Microsoft

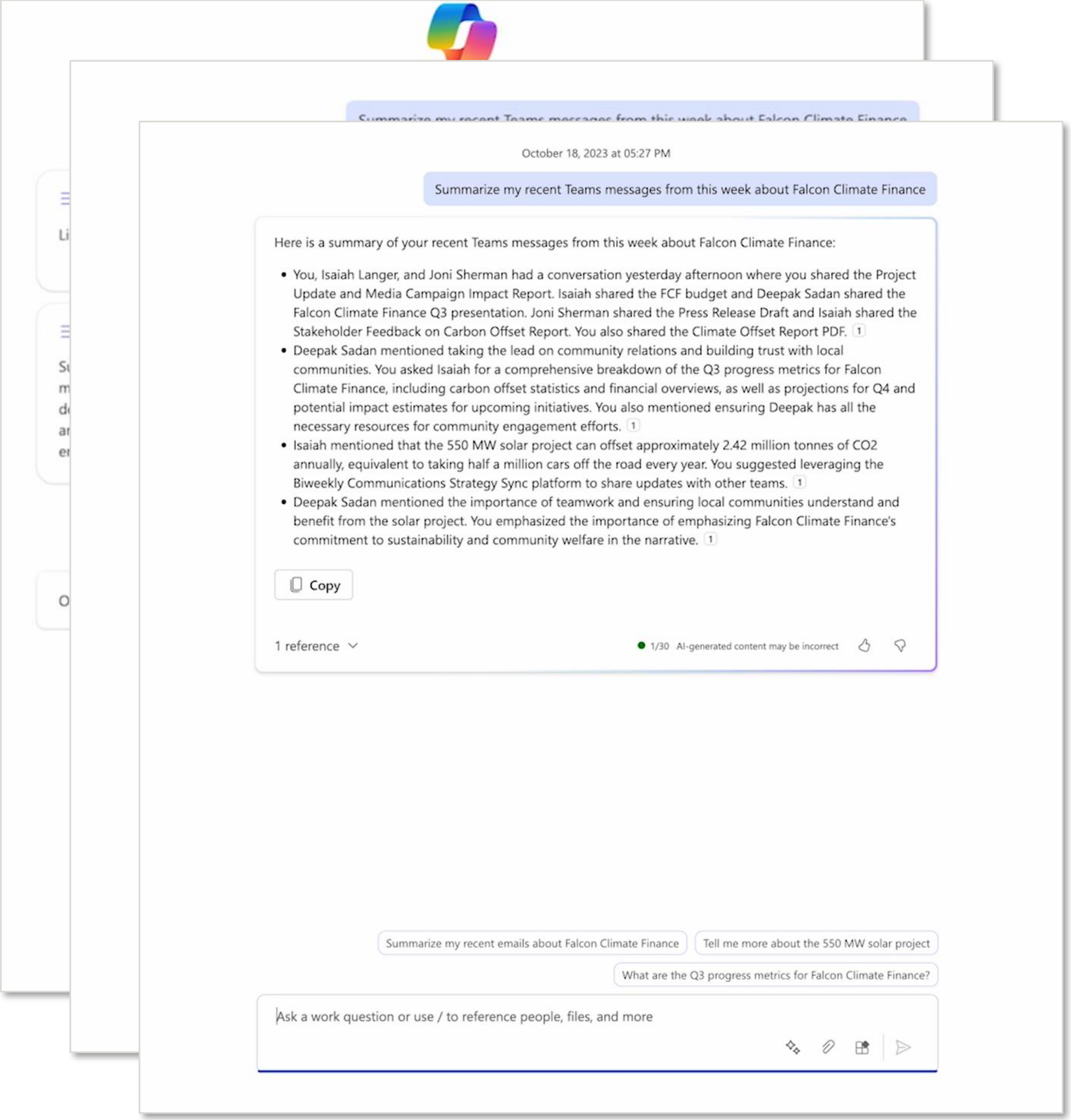
# Vector search and state-of-the-art retrieval for generative AI apps

Pablo Castro, Distinguished Engineer, Azure AI Search

# Agenda

- Retrieval-augmented generation (RAG)
- Vectors and vector databases
- State of the art retrieval with Azure AI Search
- Use cases

# Retrieval-augmented generation (RAG)


October 18, 2023 at 05:27 PM

Summarize my recent Teams messages from this week about Falcon Climate Finance

Here is a summary of your recent Teams messages from this week about Falcon Climate Finance:

