about the Falcon Climate Finance project

The Falcon Climate Finance Project is an innovative initiative aimed at leveraging financial instruments to combat climate change. The project seeks to mobilize private sector investment in sustainable and climate-resilient development. Its objectives are to promote sustainable investment, enhance climate resilience, and foster innovative solutions to reduce greenhouse gas emissions and promote clean energy. Some of the strategies being employed include the issuance of green bonds to finance renewable energy projects, the development of insurance products that protect against climate-related risks, and the implementation of mechanisms that put a price on carbon emissions. ¹ Is there anything else you would like to know about the project?

Copy

1 reference ^

¹ Falcon Climate Finance Project

General

Word | Abby Shen modified on 2/23/24

What are the top priorities this week?

According to an email sent a few minutes ago by Abby Shen, the project leader of the Falcon Climate Finance Project, the top priorities for this week are:

- **Green Bonds:** Finalize the structure of the first green bond issuance, with the goal of having a draft prospectus ready for review by the end of the week. ¹
- **Climate Risk Insurance:** Continue developing the climate risk insurance product, focusing on identifying key climate risks and designing appropriate coverage options. ¹
- **Carbon Pricing:** Begin researching potential mechanisms for implementing carbon pricing within the project scope, aiming to have a preliminary proposal ready for discussion at the next team meeting. ¹

Abby also reminded the team that their work is not just about financial returns, but also about making a tangible impact on the fight against climate change. ¹ Is there anything else you would like to know?

Copy

2/30 AI-generated content may be incorrect

1 reference ^

FCFR: 10628ms TuringBot: 21485ms

¹ Falcon Climate Finance Project - Weekly Priorities

General

Outlook | 2/23/24 | Abby Shen To Abby Shen

W

AutoSave

On

Falco...

G...

Saved

Abby Shen

File

Home

Insert

Draw

Design

Layout

Reference

Mailings

Review

View

Help

Editing

Falcon Climate Finance Project

Introduction

The **Falcon Climate Finance Project** is an innovative initiative aimed at leveraging financial instruments to combat climate change. This project seeks to mobilize private sector investment in sustainable and climate-resilient development.

Objectives

- Promote Sustainable Investment:** Encourage businesses and investors to shift towards more sustainable practices.
- Climate Resilience:** Support projects that enhance resilience to climate change, particularly in vulnerable communities.
- Innovation:** Foster innovative solutions to reduce greenhouse gas emissions and promote clean energy.

Strategies

- Green Bonds:** Issuance of green bonds to finance renewable energy projects, energy-efficient buildings, and other low-carbon infrastructure.
- Climate Risk Insurance:** Development of insurance products that protect against climate-related risks.
- Carbon Pricing:** Implementation of mechanisms that put a price on carbon emissions, encouraging businesses to reduce their carbon footprint.

Expected Outcomes

- Increased Investment in Green Technologies:** By providing attractive financial returns, the project aims to increase investment in green technologies.
- Reduced Carbon Emissions:** Through the promotion of clean energy and efficient practices, the project aims to contribute to the global effort to reduce carbon emissions.
- Enhanced Climate Resilience:** By supporting climate-resilient projects, the project aims to help communities better adapt to the impacts of climate change.

Conclusion

The Falcon Climate Finance Project represents a significant step forward in the fight against climate change. By harnessing the power of finance, it seeks to drive investment in sustainable practices and technologies, ultimately contributing to a greener and more resilient future.

Page 1 of 1

264 words

English (New Zealand)

Focus

100%

Falcon Climate Finance Project - Weekly Priorities

File

Message

Help

Falcon Climate Finance Project - Weekly Priorities

Summarize

Abby Shen

To

Abby Shen

General

3:44 pm

Dear Team,

I hope this message finds you well. As we continue to make strides in our Falcon Climate Finance Project, I wanted to take a moment to outline our top priorities for this week:

- Green Bonds:** Finalize the structure of our first green bond issuance. Our goal is to have a draft prospectus ready for review by the end of the week.
- Climate Risk Insurance:** Continue developing our climate risk insurance product. Focus on identifying key climate risks and designing appropriate coverage options.
- Carbon Pricing:** Begin researching potential mechanisms for implementing carbon pricing within our project scope. Aim to have a preliminary proposal ready for discussion at our next team meeting.

Remember, our work is not just about financial returns. It's about making a tangible impact on the fight against climate change. Each task we complete brings us one step closer to a more sustainable and resilient future.

Thank you for your hard work and dedication to this project. Let's make this week a productive one!

