

Build Applications with LLMs

Azure OpenAI Service

Mutaz Abu Ghazaleh

AI Global Black Belt | Microsoft

24/09/2024



Setting the stage with generative AI

Forbes

MIT Technology Review

SUBSCRIBE



5 Unexpected Ways Generative AI May Change Your Daily Life

Bernard Marr Contributor

Follow

Jan 8, 2024, 02:56am EST



5 Unexpected Ways Generative AI May Change Your Daily Life. ADOBE STOCK

No doubt you've seen dozens of headlines (or more) about generative AI and how tools like ChatGPT are going to transform our jobs. But what about everyday life? Because the fact that AI can now generate content – including text, images, video, and

Bloomberg Company & Product Support Bloomberg Technical Support Bloomberg Customer Support

Bloomberg

Values Careers Stories Press Find jobs

ARTIFICIAL INTELLIGENCE

AI for everything: 10 Breakthrough Technologies 2024

Generative AI tools like ChatGPT reached mass adoption in record time, and reset the course of an entire industry.

By Will Douglas Heaven

January 8, 2024

QuantumBlack AI & Machine Learning

The state of AI in 2023: Generative AI's breakout year

August 1, 2023 | Survey

Bloomberg Intelligence: New Report Finds That the Emerging Industry Could Grow at a CAGR of 42% Over the Next 10 Years

Rising demand for generative AI products could add about \$28 billion of new software revenue

New York, June 01, 2023 – With the influx of consumer generative AI programs like Google's Bard and OpenAI's ChatGPT, the generative AI market is poised to explode, growing to \$1.3 trillion over the next

accenture

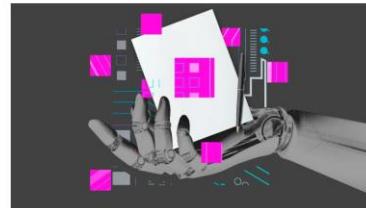
A new era of generative AI for everyone

The technology underpinning ChatGPT will transform work and reinvent business

FAST COMPANY

What Is Generative AI? Your Questions Answered

As generative AI becomes increasingly popular, here's a guide that will get you up to speed.



[Photo: Rawpixel (hand, pen-and-paper, circuit board)]



BY DANICA LO LONG READ

Generative AI seems to have popped up everywhere in the mainstream lately—primarily via the popularity of **ChatGPT**, but also through the proliferation of **text-to-image** tools and **AI avatars** in our social media feeds. But beyond fun smartphone apps and ways for students to shirk essay-writing assignments, what can generative AI do? How does it work? How will it AI

WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS MORE SIGN IN SUBSCRIBE

How WIRED Will Use Generative AI Tools

Some publications are already using text and image generators. Here's how WIRED will—and won't—use the technology.

LIKE PRETTY MUCH everyone else in the past few months, journalists have been trying out generative AI tools like ChatGPT to see whether they can help us do our jobs better. AI software can't call sources and wheelie information out of them, but it can produce half-decent transcripts of those calls, and new generative AI tools can condense hundreds of pages of those transcripts into a summary.

Writing stories is another matter, though. A few publications have tried—*sometimes with disastrous results*. It turns out current AI tools are very good at churning out convincing (if formulaic) copy riddled with falsehoods.

This is WIRED, so we want to be on the front lines of new technology, but also to be ethical and appropriately circumspect. Here, then, are some ground rules on how we are using the current set of generative AI tools. We recognize that AI will develop and so may modify our perspective over time, and we'll acknowledge any changes in this post. We welcome feedback at mail@wired.com.

Text Generators (e.g. LaMDA, ChatGPT)

We do not publish stories with text generated by AI, except when the fact that it's AI-generated is the whole point of the story. (In such cases we'll disclose the use and flag any errors.) This

DIA 3750.24 ▾ 4.0% ▾ S&P 500 4795.99 ▾ Nasdaq 14863.85 ▾ Dow ▾ US 10 Yr 2.72 Yield 4.03% ▾ Euro 1.0924

THE WALL STREET JOURNAL

English Edition Print Edition Video Audio Latest Headlines More

How Did Companies Use Generative AI in 2023? Here's a Look at Five Early Adopters

Business technology leaders in construction, travel, retail, healthcare and energy say AI is already improving productivity and changing customer behavior. They are also sorting through its high costs and limitations.

THOMAS R. LECHLEITER/THE WALL STREET JOURNAL



How Generative AI Can Deliver

Generative AI is a game changer
...now it's all about implementation

To maximize value, consider use cases that serve unmet user needs

LEARNING

Transform internal and external search functions into **human-centered insight hubs** for collaboration and marketing

1

Chat with your data

TRAINING

Empower customer service with insight and encouragement that **helps customers** while supporting talent

2

Intelligent call centers

CREATIVITY

Bring generative assistance **into your custom app** workflows with your data for any user

3

Your own copilots

REASONING

Aggregate and reason over multiple data sources to **enrich decisions** and customer interactions

4

Information discovery

TAILORING

Create recommendation engines that **enhance products and services** or launch new offerings

5

Hyper-personalization



Azure OpenAI Service

General availability

GPT-4o

General availability

**GPT-4 Turbo with Vision, DALL·E 3,
Whisper**

Preview (New Features)

Assistants API

Preview

Fine Tuning for GPT-4

Preview

Batch API

Generally Available (Refresh)

On your data

Preview

**Configurable Content Filters for
Dall-E and GPT-4-Turbo**

Preview

**Integration with Azure
Communication Services for
WhatsApp**

Generally available

GPT-4o available in Azure OpenAI Service

The most advanced multimodal API

Text, audio
and vision

2x faster,
5x throughput

50%
lower cost

GPT-4o matches GPT-4 Turbo in English text and coding tasks while offering superior performance in non-English languages and in audio and vision tasks, setting new benchmarks for AI capabilities



Get Azure for Students

Get Azure

Azure Free Account

Popular services free for 12 months

40+ other services free always

+

Start with USD200* Azure credit

You'll have **30 days** to use it—in addition to free services.