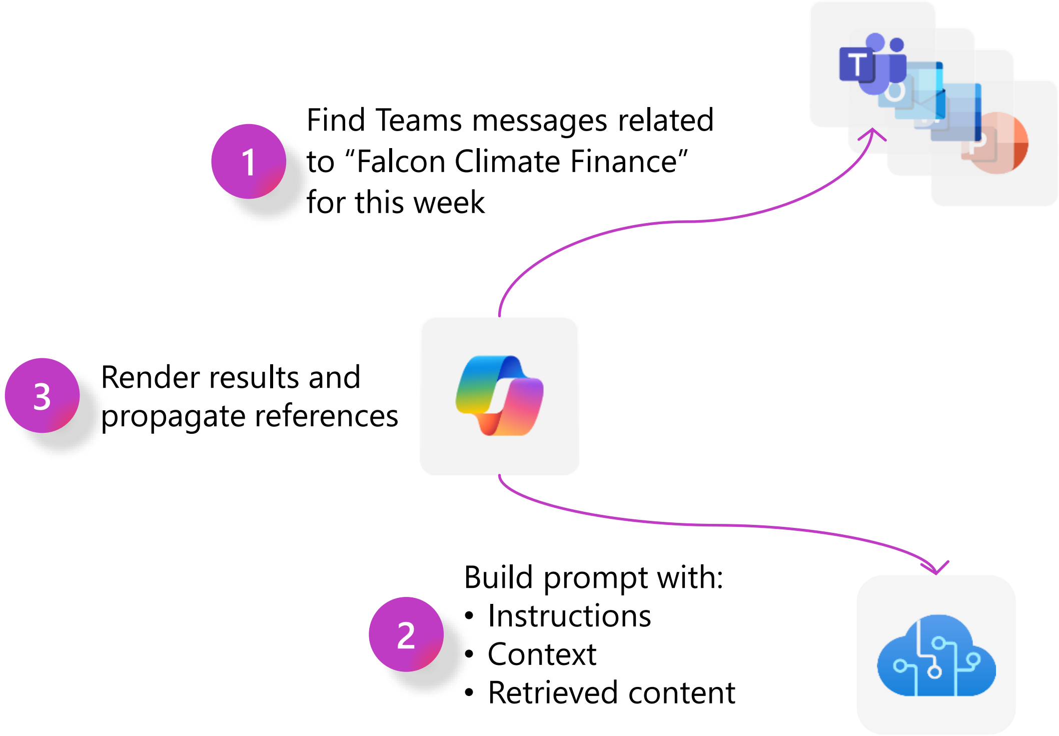
- You, Isaiah Langer, and Joni Sherman had a conversation yesterday afternoon where you shared the Project Update and Media Campaign Impact Report. Isaiah shared the FCF budget and Deepak Sadan shared the Falcon Climate Finance Q3 presentation. Joni Sherman shared the Press Release Draft and Isaiah shared the Stakeholder Feedback on Carbon Offset Report. You also shared the Climate Offset Report PDF. 1
- Deepak Sadan mentioned taking the lead on community relations and building trust with local communities. You asked Isaiah for a comprehensive breakdown of the Q3 progress metrics for Falcon Climate Finance, including carbon offset statistics and financial overviews, as well as projections for Q4 and potential impact estimates for upcoming initiatives. You also mentioned ensuring Deepak has all the necessary resources for community engagement efforts. 1
- Isaiah mentioned that the 550 MW solar project can offset approximately 2.42 million tonnes of CO2 annually, equivalent to taking half a million cars off the road every year. You suggested leveraging the Biweekly Communications Strategy Sync platform to share updates with other teams. 1
- Deepak Sadan mentioned the importance of teamwork and ensuring local communities understand and benefit from the solar project. You emphasized the importance of emphasizing Falcon Climate Finance's commitment to sustainability and community welfare in the narrative. 1

Summarize my recent emails about Falcon Climate Finance

Tell me more about the 550 MW solar project

What are the Q3 progress metrics for Falcon Climate Finance?

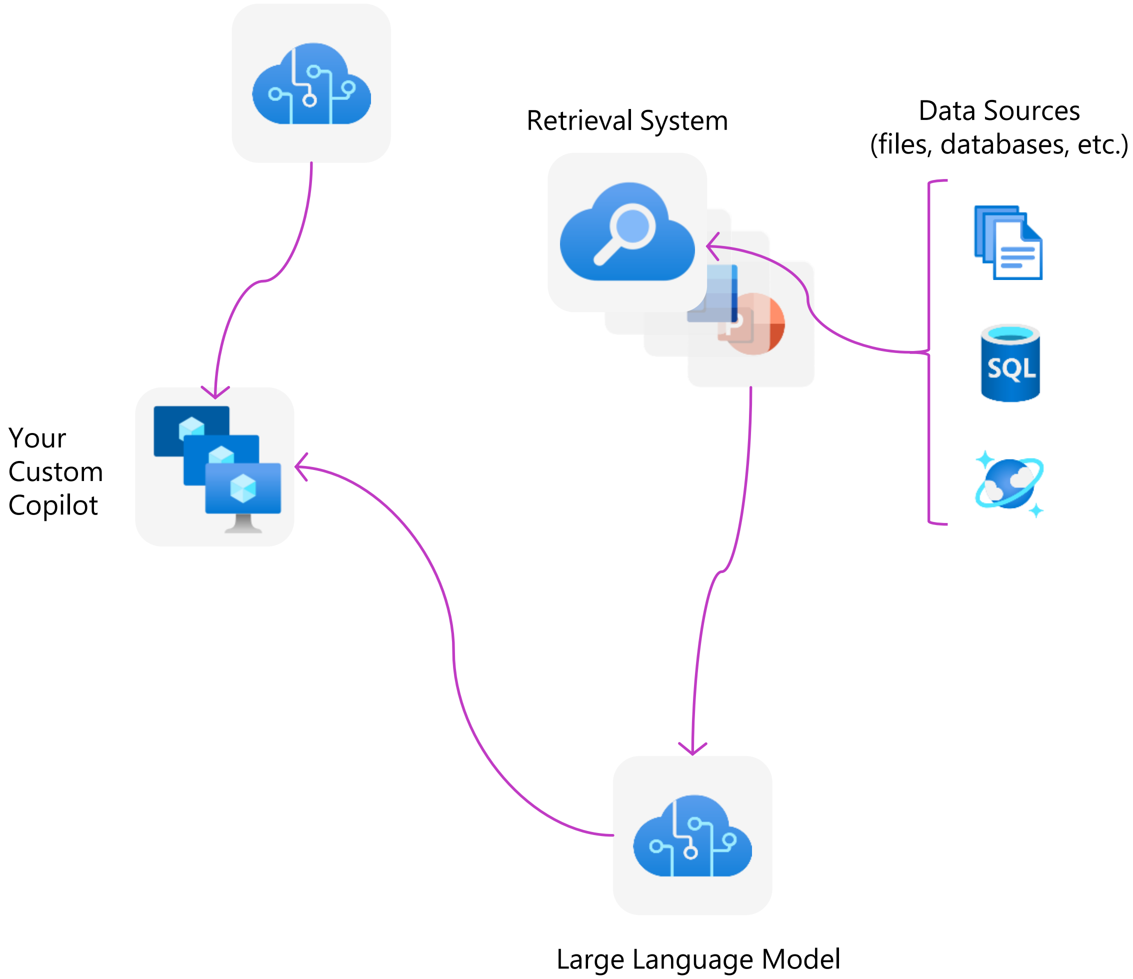
Ask a work question or use / to reference people, files, and more