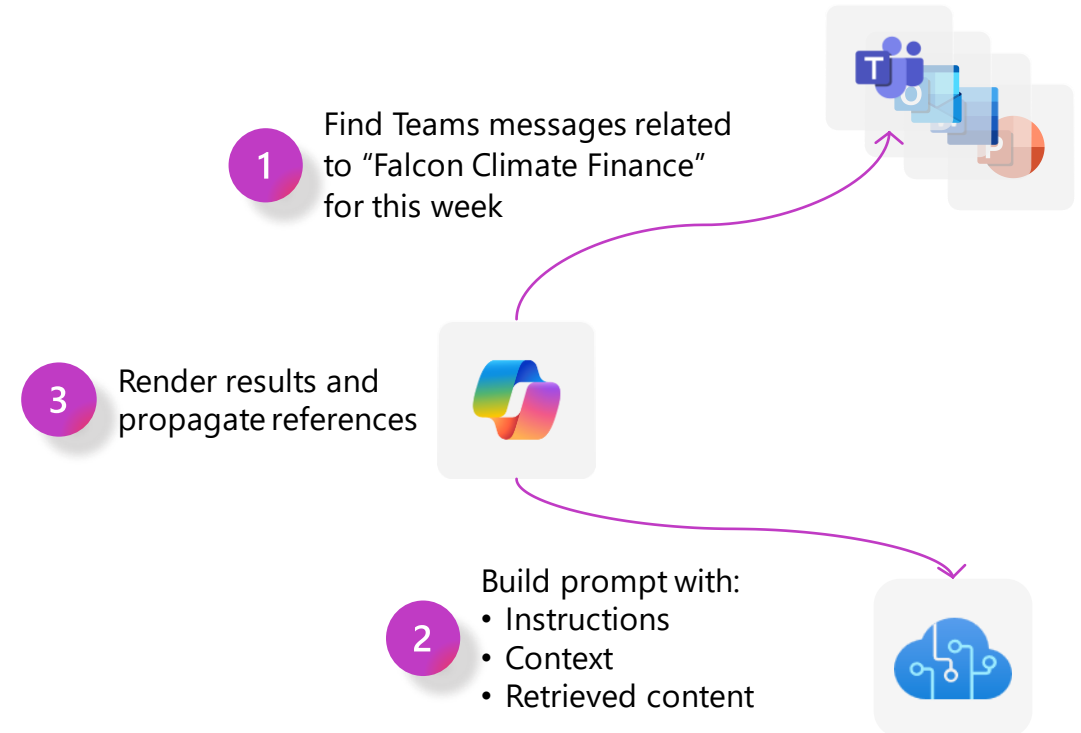
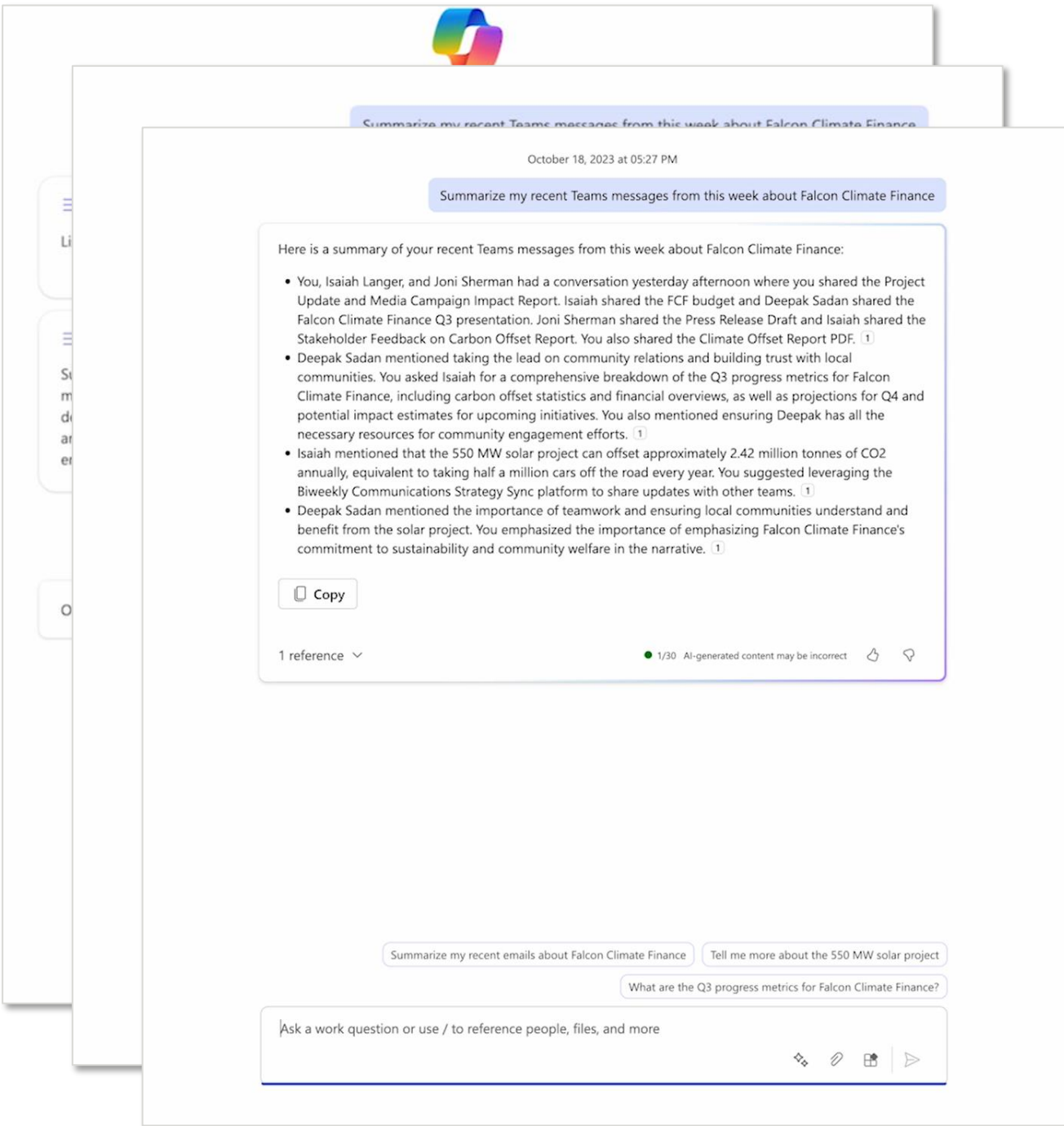
Best,