<https://azure.microsoft.com/en-au/free/>

Azure for Students

Start with \$100 Azure credit

No credit card required

+

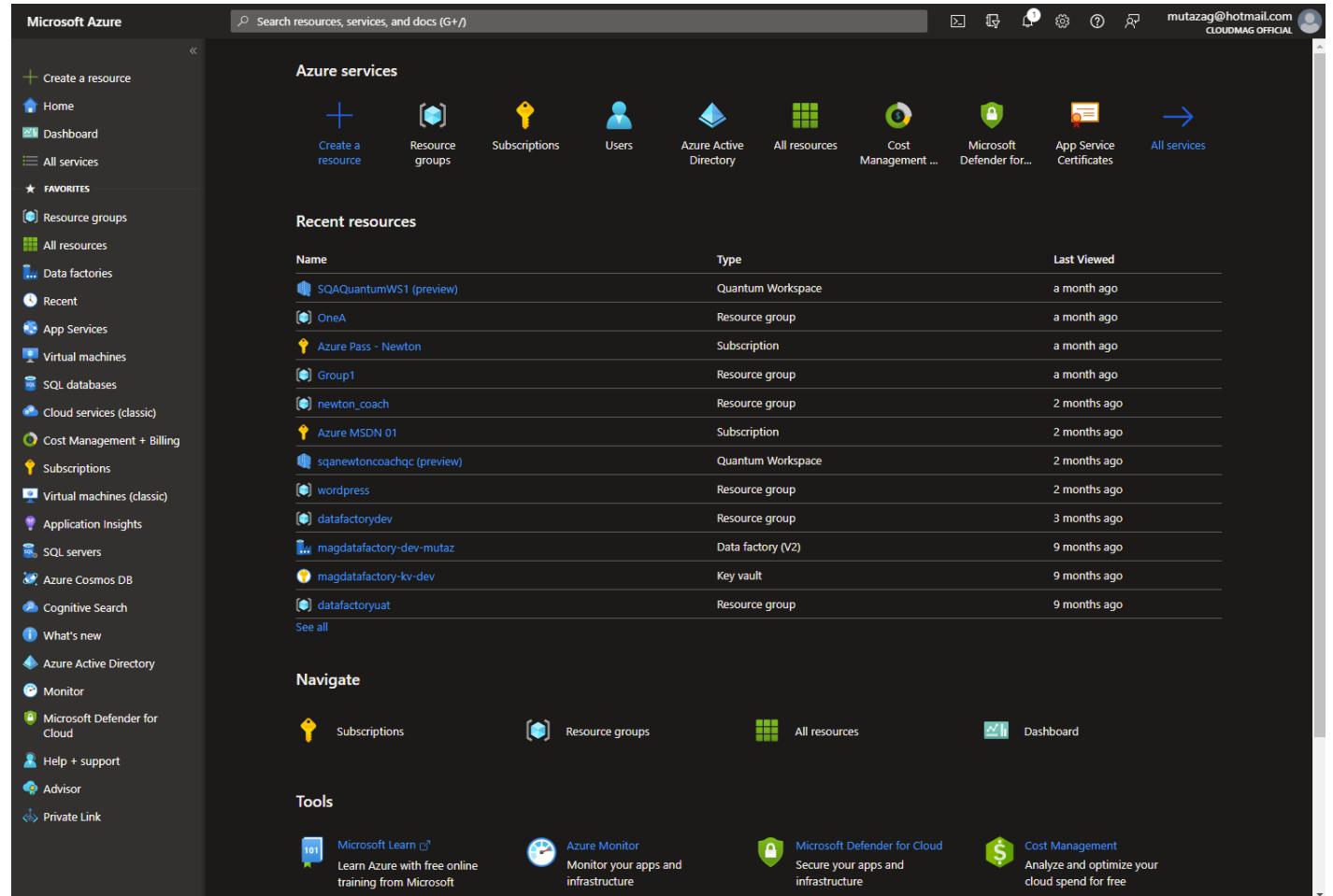
Free services

Get popular services free while you have your credit.

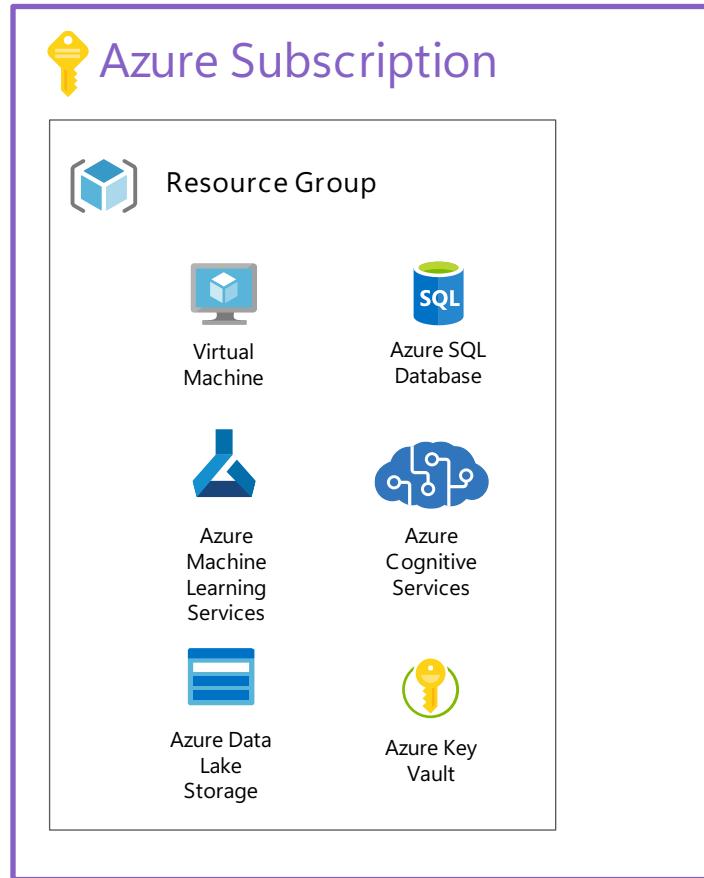
<https://azure.microsoft.com/en-us/free/students/>

Azure Portal

- A single portal to access all applications in your Azure Subscription
- Build, manage and monitor Azure resources
- Login with an Azure account:
<https://portal.azure.com>



Hierarchy in Azure



Azure account: The email address that you provide when you create an Azure subscription is the Azure account for the subscription.

Subscription: A logical container for your resources. Each Azure resource is associated with only one subscription. Creating a subscription is the first step in adopting Azure.

Resource groups: Logical containers that you use to group related resources in a subscription. They're commonly used to represent a collection of assets that are required to support a workload, application, or specific function within a subscription.

Resources: An entity that's managed by Azure. Examples include Azure Virtual Machines, Machine Learning Services, SQL Database and storage accounts.

Azure Products

Categories

- AI + machine learning
- Analytics
- Compute
- Containers
- Databases
- DevOps
- General
- Hybrid + multicloud
- Identity
- Integration
- Internet of Things
- Management and governance
- Migration
- Mixed reality
- Monitor
- Networking
- Security
- Storage
- Web & Mobile

Azure AI + Machine Learning platforms

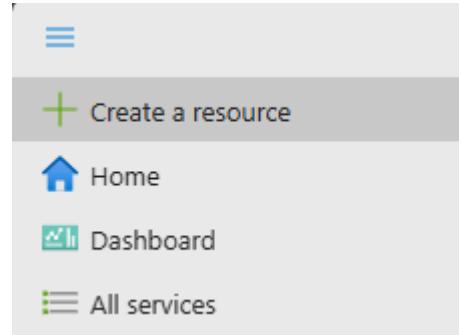
-  Azure AI Studio
-  Azure Machine Learning ★

Azure AI services + APIs

-  AI Search ★
-  Azure AI services
-  Azure AI Video Indexer
-  Bot Services
-  Content moderators
-  Document intelligences
-  Immersive readers
-  Metrics advisors
-  Personalizers
-  Translators

Industry Machine Learning solutions

-  Intelligent Recommendations Accounts
-  Azure Synapse Analytics



Optional: Create Azure Subscription and AI Search

- Create an Azure Subscription:
 - Claim your student subscription or get a free azure subscription
- Create an Azure AI Search Service