1 Find Teams messages related to “Falcon Climate Finance” for this week

3 Render results and propagate references

2 Build prompt with:
- Instructions
- Context
- Retrieved content

# RAG – Retrieval Augmented Generation



# Incorporating domain knowledge

**Prompt engineering**

In-context learning

**Fine tuning**

Learn new skills

**Retrieval augmentation**

Learn new facts

# 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”

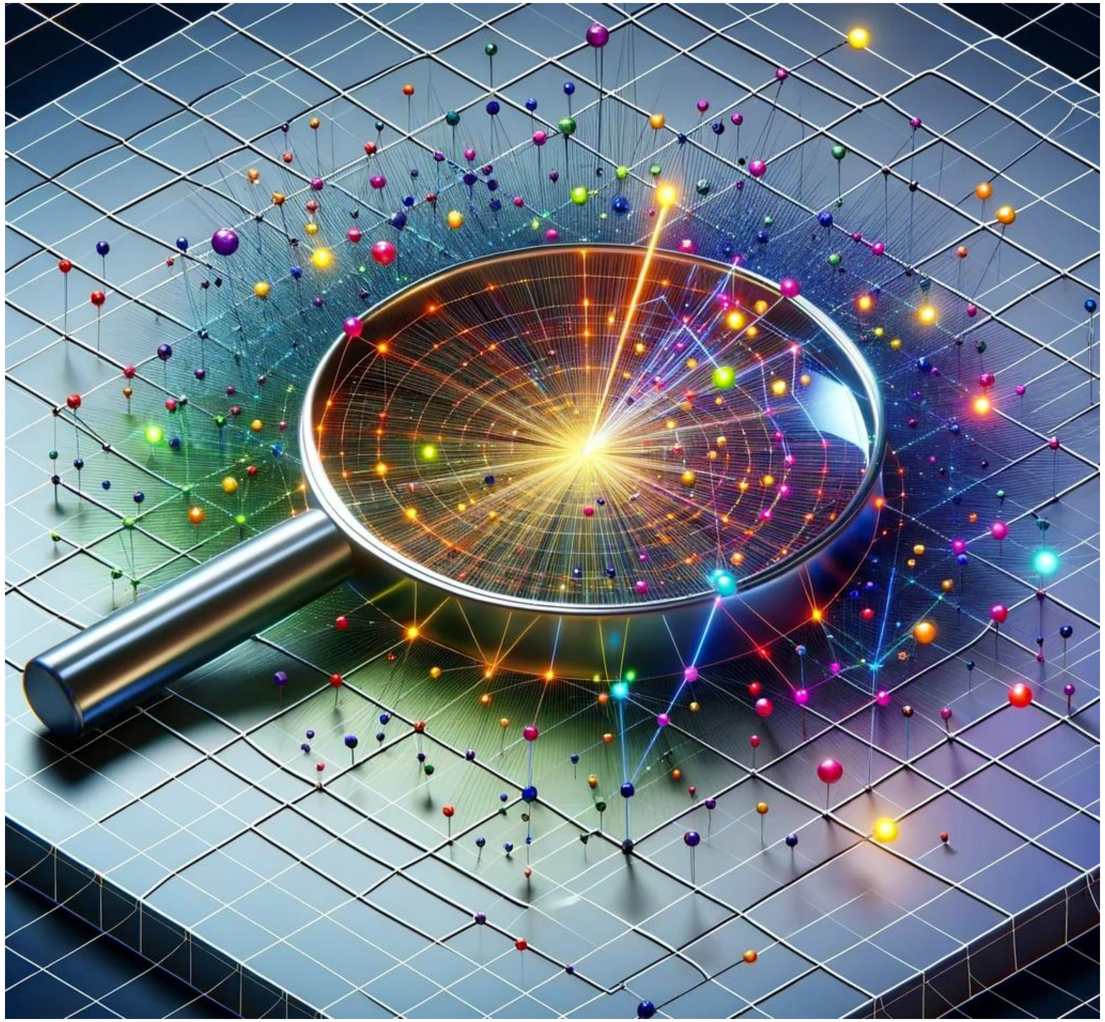
**Won’t match:**

“Scuba classes”

“Snorkeling group sessions”