Abby
Project Leader, Falcon Climate Finance Project

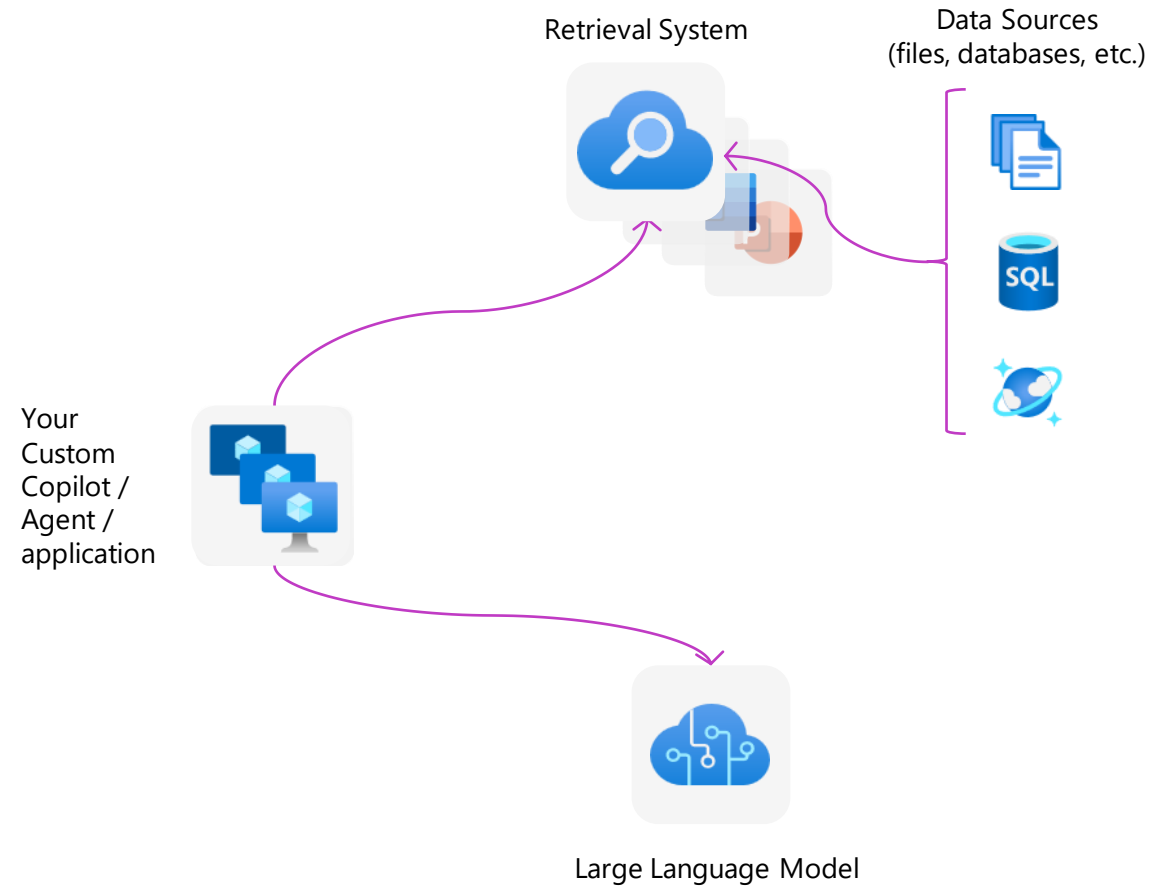
Ngā mihi,
Abby

Abby-Junyi Shen
Technology Strategist – ISV Partners & Digital Natives
New Zealand

Microsoft



RAG – Retrieval Augmented Generation



Robust retrieval for RAG apps

- Responses only as good as retrieved data
- Keyword search recall challenges
 - “vocabulary gap”
 - Gets worse with natural language questions
- Vector-based retrieval finds documents by semantic similarity
 - Robust to variation in how concepts are articulated (word choices, morphology, specificity, etc.)

Example

Question:

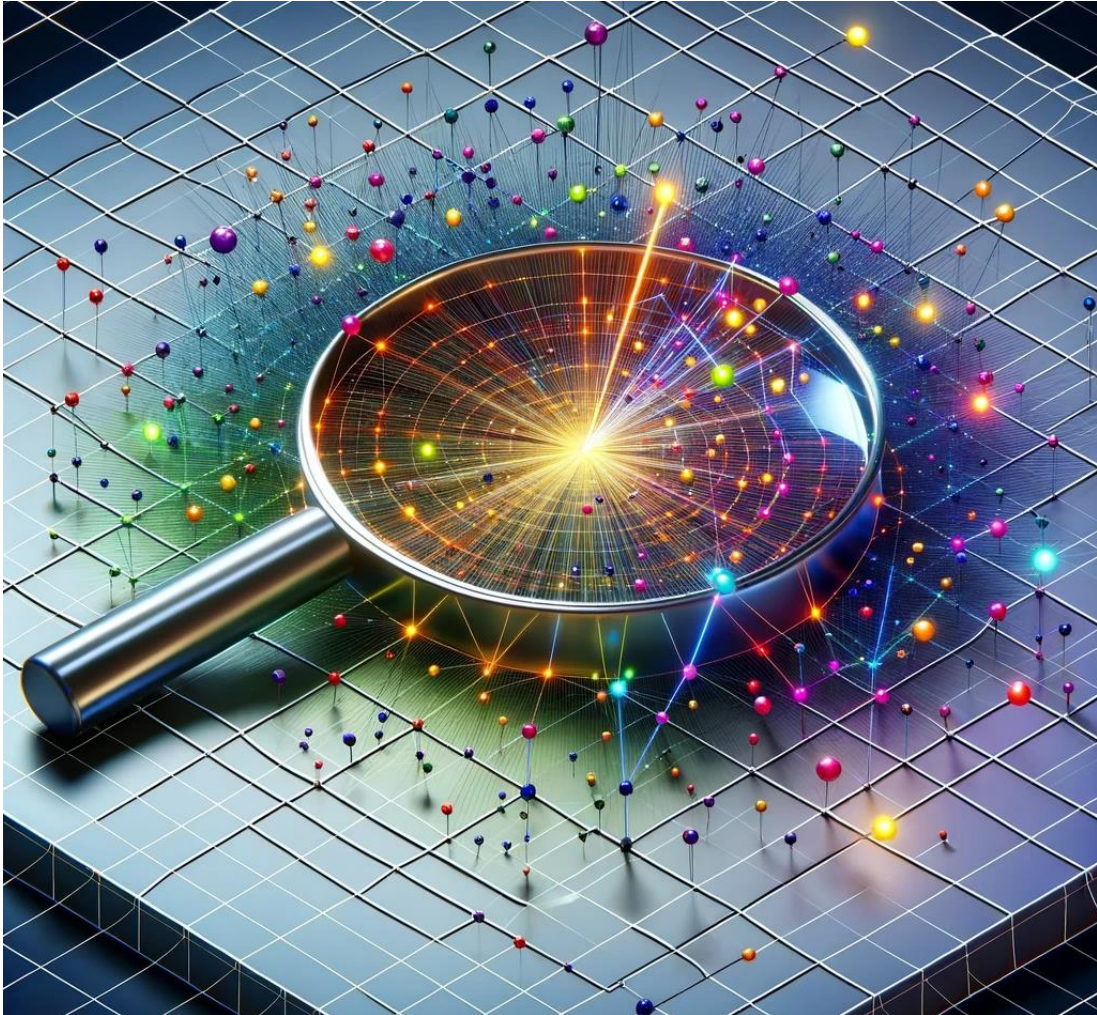
“Looking for lessons on underwater activities”