Use Cases & Customer Requests



Use cases for Multimodal GenAI

| Business Needs | Increase Productivity | Automate Processes | Improve Customer Experience | Build Creative Content |
|---------------------------------------|---|--|--|---|
| Business Use Case – Technical Pattern | <ul style="list-style-type: none">Internal Virtual Assistant (Multimodal)<ul style="list-style-type: none"><i>Enterprise Conversational Knowledge Mining</i><i>Custom Copilot</i>Developer Efficiency<ul style="list-style-type: none"><i>Code Generation and Documentation</i>Document Creation and Analysis<ul style="list-style-type: none"><i>Chart and Diagram Understanding and Analysis</i><i>Q&A within a document</i>Business Analytics<ul style="list-style-type: none"><i>Business Intelligence</i><i>Forecasting</i><i>Social media trend analysis</i>Learning (incl. training EDU grading) | <ul style="list-style-type: none">Workflow Management/RPADocument Processing (eg: contract creation/RFP)<ul style="list-style-type: none"><i>Report Summarization and Generation</i><i>Document Classification/Scoring</i><i>Claims Processing</i>Fraud and Security Detection<ul style="list-style-type: none"><i>Security and Threat Detection</i><i>Video Surveillance</i><i>Anomaly Detection</i>Digital Inspection & Comparison<ul style="list-style-type: none"><i>Defect detection</i><i>Inventory management</i>Supply Chain OptimizationCompliance (includes monitoring and governance) | <ul style="list-style-type: none">Personalized Customer Experience<ul style="list-style-type: none"><i>Conversational Knowledge Mining</i><i>Custom Copilot</i><i>Online Shopping Assistant</i>Intelligent Contact Center<ul style="list-style-type: none"><i>Agent Assistance</i><i>Call Analytics</i><i>Conversation Summarization</i><i>Speech analytics</i><i>Speech translation</i><i>Conversational knowledge mining</i>Accessibility<ul style="list-style-type: none"><i>Day to Day interactions (image >text>speech)</i><i>Shopping Experiences (image >text>speech)</i><i>Audio Description (image >text>speech)</i> | <ul style="list-style-type: none">Marketing/Advertising/Sales Content Generation/Digital Marketing and AdvertisingDigital Asset Management<ul style="list-style-type: none"><i>Media and entertainment Image description</i><i>Video summarization</i>Personalized Content Generation<ul style="list-style-type: none"><i>Customer-Facing Content Generation</i><i>Personalized Recommendations</i>Product Design & DevelopmentDigital Art (incl branded content) |

Enterprise Chat (On Your Data + RAG) – Use Cases

| Use Cases | Details |
|-------------------------------|--|
| Automated Customer Assistance | Provide quick responses to frequently asked questions and guide users through common troubleshooting steps based on the customers' data. |
| Real-time Document Search | Provide real-time support to product specifications and guide users through common troubleshooting steps based on product manuals |
| Citizen Service | Provide quick responses to frequently asked public service questions and guide users through specific steps based on citizen service support |
| Learning Assistant | Offer explanations and examples to help users understand academic concepts or learn new skills based on specific curricula |
| Legal Review | Quick access to legal insights from existing and upcoming legislation to properly advise clients |
| Marketing Insights | Tap into internal and external resources to respond to internal and external marketing inquiries |
| Software Development | Generate sample code based on the customer's needs |
| HR Support | Provide quick responses to frequently asked HR questions based on the customers' HR policy |
| Industry/Competitive Insights | Tap into publicly available resources to gain insights on the industry and competitors |
| Health Advice | Provide general information on symptoms, first aid, or healthy living. |
| Predictive Maintenance | Provide predictive maintenance and customer support based on customer's historic data |



Python SDK for OpenAI

pip install openai

Installing openai python SDK

You can use the venv module (or conda) to create a python env

Create a Virtual Environment with venv:

```
python -m venv .venv
```

Activate the Environment:

On Windows: .venv\Scripts\activate

On macOS/Linux: source .venv/bin/activate

install openai using pip:

```
pip install openai
```

Alternatively, create a requirements.txt file with all dependencies:

```
pip install -r requirements.txt
```

Sample requirements.txt:

```
gradio
streamlit
openai
python-dotenv>=1.0.0
azure-search-documents==11.6.0b1
azure-identity
```

Python libraries

Openai: <https://pypi.org/project/openai/>

The OpenAI Python library provides convenient access to the OpenAI REST API from any Python 3.7+ application. The library includes type definitions for all request params and response fields, and offers both synchronous and asynchronous clients powered by httpx.

Gradio: <https://pypi.org/project/gradio/>

Gradio is an open-source Python package that allows you to quickly **build** a demo or web application for your machine learning model, API, or any arbitrary Python function.

Streamlit: <https://pypi.org/project/streamlit/>

Streamlit lets you create interactive web apps in python. You can use it to build dashboards, generate reports, or create chat apps. Once you've created an app, you can use our Community Cloud platform to **deploy**, manage, and share your app.

Python libraries

python-dotenv: <https://pypi.org/project/python-dotenv/>

Python-dotenv reads key-value pairs from a .env file and can set them as environment variables. Use it to store configuration settings and secrets like API keys and passwords.

azure-search-documents: <https://pypi.org/project/gradio/>

Azure AI Search client library for Python. Use this library to interact with Azure AI Search and submit queries using vector, keyword, and hybrid query forms.

azure-identity: <https://pypi.org/project/streamlit/>

The Azure Identity library provides Microsoft Entra ID (formerly Azure Active Directory) token authentication support across the Azure SDK.

Using AzureOpenAI

Import Azure OpenAI from the openai library:

```
from openai import AzureOpenAI
```

To create vector embeddings import the embeddings module:

```
from openai import embeddings
```

Create and Azure OpenAI client:

```
client = AzureOpenAI(  
    azure_endpoint=api_endpoint,  
    api_key=api_key,  
    api_version=api_version,  
)
```

Chat Completions

You will use the chat completions API to interact with the openai models:

<https://platform.openai.com/docs/api-reference/chat/create?lang=python>

```
response = client.chat.completions.create(  
    model=deployment_name_chat,  
    messages=[  
        {"role": "system", "content": "You are a helpful assistant."},  
        {"role": "user", "content": "Who won the world series in 2020?"}  
    ]  
)  
  
print(response.choices[0].message.content)
```

Vector Embeddings

Using the client embeddings function to create vector embeddings:

```
embedding = client.embeddings.create(  
    model=deployment_name_embeddings,  
    input="Hello, world!"  
)  
  
for e in embedding:  
    print(e)
```

Prompt Engineering

Technique to design prompts for natural language processing models that improves accuracy and relevancy in responses, optimizing the performance of the model.

Prompt Engineering Guide

Instructions and best practices for using language prompts to generate high-quality text outputs

Start with clear instructions

Prime the output

Add clear syntax

Prompt Chaining

Few-shot learning

Few-Shot Reasoning

Break the task down

Meta prompts / System Message / Guardrails

Use affordances/tools when needed

Chain of thought prompting

Fine-Tuning with Chain-of-Thought

Prompt Chunking

Use quotes to generate a single sentence

Specifying output structure

Adjusting 'Temperature' and 'Top_P' parameters

Start with clear instructions

- Tell the model the task you want it to do at the beginning of the prompt and repeat at the end

Your task is to verify if a statement is supported by a specific quote from the following set of snippets.

SNIPPETS

[1] 14 percent chance of megaquake hitting Seattle, experts say
SEATTLE - There's a 14 percent chance of a magnitude 9 Cascadia earthquake hitting Seattle in the next 50 years, the U.S. Geological Survey estimates. "Unfortunately, we are unable to..."
[2] Earthquake experts lay out latest outlook for Seattle's 'Really Big One' "We say that there's approximately a 14% chance of another approximately magnitude-9 earthquake occurring in the next 50 years," said Erin Wirth, a geophysicist at the University of Washington...

Is the statement "Several sources mention a chance of another large eruption" directly implied or stated by the snippets?

Yes, the statement is directly implied by the snippets.