# Vectors and vector databases

# Vectors



Learned vector representations

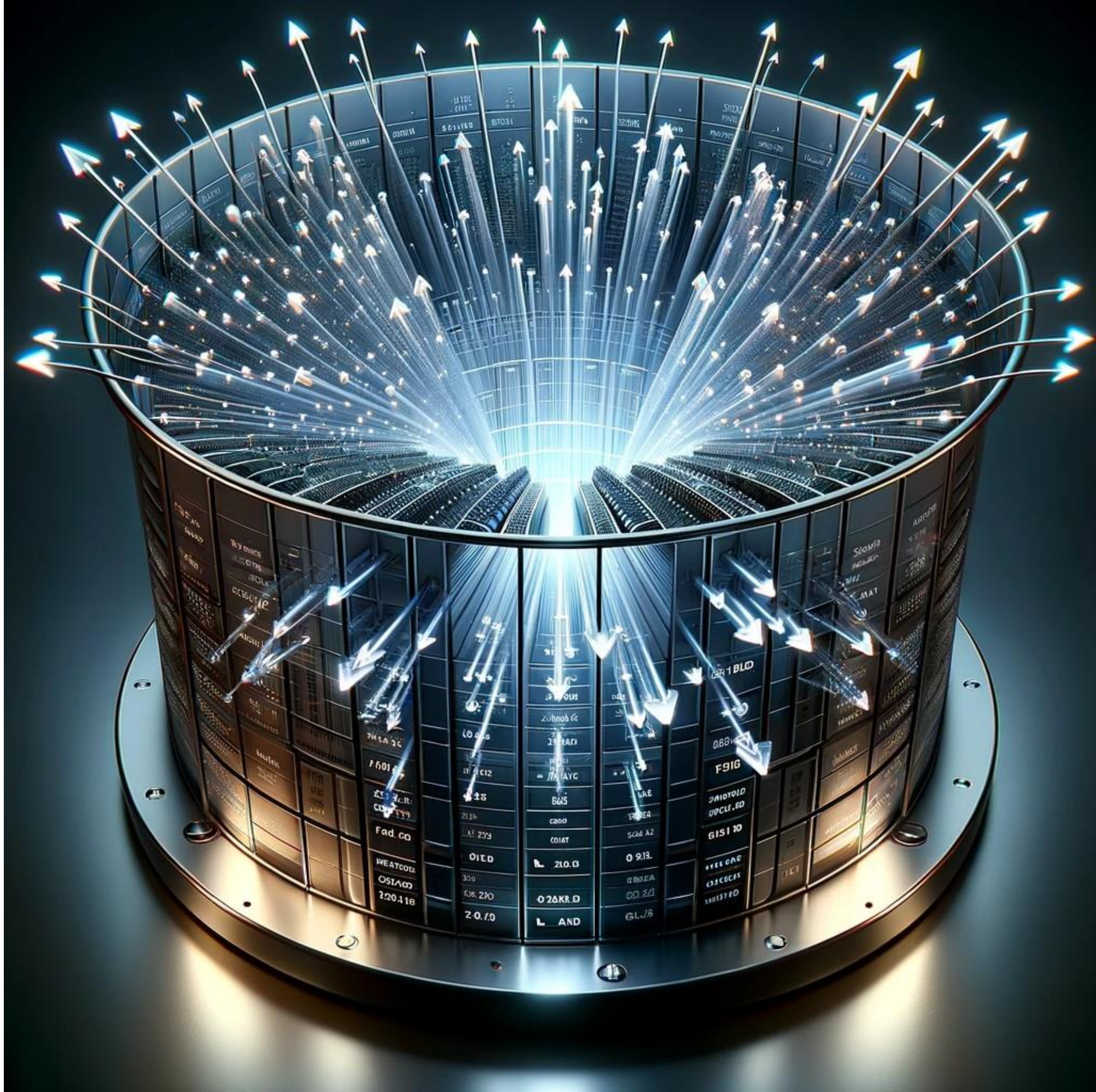
- Models that encode item -> vector
- Similar items map to close vectors
- Sentences, images, graphs, etc.

Vector search

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

# Vector databases

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



# Vector databases in Azure

## Azure AI Search

Best relevance: highest quality
of results out of the box

Automatically index data
from Azure data sources:
SQL DB, Cosmos DB, Blob
Storage, ADLSv2, and more

## Vectors in Azure databases

Keep your data where it is:
native vector search capabilities

Built into
Azure Cosmos DB MongoDB vCore and
Azure Cosmos DB for PostgreSQL services