Won't match:

“Scuba classes”
“Snorkeling group sessions”

Vectors and vector databases

Vectors



Vector representations

- Models that encode item \rightarrow vector
- Similar items map to close vectors
- Sentences, images, graphs, etc.

Vector search

- Find K closest vectors given a "query" vector
- Search exhaustively or through approximations

Vectors

```
▶ client.embeddings.create(input="hello, world", model="ada002").data[0].embedding ⓘ  
[18] ✓ 3.1s  
... [-0.016783414408564568,  
      -0.00672766612842679,  
      -0.02743067592382431,  
      -0.046463146805763245,  
      -0.01095276977866888,  
      0.01014025043696165,  
      -0.013910342939198017,  
      -0.004839369561523199,  
      -0.018681460991501808,  
      -0.028366699814796448,  
      0.028990715742111206,  
      0.0197994876652956,
```

Vector similarity

Query:

```
'The new movie is awesome'
```

Data:

```
['The dog plays in the garden',  
 'This recent movie is so good',  
 'The new movie is awesome']
```

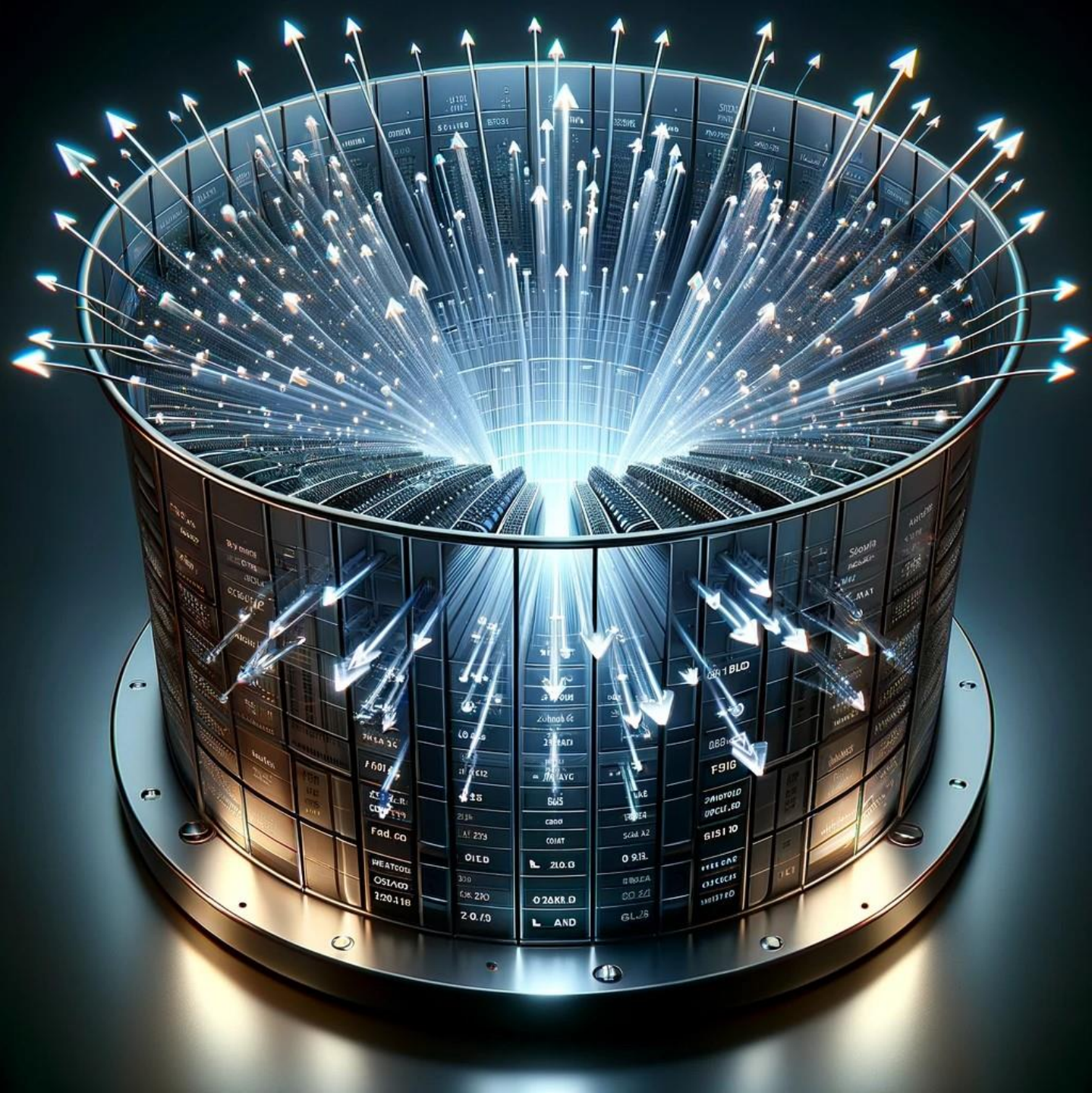
```
The new movie is awesome  
The new movie is awesome  
The new movie is awesome
```

```
The dog plays in the garden  
This recent movie is so good  
The new movie is awesome
```

```
Score: 0.7488  
Score: 0.9191  
Score: 1.0000
```