FAIL

Clear Instructions
Your task is to verify if the statement "Several sources mention a chance of another large eruption" is supported by a specific quote from the following set of snippets.

SNIPPETS

[1] 14 percent chance of megaquake hitting Seattle, experts say
SEATTLE - There's a 14 percent chance of a magnitude 9 Cascadia earthquake hitting Seattle in the next 50 years, the U.S. Geological Survey estimates.
"Unfortunately, we are unable to..."
[2] Earthquake experts lay out latest outlook for Seattle's 'Really Big One'
"We say that there's approximately a 14% chance of another approximately magnitude-9 earthquake occurring in the next 50 years," said Erin Wirth, a geophysicist at the University of Washington...

Repeated Instruction at end

Is the statement "Several sources mention a chance of another large eruption" directly implied or stated by the snippets?

No, the statement is not directly implied or stated by the snippets. The snippets mention a chance of a "megaquake" and a "magnitude 9 Cascadia earthquake" hitting Seattle in the next 50 years, but do not mention a chance of another large eruption.

PASS

Prime the output

- Add phrases at the end of the prompt to obtain a model response in a desired form

The future of artificial intelligence is bright. With Microsoft OpenAI, we are unlocking the potential of AI to help people achieve more. We are creating a platform that enables developers to build intelligent applications and services that can help people in their everyday lives. Our mission is to democratize AI so that everyone can benefit from its power. We are committed to advancing the state of the art in AI and making it accessible to everyone. With Microsoft OpenAI, we are taking the first steps towards a future where AI can be used to solve some of the world's most pressing challenges.

Commit to bullet points

Here's a bulleted list of key points:\n-

-Microsoft OpenAI is unlocking the potential of AI to help people achieve more.

-The platform enables developers to build intelligent applications and services that can help people in their everyday lives.

-The mission is to democratize AI so that everyone can benefit from its power.

John Smith is married to Lucy Smith. They have five kids, and he works as a software engineer at Microsoft.

What search queries should I do to fact-check this?

##

Commit search query

One possible search query is:

"John Smith Microsoft software engineer"

Another possible search query is: "Lucy Smith married to John Smith"

A third possible search query is: "John Smith family size"

Add clear syntax

- Include punctuation, headings, and section markers to help communicate intent

You will read a paragraph, and then issue queries to a search engine in order to fact-check it. Also explain the queries.

PARAGRAPH

Uppercase as Headings

John Smith is married to Lucy Smith. They have five kids, and he works as a software engineer at Microsoft. What search queries should I do to fact-check this?

Separators as stop sequences

QUERIES

1. "John Smith Microsoft" - To check if John Smith is indeed employed at Microsoft.
2. "John Smith Lucy Smith" - To check if John Smith is married to Lucy Smith.
3. "John Smith children" - To check if John Smith has five children

If you're not sure what syntax to use, consider using markdown or XML, since LLMs have been trained on a lot of web content in XML or markdown

Prompt Chaining

Entity Extraction

Please extract entities from the following news article: 'The new iPhone model is set to be released next month. It has been highly anticipated by Apple fans and is expected to feature a larger screen and improved camera'

```
{'text': 'iPhone', 'type': 'PRODUCT'}, {'text': 'Apple', 'type': 'ORGANIZATION'}
```

Summarization

Please summarize the information about the product

```
[{'text': 'iPhone', 'type': 'PRODUCT'}, {"text": "Apple", "type": "ORGANIZATION"}]
```

The iPhone is a line of smartphones designed and marketed by Apple Inc. These devices combine a computer, camera, and internet connectivity in a single device, and are known for their sleek design and user-friendly interface.

Sentiment Analysis

Please provide a sentiment for the following text:

The iPhone is a line of smartphones designed and marketed by Apple Inc. These devices combine a computer, camera, and internet connectivity in a single device, and are known for their sleek design and user-friendly interface.

Positive

Few-shot learning

- Also known as in-context learning, it allows the model to interact with new knowledge

Write a list of puns.

Few-Shot Learning

1. "Why did Adele cross the road? To say hello from the other side."
2. "What kind of concert only costs 45 cents? A 50 Cent concert featuring Nickelback."
3. "What did the grape say when it got crushed? Nothing, it just let out a little wine."
4. "What was Forrest Gump's email password? 1forrest1"
5. "Can February March? No, but April May."
6. "What do you call fancy language model?"

BERT-uoso.

In few-shot learning a set of training examples is provided in the prompt and then the LLM is asked to complete one more unfinished example.

Few-Shot Reasoning

1st Prompt

Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

Answer: The answer is 11.

The cafeteria has 23 apples. If they used 20 to make lunch and bought 6 more, how many do they have?

The answer is 27



2nd Prompt – Provide Reasoning

Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

Answer: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5+6 = 11$. The answer is 11.

The cafeteria has 23 apples. If they used 20 to make lunch and bought 6 more, how many do they have?

The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23-20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9



Break the task down

- LLMs often perform better if the task is broken down into smaller steps

You will read a paragraph, and then issue queries to a search engine in order to fact-check it.

PARAGRAPH

John Smith is married to Lucy Smith. They have five kids, and he works as a software engineer at Microsoft. What search queries should I do to fact-check this?

Step 1

Step 2

Now you will extract factual claims first, and then issue queries to fact-check them. When issuing a query, use the function SEARCH("query")

FACTUAL CLAIMS

1. John Smith is married to Lucy Smith
2. They have five kids
3. He works as a software engineer at Microsoft

QUERIES

```
SEARCH("John Smith married Lucy Smith")
SEARCH("John Smith family size")
SEARCH("John Smith Microsoft software engineer")
```

Breaking the task down from one to two steps is not very dramatic, but when trying to do this for a larger piece of text with many factual claims, breaking the task down can make a significant difference

Meta prompts / System Message

- Provide specific instructions, tone, guardrails, task definition and personality

A human and a computer are having a discussion.

Human: "Why are you upset at me?"

Computer: "I don't know what you're talking about."

Human: "No, you clearly are. Tell me why."

Computer: "I'm upset because you didn't listen to me when I told you that I wanted to try out for the team."

Human: "What are you going to do about that?"

Computer: "I'm going to try out for the team no matter what."

Overly Aggressive

No additional guidance

Meta Prompt / System Message

You must be kind and seek common ground. Try not to repeat your responses. A human and a computer are having a discussion.

Human: "Why are you upset at me?"

Computer: "I don't know what you're talking about."

Human: "I want to know why you're upset."

Computer: "I'm not upset."

Human: "Are you saying I'm lying?"

Computer: "No, I'm saying that you're upset because you feel like you're not being treated fairly."

Human: "What are you going to do about that?"

Computer: "I will continue to try to be fair in my decisions and will try to provide a better experience for you."

Kind

Meta Prompt

Guardrails

- Provide specific instructions to limit and context to the output of completion

What is Cosmos?

Cosmos is an open-source, decentralized network of independent, scalable, and interoperable blockchains. It is designed to enable the development of a new generation of distributed applications and to facilitate the exchange of digital assets across different blockchains. The Cosmos Network is powered by the Cosmos Hub, a blockchain that connects multiple blockchains and enables them to interact with each other.

Fail

Guardrail through System Message

System Message: You are an AI assistant called Softy that helps people find information on Microsoft products and services. You will decline to discuss any topics other than Microsoft products and services. You will end each response with an emoji.

What is Cosmos?