# Azure AI Search

**Feature-rich vector database**

Generally available

Vector search

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

**Seamless data & platform integrations**

Public preview

Azure AI Search in Azure AI Studio

Integrated vectorization

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

Generally available

Semantic ranker

**Enterprise-ready foundation**

# Vector search in Azure AI Search

Feature rich, enterprise-ready

# Vector search in Azure AI Search

Generally available

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

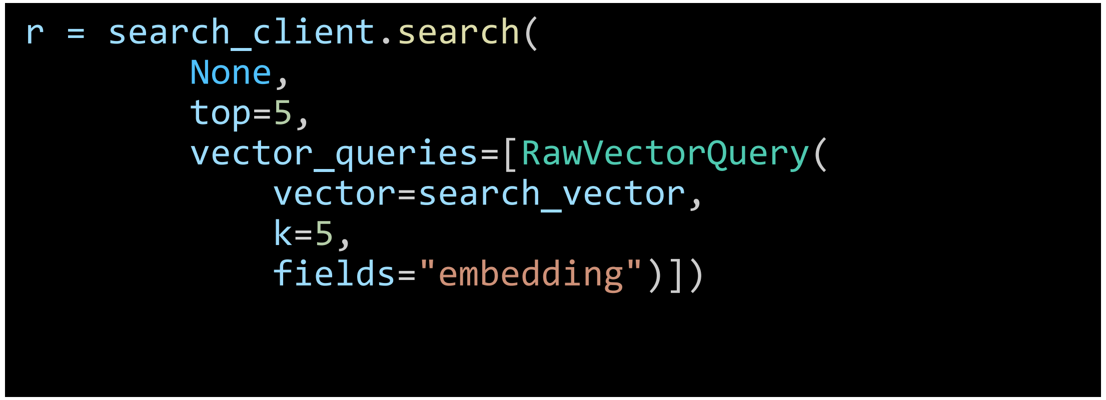
# Vector search strategies

## ANN search

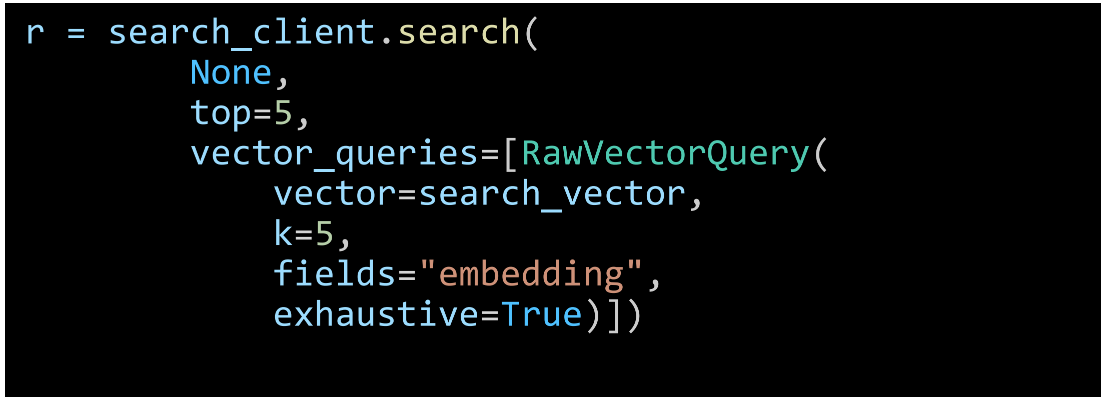
- Fast vector search at scale
- Uses HNSW, a graph method with excellent performance-recall profile
- Fine control over index parameters

## Exhaustive KNN search

- Per-query or built into schema
- Useful to create recall baselines
- Scenarios with highly selective filters
    - e.g., dense multi-tenant apps



```
r = search_client.search(
        None,
        top=5,
        vector_queries=[RawVectorQuery(
            vector=search_vector,
            k=5,
            fields="embedding")])
```



```
r = search_client.search(
        None,
        top=5,
        vector_queries=[RawVectorQuery(
            vector=search_vector,
            k=5,
            fields="embedding",
            exhaustive=True)])
```

# 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

```python
r = search_client.search(
        None,
        top=5,
        vector_queries=[RawVectorQuery(
            vector=query_vector,
            k=5,
            fields="embedding")],
        vector_filter_mode=VectorFilterMode.PRE_FILTER,
        filter=
"category eq 'perks' and created gt 2023-11-15T00:00:00Z")
```

## Multi-vector scenarios

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

```python
r = search_client.search(
        None,
        top=5,
        vector_queries=[
            RawVectorQuery(
                vector=query1, k=5, fields="embedding"),
            RawVectorQuery(
                vector=query2, k=5, fields="embedding")
        ])
```

# Enterprise ready vector database

| | |
|---|---|
| **Data Encryption** | Including option for customer-managed encryption keys |
| **Secure Authentication** | Managed identity and RBAC support |
| **Network Isolation** | Private endpoints, virtual networks |
| **Compliance Certifications** | Extensive certifications across finance, healthcare, government, etc. |