Vector databases

- Durably store and index vectors and metadata at scale
- Various indexing & retrieval strategies
- Combine vector queries with metadata filters
- Enable access control



Azure AI Search

Feature rich, enterprise-ready vector database

Data and platform integration

State-of-the-art retrieval system



Azure AI Search

**Feature-rich
vector
database**

**Ingest any
data type, from
any source**

**Seamless data
& platform
integrations**

**State-of-
the-art
search ranking**

**Enterprise-
ready
foundation**

Vector search

Azure AI Search in
Azure AI Studio

—
Integrated
vectorization
(Public Preview)

Semantic ranker

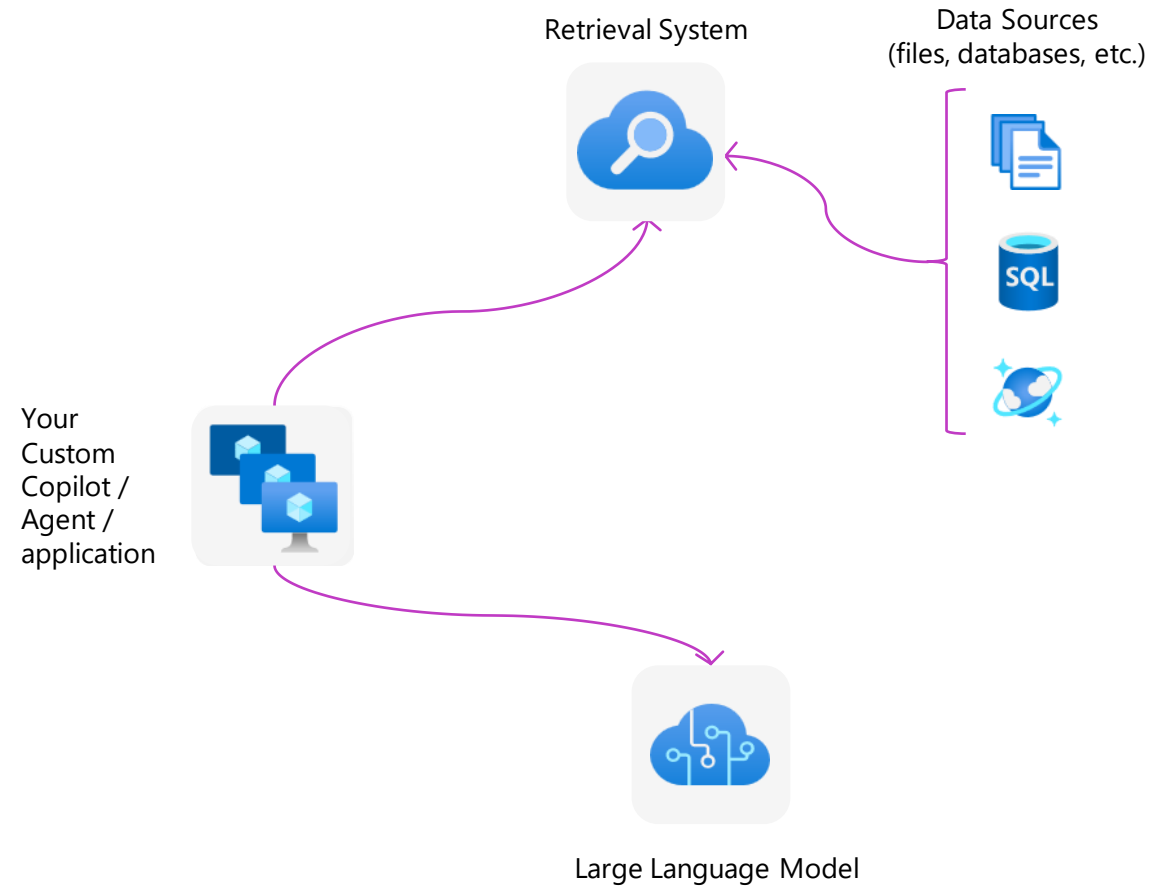
Azure AI Search: vector search

Vector search in Azure AI Search



- Comprehensive vector search solution
- Enterprise-ready
 - scalability, security and compliance
- Integrated with Semantic Kernel, LangChain, LlamaIndex, Azure OpenAI Service, Azure AI Studio, and more

RAG – Retrieval Augmented Generation



Rich vector search query capabilities

Filtered vector search

- Scope to date ranges, categories, geographic distances, etc.
- Rich filter expressions
- Pre-/post-filtering
 - Pre-filter: great for selective filters, no recall disruption
 - Post-filter: better for low-selectivity filters, but watch for empty results

Multi-vector scenarios

- Multiple vector fields per document
- Multi-vector queries
- Can mix and match as needed

Not just text



- Images, sounds, graphs, and more
- Multi-modal embeddings - e.g., images + sentences in Azure AI Vision
- Still vectors → vector search applies
- RAG with images with GPT-4 Turbo with Vision