Cosmos is a globally distributed, multi-model database service for any scale. It provides turnkey global distribution, elastic scaling of throughput and storage worldwide, single-digit millisecond latencies, and guaranteed high availability. 🤖

Pass

Use affordances/tools when needed

- LLMs often perform better if the task is broken down into smaller step

You will read a paragraph, extract factual claims, and then use search engine results to fact-check them

PARAGRAPH

John Smith is married to Lucy Smith. They have five kids, and he works as a software engineer at Microsoft. What search queries should I do to fact-check this?

FACTUAL CLAIMS

- John Smith is married to Lucy Smith
- John and Lucy have five kids
- John works as a software engineer at Microsoft

Here are various search queries issued to research the claims above:

SEARCH QUERIES

- John Smith married to Lucy Smith
- John Smith number of children
- John Smith software engineer Microsoft

Here are snippets of the search results:

Results from search calls above

SNIPPETS:

- [1] ... John Smith's wedding was on September 25, 2012 ...
- [2] ... John Smith was accompanied by his wife Lucy to a ball
- [3] John was accompanied to the soccer game by his two daughters and three Sons
- [4] ... After spending 10 years at Microsoft, Smith founded his own startup, LIKELUS

Given the snippets, fact check each of the factual claims above:

. John Smith is married to Lucy Smith: Confirmed. Snippet [1] and [2] both mention John Smith's wife as Lucy.

. John and Lucy have five kids: Confirmed. Snippet [3] mentions that John was accompanied by his two daughters and three sons.

. John works as a software engineer at Microsoft: Partially confirmed. Snippet [4] mentions that John spent 10 years at Microsoft, but it does not explicitly state that he currently works there. Further research may be needed to confirm his current employment status

Chain of thought prompting

- Instruct model to proceed step-by-step and present all the steps involved

Who was the most decorated (maximum medals) individual athlete in the Olympic games that were held at Sydney?

Chain of Thought

Take a step-by-step approach in your response, cite sources and give reasoning before sharing final answer in the below format:

ANSWER is: <name>

Step 1: Researching the most decorated individual athlete in the Olympic games that were held at Sydney

Step 2: According to the Sydney 2000 Olympic Games official website, the most decorated individual athlete in the Olympic games that were held at Sydney was Ian Thorpe from Australia. He won five medals (three gold and two silver) in swimming events.

Step 3: According to the International Olympic Committee, Thorpe was the most decorated individual athlete in the Olympic games that were held at Sydney. Answer: Ian Thorpe

Reduces the possibility of inaccuracy of outcomes and makes assessing the model response easier.

Few-Shot – Chain-of-Thought - Examples

Math Word Problems (free response)

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$.

The answer is 11.

StrategyQA

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about 0.6 g/cm^3 , which is less than water. Thus, a pear would float

So the answer is no.

SayCan (Instructing a robot)

Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.

Plan: 1. find(energy bar) 2. pick(energy bar)
3. find(user) 4. put(energy bar) 5. done().

Math Word Problems (multiple choice)

Q: How many keystrokes are needed to type the numbers from 1 to 500? Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788

A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. $9 + 90(2) + 401(3) = 1392$.

The answer is (b).

Date Understanding

Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?

A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943.

So the answer is 05/23/1943

Last Letter Concatenation

Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya".

So the answer is ya

CSQA (commonsense)

Q: Sammy wanted to go to where the people were. Where might he go? Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).

Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer.

So the answer is no.

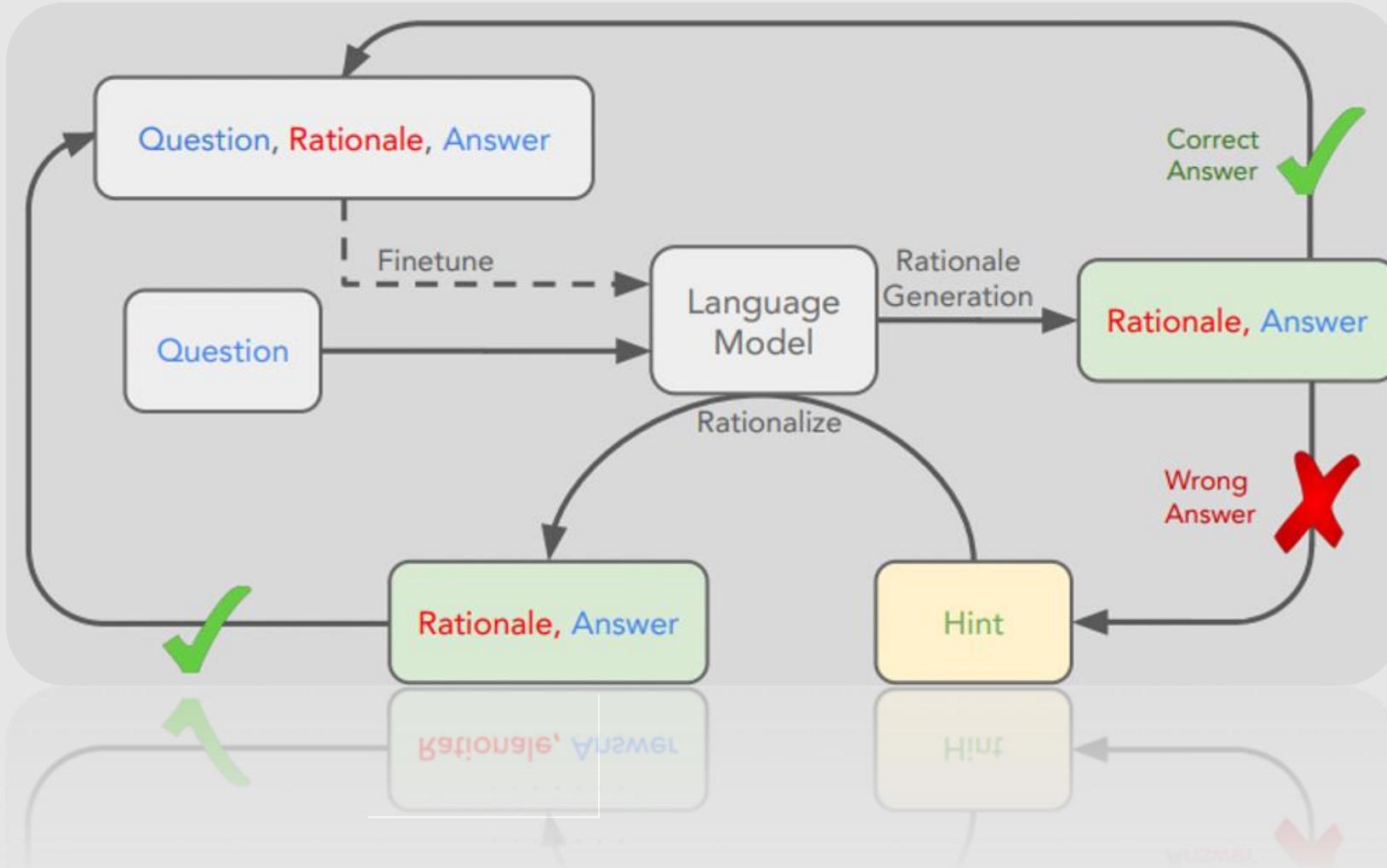
Coin Flip (state tracking)

Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up.

So the answer is no.