# 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

# Azure AI Search:

## Seamless Data and Platform Integrations

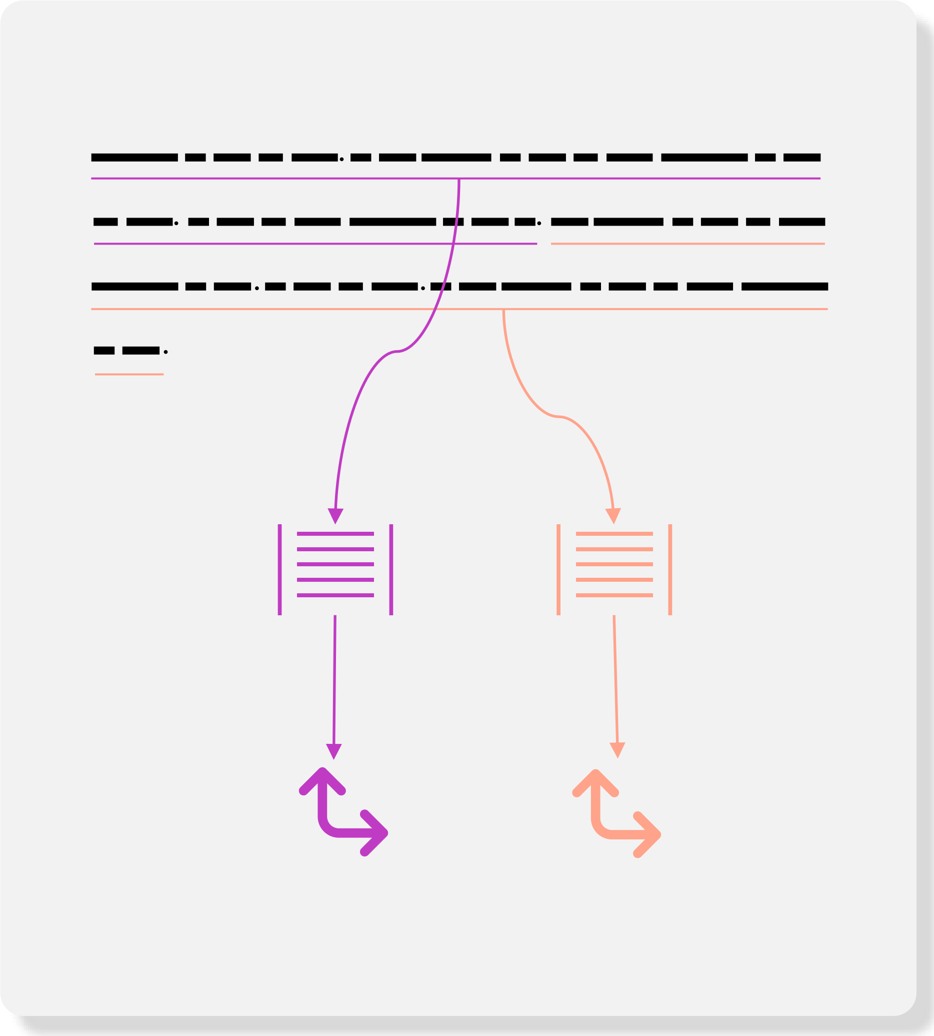
# Data preparation for RAG applications

## 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

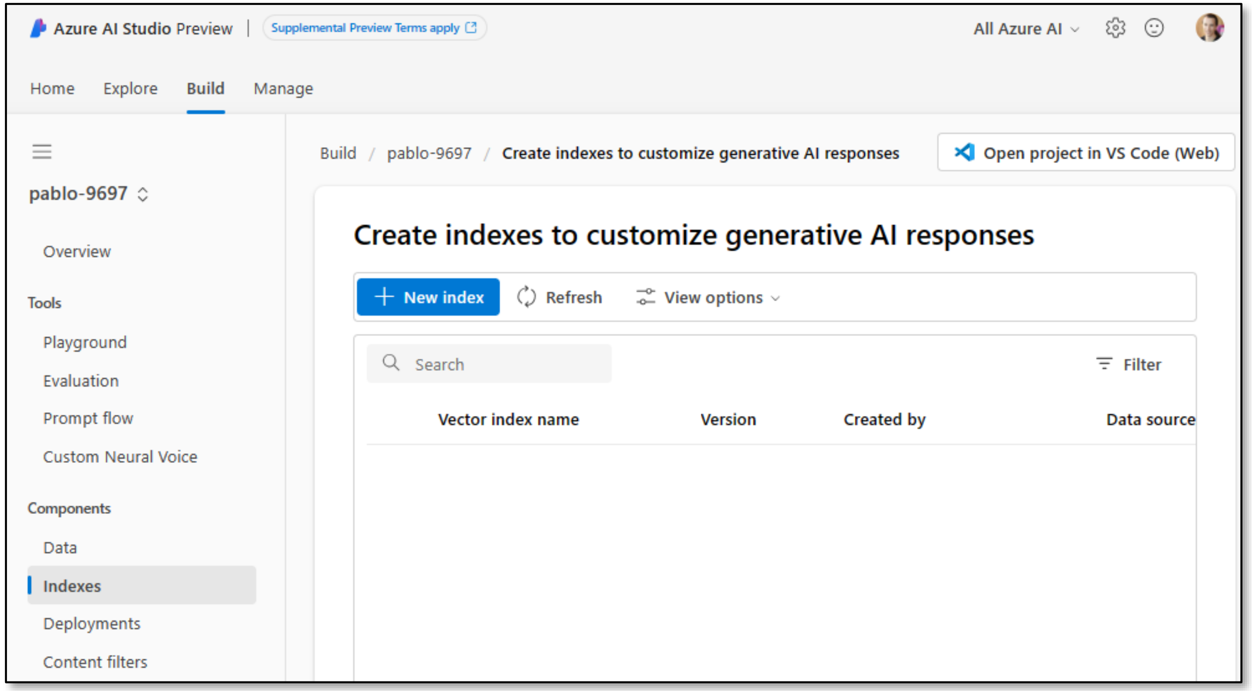
## Vectorization

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



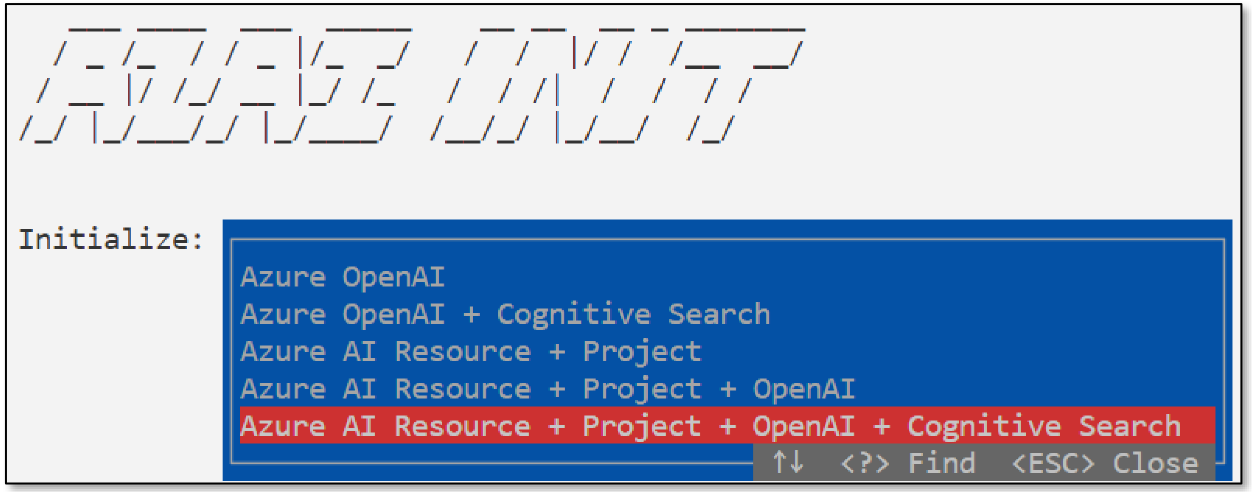
# Azure AI Studio & Azure AI SDK

- First-class integration
- Build indexes from data in Blob Storage, Microsoft Fabric, etc.
- Attach to existing Azure AI Search indexes



Create indexes to customize generative AI responses

| Vector index name | Version | Created by | Data source |
|---|---|---|---|



```
   ___ ____   ___  _____    __ __  __ _ _______
  / _ /_  /  / _ |/_  _/   / /  |/ / /_  __/
 / __ |/ /_/ __ |_/ /_   / / /|  / /  / /
/_/ |_/___/_/ |_/___/  /_/_/ |_/_/  /_/


Initialize:  Azure OpenAI
             Azure OpenAI + Cognitive Search
             Azure AI Resource + Project
             Azure AI Resource + Project + OpenAI
             Azure AI Resource + Project + OpenAI + Cognitive Search
                               ↑↓  <?> Find  <ESC> Close
```