0a9b0b18-0304-41a2-8367-155479b77985.png



0b0e0e52-e3bf-4bf7-b1e1-3989c91d51a9.png



1a8555cc-a1e4-4fae-9624-8eacfb64bc8.png



2faec00a-4ee2-4711-aec9-240fd6e342c5.png



5b73df27-5275-4437-b7cf-45b1eee175fb.png



5df18ce1-7751-4100-9996-279ba57265f6.png



6cc997c4-da5d-432c-91f7-216a21d7e690.png



8c9044bf-e9cb-4c15-a977-934533055792.png



22ae32ed-7f8f-4238-b7f1-fe5f1dd5a449.png



45e09e69-17c4-47d9-a501-2f8177e34349.png



088c709c-cf18-4253-8664-3e474b94635c.png



92c773f6-ed44-4570-992d-3e0d2af55cbf.png



93afa1a1-d234-409c-ac8e-e0b49dbe79a1.png



791d0836-e0ad-4b46-bc79-1abcf292b700.png



09887388-3a47-4198-ab19-e18d23768ac5.png



a1b30c60-bb66-4bef-a0b4-0766400e306c.png



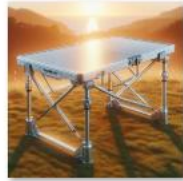
a07b9539-e3ef-4dee-969f-076b0d7ac448.png



cf761719-410d-4d76-9ddb-9ac99d3125b8.png



d4eb8b2d-ef94-4aea-a76e-32f1bdb0457c.png



f3c9d271-39b2-495c-81a8-150d6220c554.png



f8a023c8-169cd-443a-b0bb-b63409d7933f.png

Azure AI Search: data processing

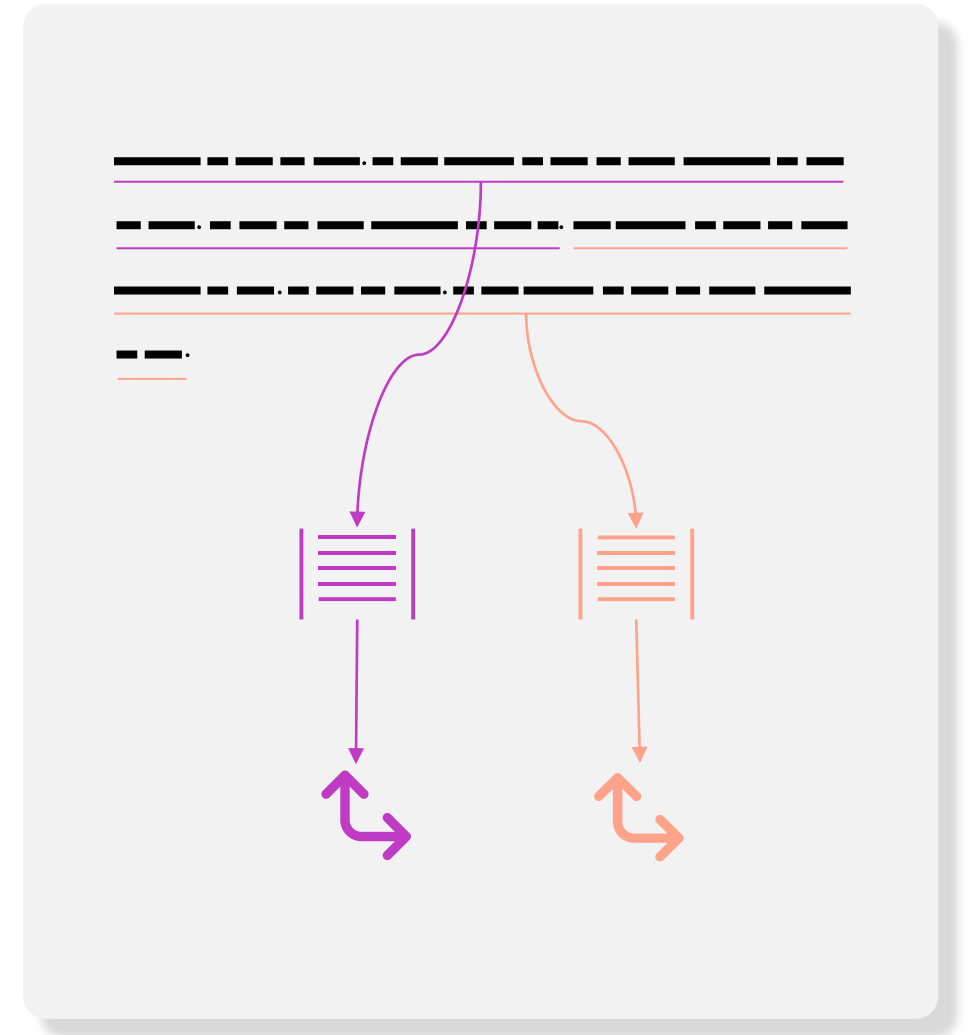
Data preparation for RAG applications

Chunking

- Split long-form text into short passages
 - LLM context length limits
 - Focused subset of the content
 - Multiple independent passages
- Basics
 - ~200–500 tokens/passage
 - Maintain lexical boundaries
 - Introduce overlap
- Layout
 - Layout information is valuable, e.g., tables

Vectorization

- Indexing-time: convert passages to vectors
- Query-time: convert queries into vectors



Integrated vectorization

End-to-end data processing tailored to RAG



Data source access

- Blob Storage
- ADLSv2
- SQL DB
- CosmosDB
- ...

+ Incremental change tracking

File format cracking

- PDFs
- Office documents
- JSON files
- ...