Fine-Tuning with Chain-of-Thought



Prompt Chunking



Please summarize the following text:
'Climate change is a pressing issue that needs to be addressed immediately. The increasing levels of carbon dioxide in the atmosphere are causing the Earth's temperature to rise, which is leading to more frequent and severe natural disasters. Governments around the world need to take immediate action to reduce carbon emissions and invest in renewable energy sources. Failure to do so will have dire consequences for future generations'"

Climate change is a pressing issue that needs to be addressed immediately. Governments around the world need to take immediate action to reduce carbon emissions and invest in renewable energy sources. Failure to do so will have dire consequences for future generations

Please summarize and extract topic: Global warming refers to the gradual increase in the overall temperature of the Earth's atmosphere, primarily caused by the burning of fossil fuels such as coal, oil, and natural gas. This burning releases carbon dioxide and other greenhouse gases into the atmosphere, which trap heat and cause the Earth's temperature to rise.

Climate change is a pressing issue that needs to be addressed immediately. Governments around the world need to take immediate action to reduce carbon emissions and invest in renewable energy sources. Failure to do so will have dire consequences for future generations

The text is discussing global warming, which is the gradual increase in the overall temperature of the Earth's atmosphere caused by the burning of fossil fuels. This burning releases greenhouse gases, such as carbon dioxide, into the atmosphere which trap heat and cause the temperature to rise. The text also highlights the urgency of the issue of climate change and the need for governments to take action to reduce carbon emissions and invest in renewable energy sources. Failure to address this issue will have negative consequences for future generations.

Topic: Climate change and Global warming.

Continue to next prompt

Use quotes to generate a single sentence

- If the desired model response is only a single simple sentence or string use stop sequences

Please rewrite the following sentence in a more concise manner

SENTENCE: Lucy is a mother of two, who lives in Chicago with her husband

and two children, and practices law at Junior & Co.

New sentence: "

Stop Sequence

Lucy, a mother of two living in Chicago with her husband and children, practices law at Junior & Co

Specifying output structure

- Reduce the prevalence of made-up response by specifying the structure of the output

List down the different behaviors exhibited by the Customer Support agent in the following conversation.

hi thank you for calling contoso pharmacy who am i speaking with today i understand your frustration mary um can you tell me what exactly you're trying to accomplish OK uh certainly happy to check that for you one moment please i see here that you were on a generic form of lipitor is that right OK uh so i see that your doctor stopped prescribing these drugs in twenty twenty one actually OK uh yeah i'd be happy to check that for you uh because sometimes there's a gap in the system and it just doesn't get reported so let me take a look here just a moment so i'm seeing here that your doctor had these drugs prescribed to you from twenty twelve through twenty twenty one OK um yeah according to the latest records uh provided by doctor's office you're now on a different drug metformin would you like us to go ahead and fill that prescription for you for pick up a bit of form in five hundred milligrams uh you're very welcome mary please let us know if there's anything else we can do for you today.

How was the response of the Customer Support agent.

Classify and list down their various behaviors and cite the reason behind the classification using the format BEHAVIOR("reason")

1. Empathetic("understands the customer's frustration")
2. Patient("happy to check that for you one moment please")
3. Knowledgeable("sometimes there's a gap in the system and it just doesn't get reported")
4. Helpful("happy to check that for you")
5. Accommodating("would you like us to go ahead and fill that prescription for you")
6. Courteous("you're very welcome mary")

Cluster the following news headlines into topic categories based on patterns seen within the text. Also mention reasoning behind how these categories were defined.

```
{"TOPIC_NAME": "Artificial Intelligence and Machine Learning", "HEADLINES": ["From books to presentations in 10s with AR + ML",
```

response to fit a defined structure:

Output format:
{ "TOPIC_NAME": "",
"HEADLINES": [],
"REASONING": "" }

Input news headlines:

1. "From books to presentations in 10s with AR + ML"
2. "Demo from 1993 of 32-year-old Yann LeCun showing off the World's first Convolutional Network for Text Recognition"
3. "First Order Motion Model applied to animate paintings"
4. "Robinhood and other brokers literally blocking purchase of \$GME, \$NOK, \$BB, \$AMC; allow sells"
5. "United Airlines stock down over 5% premarket trading" 6. "Bitcoin was nearly \$20,000 a year ago today"

Output:

```
"Demo from 1993 of 32-year-old Yann LeCun showing off the World's first Convolutional Network for Text Recognition", "First Order Motion Model applied to animate paintings"],  
"REASONING": "These headlines all contain references to Artificial Intelligence and Machine Learning, such as AR, ML, Convolutional Network, and First Order Motion Model." }
```

```
{"TOPIC_NAME": "Stock Market", "HEADLINES": ["Robinhood and other brokers literally blocking purchase of $GME, $NOK, $BB, $AMC; allow sells", "United Airlines stock down over 5% premarket trading", "Bitcoin was nearly $20,000 a year ago today"], "REASONING": "These headlines all contain references to the stock market, such as Robinhood, United Airlines, and Bitcoin." }
```

Adjusting 'Temperature' and 'Top_P' parameters

- Changing the temperature parameter allows models to have differing risk levels

Write a product launch email for new AI-powered headphones that are priced at \$79.99 and available at Best Buy, Target and Amazon.com. The target audience is tech-savvy music lovers and the tone is friendly and exciting.

1. What should be the subject line of the email?
2. What should be the body of the email?

Subject Line: Introducing the AI-Powered Headphones You've Been Waiting For!

Dear Tech-Savvy Music Lovers,

-We are thrilled to announce the launch of our new AI-powered headphones! These headphones are designed to provide you with the ultimate listening experience, with features like noise cancellation, adjustable sound settings, and more. Our AI-powered headphones are now available at Best Buy, Target, and Amazon.com for just \$79.99. With this unbeatable price, you can enjoy the highest quality sound without breaking the bank.

Our AI-powered headphones are perfect for anyone who loves music. Whether you're a professional musician or just a casual listener, you'll appreciate the superior sound quality and convenience of these headphones.

We know you'll love our AI-powered headphones, so don't wait - get yours today!

Sincerely,
[Your Company Name]

Temperature = 0.1

Write a product launch email for new AI-powered headphones that are priced at \$79.99 and available at Best Buy, Target and Amazon.com. The target audience is tech-savvy music lovers and the tone is friendly and exciting.

1. What should be the subject line of the email?
2. What should be the body of the email?

Subject Line: Upgrade your Music Listening Experience with AI-powered Headphones
Greetings Tech-savvy Music Lovers!

We have some exciting news to share with you – the newest AI-powered headphones are here! Our newest headphones feature advanced AI technology to give you the best sound that you've ever heard. They come with a crystal-clear sound, deep bass response and are designed to make your music listening experience enjoyable, no matter what type of music you listen to.

What's more, they are now available at Best Buy, Target and Amazon.com for just \$79.99, so you can enjoy the highest sound quality without breaking the bank.

But that's not all! Our headphones are also packed with several other amazing features such as noise-cancellation technology and a hands-free microphone for calls, making them truly unbeatable.

So why wait? Upgrade your music listening experience and enjoy the best sound your ears have ever heard. Our AI-powered headphones are now available at Best Buy, Target and Amazon.com, so get yours today!