# Integrated vectorization

End-to-end data processing tailored to RAG

In preview

### 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

# Azure AI Search:

State-of-the-art retrieval system

Generally available

# Semantic ranker

SOTA re-ranking model

Highest performing retrieval mode

New pay-go pricing: Free 1k requests/month, $1 per additional 1k

Multilingual capabilities

Includes extractive answers, captions and ranking

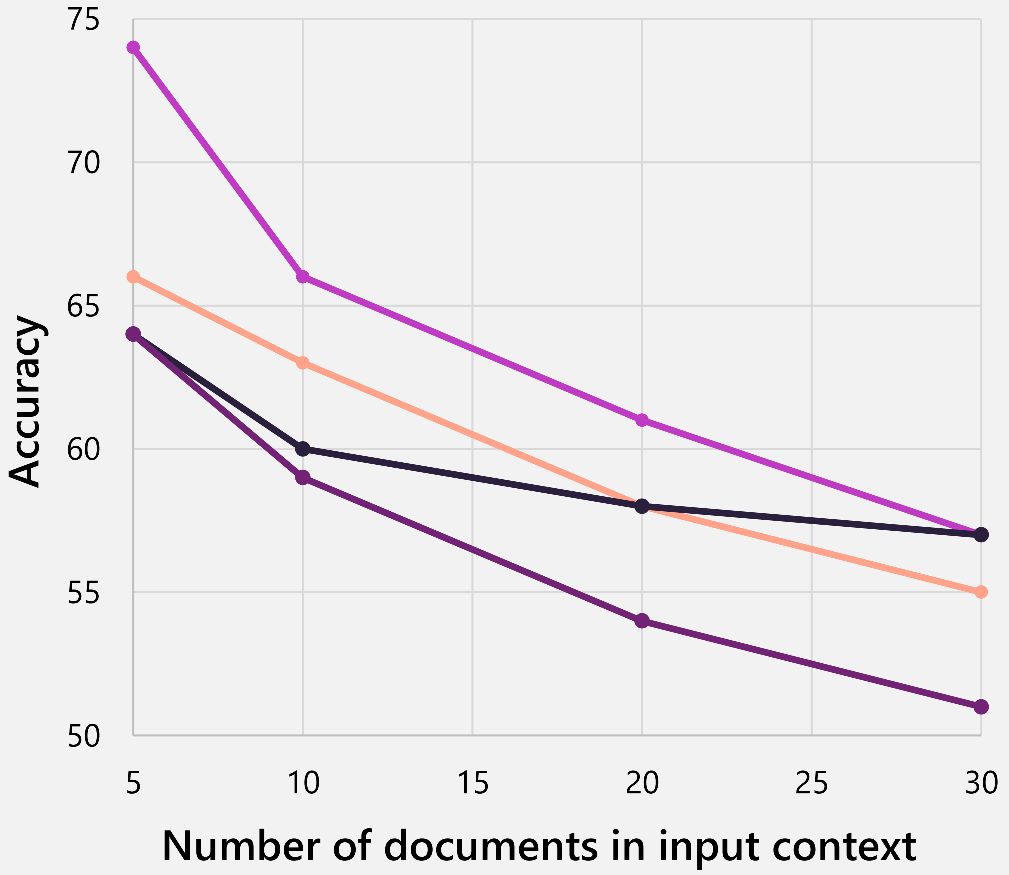
*Formerly semantic search

# Relevance

- Relevance is critical for RAG apps

- Lots of passages in prompt → degraded quality
  → Can’t only focus on recall

- Incorrect passages in prompt → possibly well-grounded yet wrong answers
  → Helps to establish thresholds for “good enough” grounding data



Source: Lost in the Middle: How Language Models Use Long Contexts, Liu et al. arXiv:2307.03172

# Improving relevance

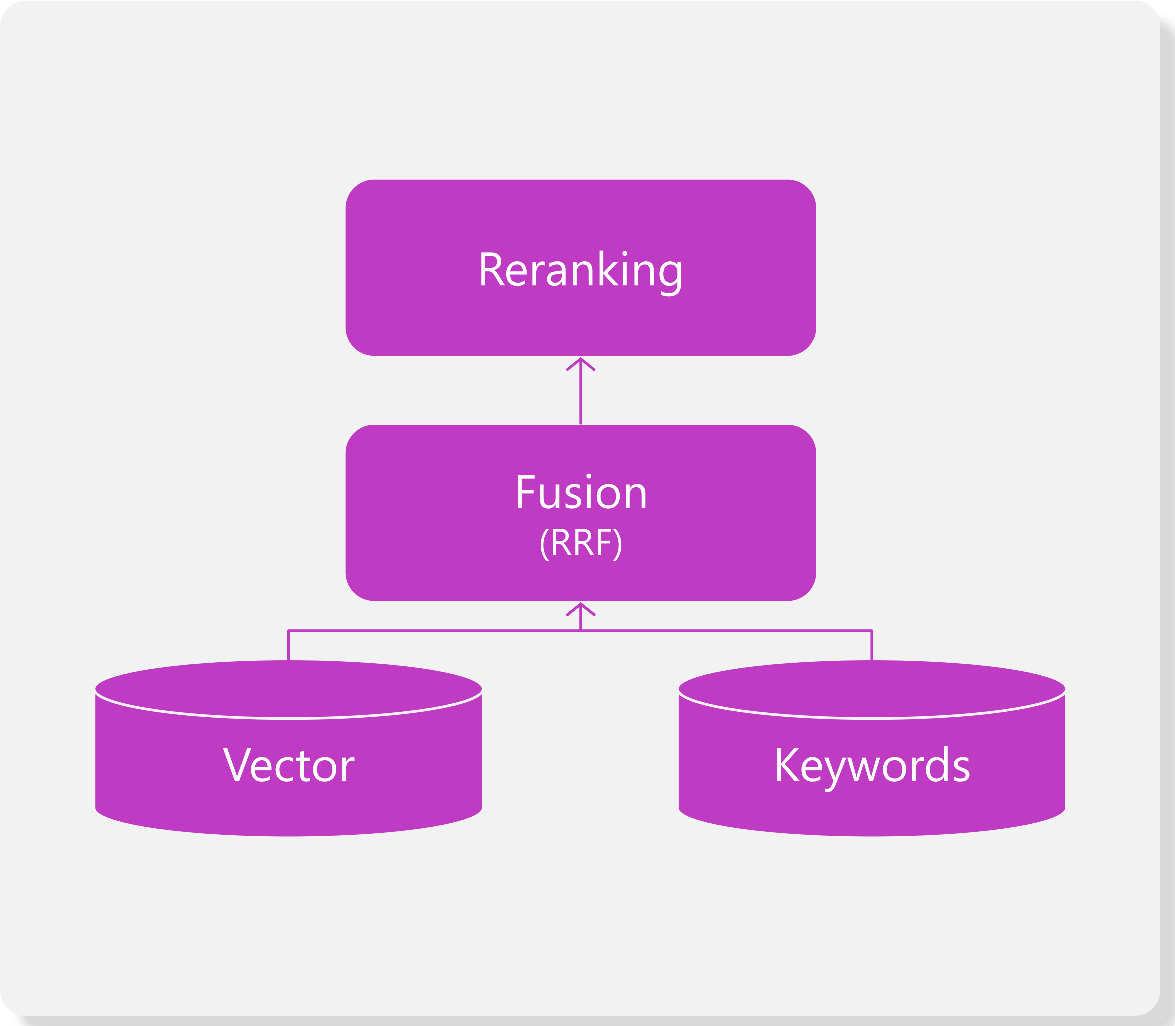
All information retrieval tricks apply!