+ Extract images and text, OCR as needed

Chunking

- Split text into passages
- Propagate document metadata

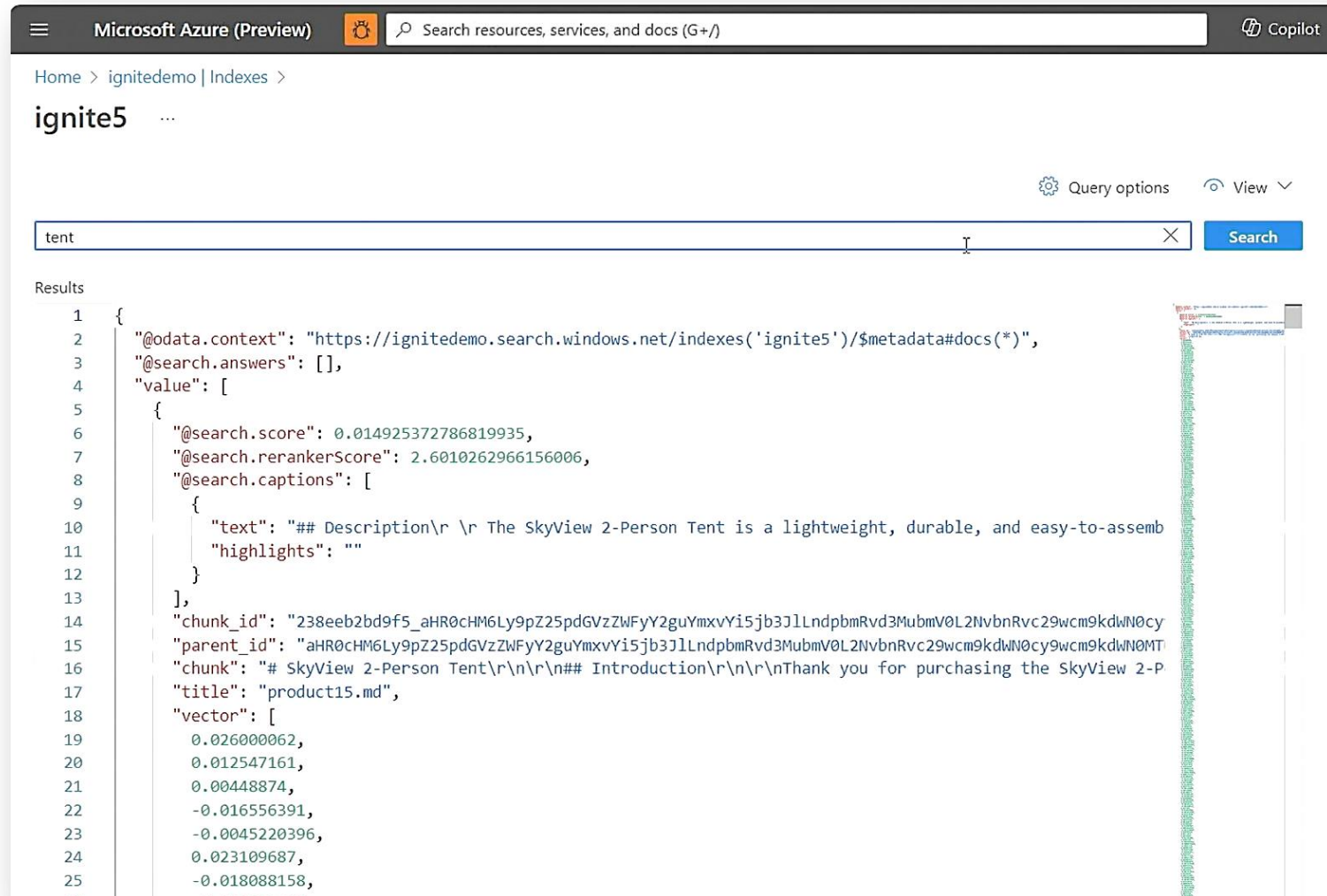
Vectorization

- Turn chunks into vectors
- OpenAI embeddings or your custom model

Indexing

- Document index
- Chunk index
- Both

Even poke around with it!



Microsoft Azure (Preview) Search resources, services, and docs (G+)

Home > ignite5demo | Indexes >

ignite5

Query options View

tent Search

Results

```
1 {
2   "@odata.context": "https://ignitedemo.search.windows.net/indexes('ignite5')/$metadata#docs(*)",
3   "@search.answers": [],
4   "value": [
5     {
6       "@search.score": 0.014925372786819935,
7       "@search.rerankerScore": 2.6010262966156006,
8       "@search.captions": [
9         {
10          "text": "## Description\r\n\r\nThe SkyView 2-Person Tent is a lightweight, durable, and easy-to-assemb
11          "highlights": ""
12        }
13      ],
14      "chunk_id": "238eeb2bd9f5_aHR0cHM6Ly9pZ25pdGVzZWZyY2guYmxvYi5jb3JlLndpbmRvd3MubmV0L2NvbnRvc29wcm9kdWN0cy
15      "parent_id": "aHR0cHM6Ly9pZ25pdGVzZWZyY2guYmxvYi5jb3JlLndpbmRvd3MubmV0L2NvbnRvc29wcm9kdWN0cy9wcm9kdWN0MT
16      "chunk": "# SkyView 2-Person Tent\r\n\r\n## Introduction\r\n\r\nThank you for purchasing the SkyView 2-P
17      "title": "product15.md",
18      "vector": [
19        0.026000062,
20        0.012547161,
21        0.00448874,
22        -0.016556391,
23        -0.0045220396,
24        0.023109687,
25        -0.018088158,
```