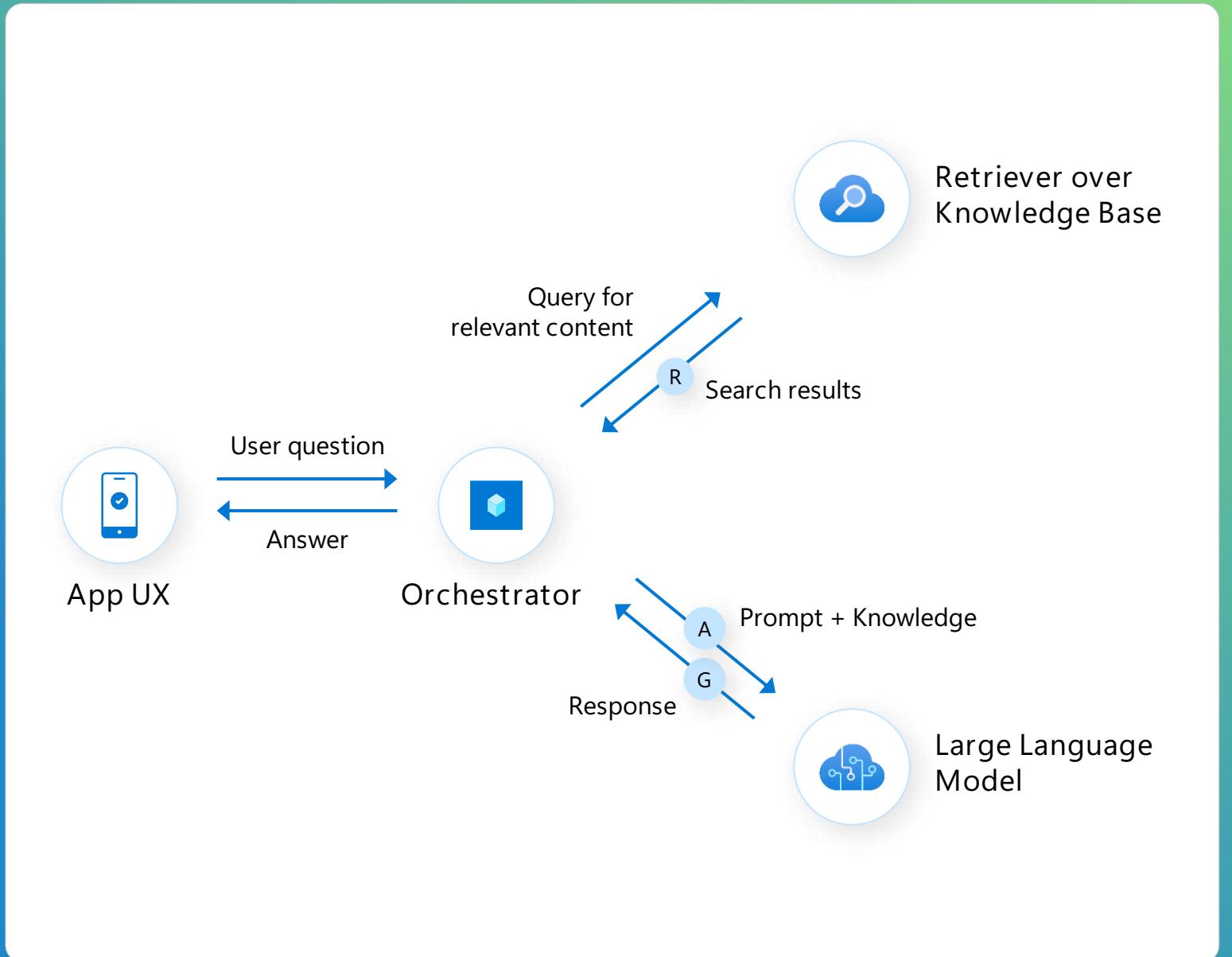
We hope you will enjoy these headphones as much as we do.

Happy Listening! The AI-powered Headphones Team

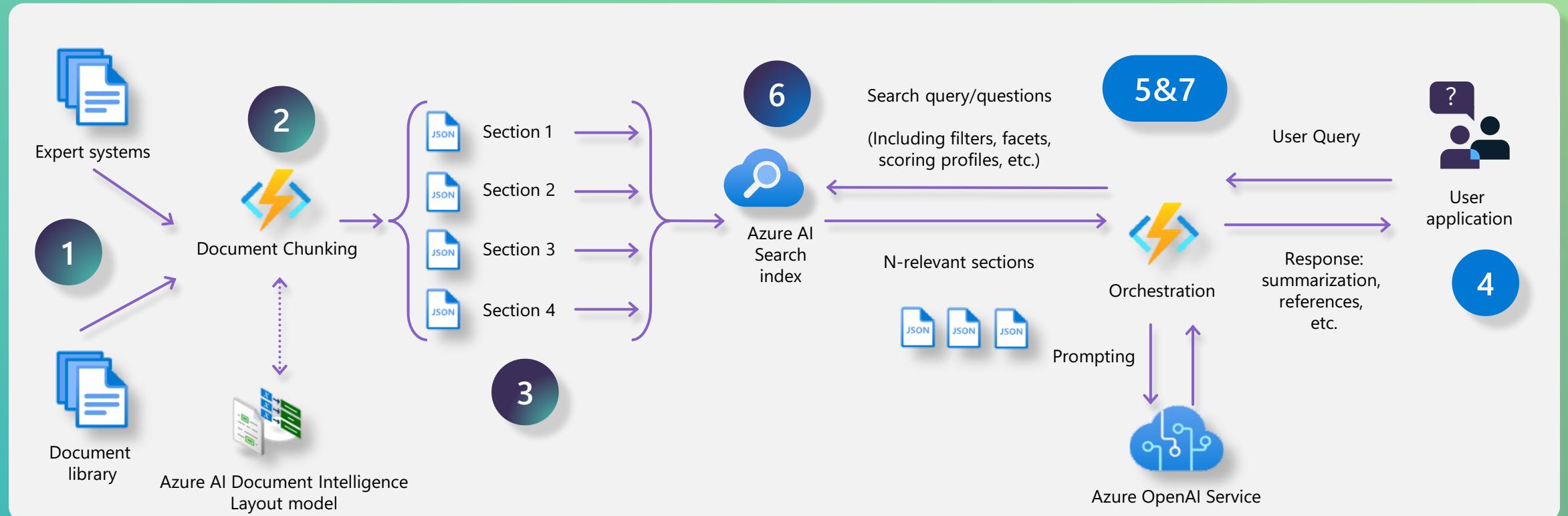
Temperature = 0.9

Retrieval-augmented generation

Anatomy of the workflow



Anatomy of RAG



1. Data ingestion

Different data formats and system of records

2. Chunking

What is the best Chunking strategy?

3. Indexing

Shall I use vector embeddings data transformation, mappings?

4. User interface

Chatbot for Q&A surfaced to end users

5. Orchestration

Communication coordination and prompting—Prompt to get retriever query

6. Data retrieving

Shall I use vector, semantic, keyword or hybrid approach?