Complete search stacks do better:

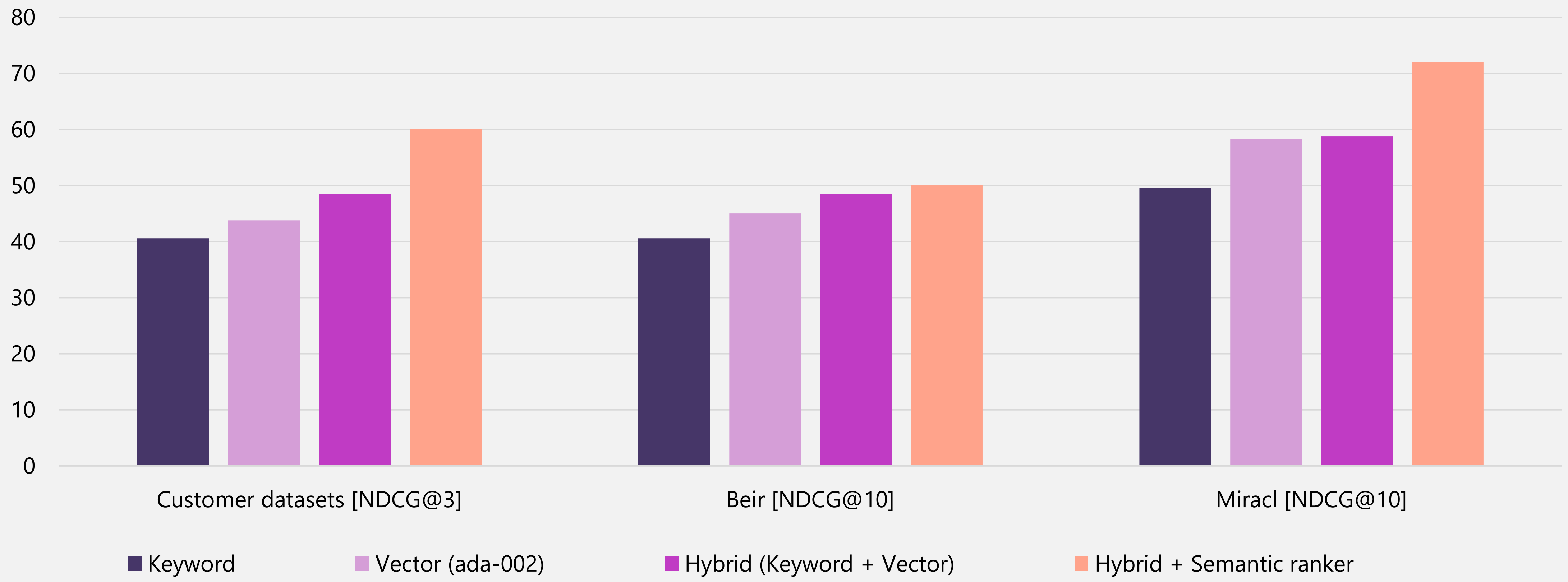
- Hybrid retrieval (keywords + vectors) > pure-vector or keyword
- Hybrid + Reranking > Hybrid

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

Announced July 2023

# Azure Migrate and Modernize & Azure Innovate

Offerings spanning Migration to Innovation in one place

## **Comprehensive resources** in one place

- **Extensive guidance** optimized approach from start to finish with assessments, proof of concepts, pilots, tooling, deployment
- **Free automated tooling** provides you with discovery, assessment, business case analysis, planning, migration, and modernization capabilities
- **Proven technical frameworks** to help design optimized workloads with security and cost recommendations built throughout

## **Direct access to experts** and funding

- Access to validated, specialized partners with advanced capabilities to help with all stages from planning to deployment
- Microsoft-led delivery for rapid rehost migrations, and specialized partners for more complex workloads
- Rich Investments and funding across every stage, from planning to deployment to maximize savings

## **Extensive coverage** —from migration to innovation

- **End-to-end coverage** including migrating or modernizing Windows Server & SQL Server, Linux, Oracle, SAP, HPC, analytics, AI, and more
- Built for all, no matter the size of your business or industry

**Learn more!**
**aka.ms/AzureHeroOfferings**