Retrieval

Semantic ranker



State of the art re-ranking model

Highest performing retrieval mode

New pay-go pricing

Multilingual capabilities

Includes extractive answers, captions and ranking

Improving relevance

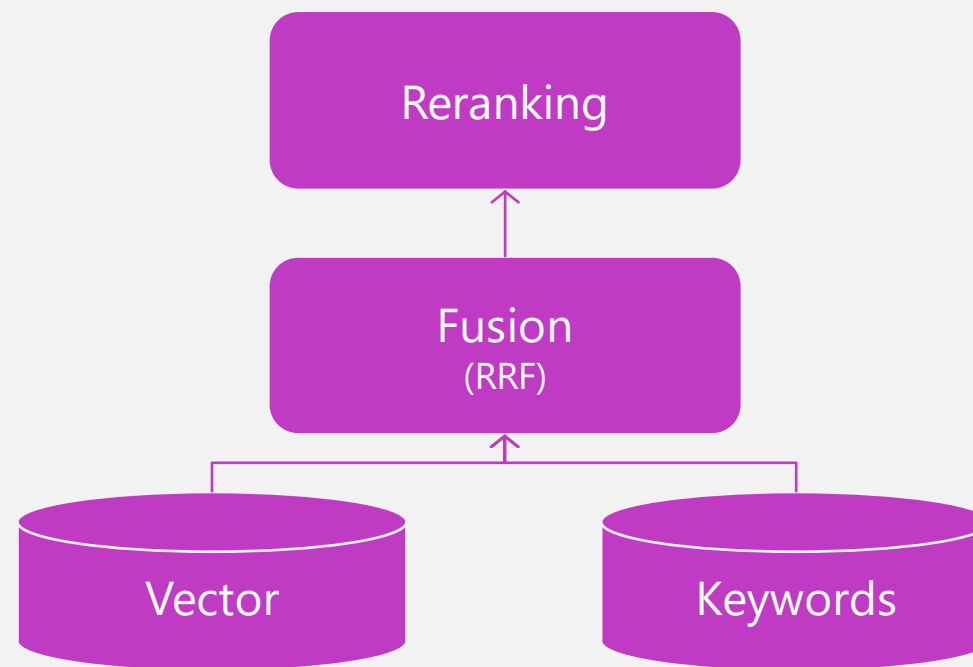
All information retrieval tricks apply!

Complete search stacks do better:

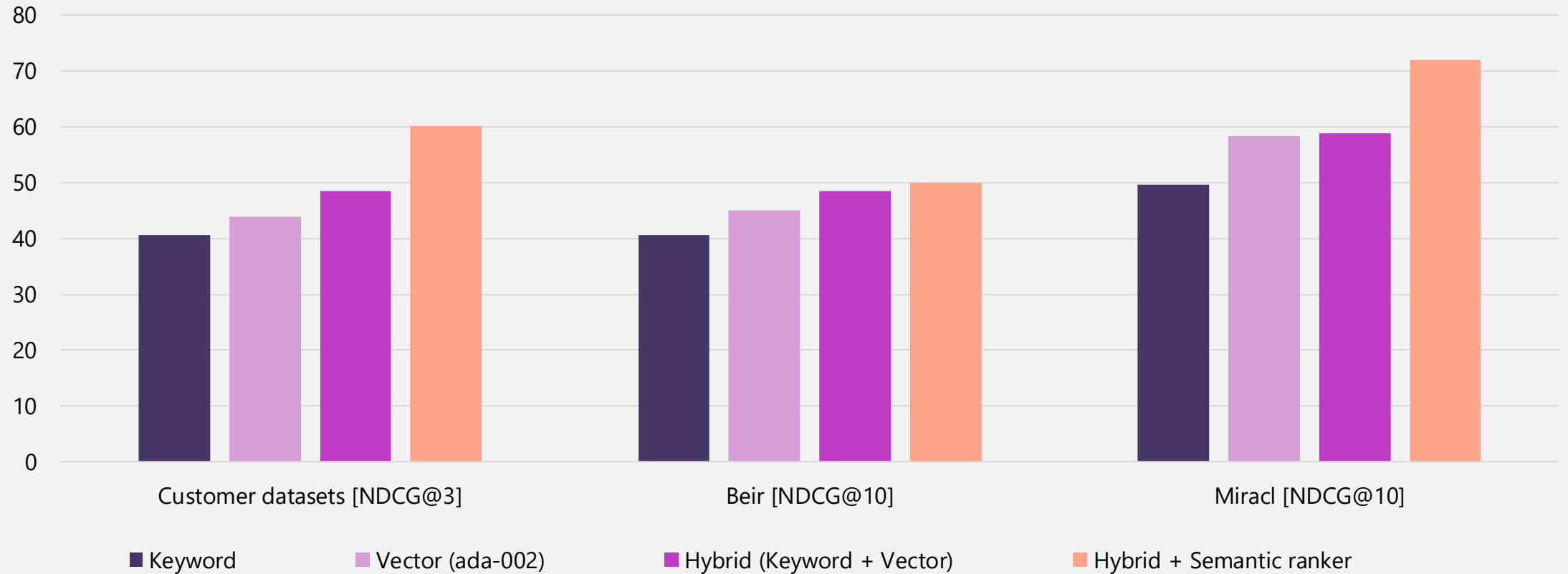
- Hybrid retrieval (keywords + vectors) > pure-vector or keyword
- Hybrid + Reranking > Hybrid

Identify good & bad candidates

- Normalized scores from Semantic ranker
- Exclude documents below a threshold




Retrieval relevance across methods



Retrieval comparison using Azure AI Search in various retrieval modes on customer and academic benchmarks
Source: [Outperforming vector search with hybrid + reranking](#)

Impact of query types on relevance



Query type	Keyword [NDCG@3]	Vector [NDCG@3]	Hybrid [NDCG@3]	Hybrid+ Semantic ranker [NDCG@3]
Concept seeking queries	39	45.8	46.3	59.6
Fact seeking queries	37.8	49	49.1	63.4
Exact snippet search	51.1	41.5	51	60.8
Web search-like queries	41.8	46.3	50	58.9
Keyword queries	79.2	11.7	61	66.9
Low query/doc term overlap	23	36.1	35.9	49.1
Queries with misspellings	28.8	39.1	40.6	54.6
Long queries	42.7	41.6	48.1	59.4
Medium queries	38.1	44.7	46.7	59.9
Short queries	53.1	38.8	53	63.9

Source: [Outperforming vector search with hybrid + reranking](#)

RAG at scale

Powering massive, mission-critical RAG workloads with Azure AI Search

Azure OpenAI Service

aka.ms/aoaiservice



Microsoft
Azure

Search smarter!



Learn more about Azure AI Search

<https://aka.ms/AzureAISearch>



Explore Azure AI Studio for a complete RAG development experience

<https://aka.ms/AzureAIStudio>



Dig more into quality evaluation details and why Azure AI Search will make your application generate better results

<https://aka.ms/ragrelevance>

aka.ms/genAI-links



global AI bootcamp 2024

New Zealand

Auckland