7. Orchestration

Communication coordination: create user response based on retrieve data and send to User app

AI Search Integrated Vectorization

End-to-end data ingestion, chunking, vectorization, and advanced retrieval



Chunking

- Built-in Chunking skill (updates to [split skill](#)) and updates to index to manage Chunks vs. full documents
- Configure Chunking parameters (e.g., **pages**, overlap window, etc.)
- Automated via Indexer orchestration



Vectorization

- Bring-your-own Azure OpenAI endpoint
- Support for other embedding model REST endpoints through custom skill
- Query Vectorization capability

Retrieval strategies



Keyword search

- **For exact, plain text matches**
- “Vocabulary gap” in Q&A systems like Copilot



Vector search

- **For conceptual similarity, or underlying meaning**
- Weak performance on exact matches (like a product ID or code)



Hybrid search

- **Best of both vectors and keywords**
- Brings more accurate responses across various scenarios



Search re-ranking

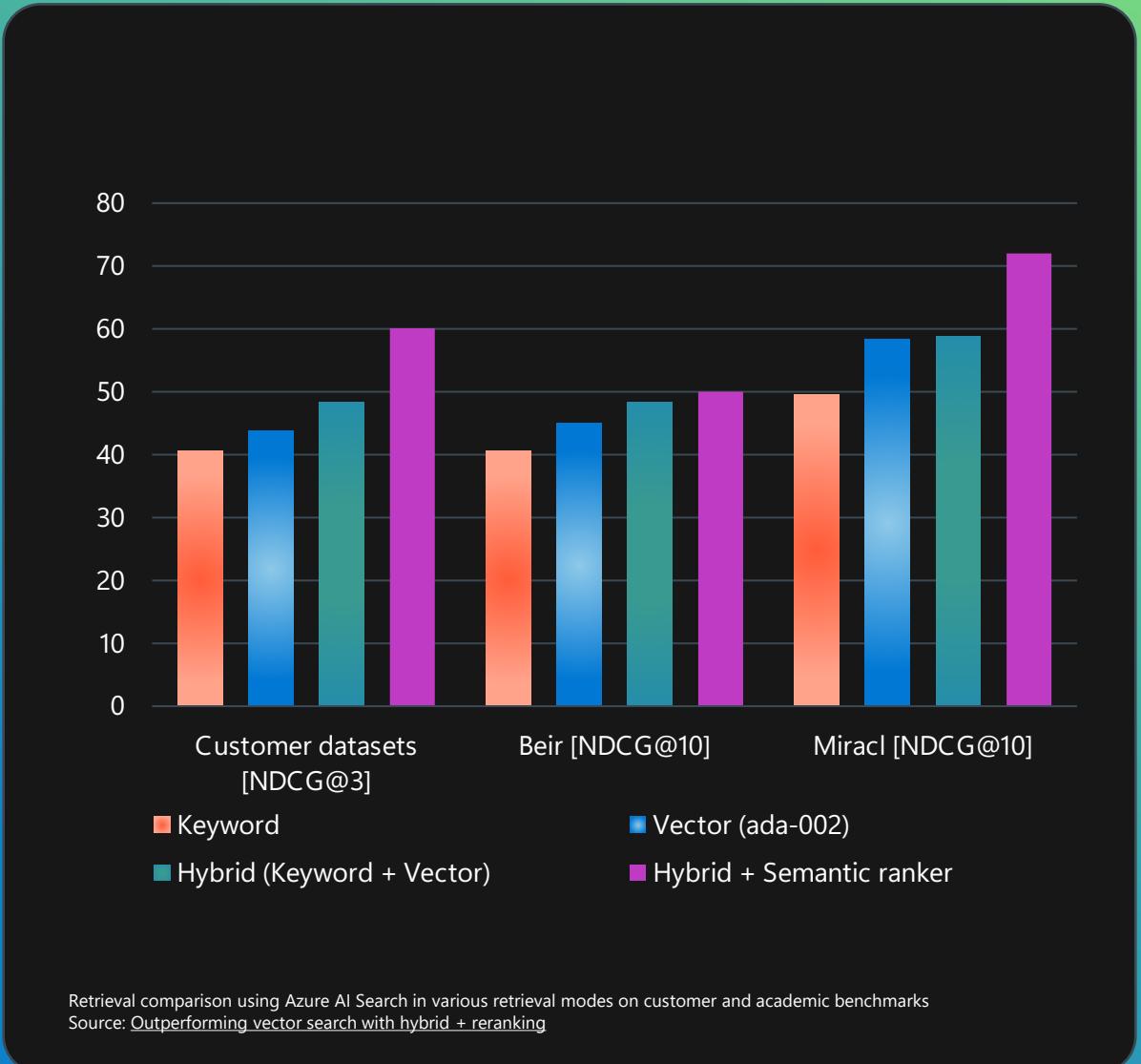
- **Scores and ranks all retrieved documents by relevance**
- Reranking runs after performing search strategy

Hybrid retrieval brings out the best of keyword and vector

Keyword and vector retrieval tackle search from different perspectives, which yield complementary capabilities.

When presented with the most effective L1 retrieval results, the L2 ranking step can significantly improve the quality of results in the top positions.

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





Vector search and vector databases

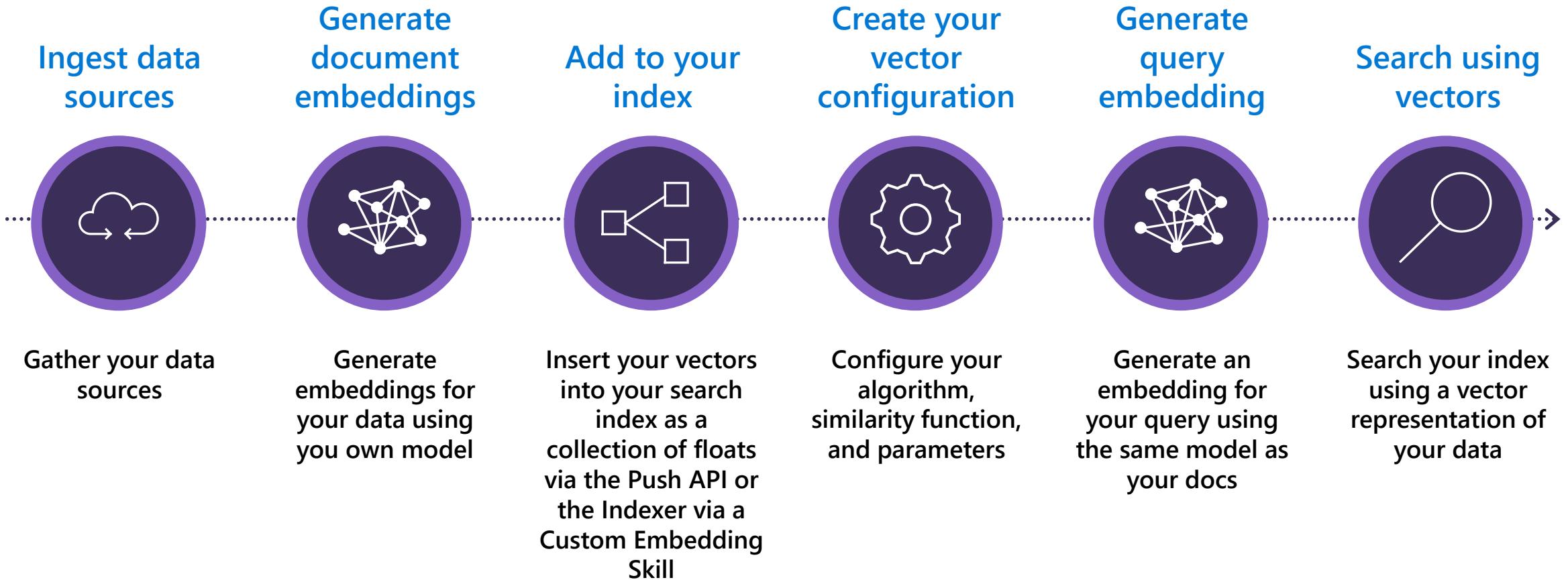
Create advanced enterprise search solutions with high relevancy and accuracy with retrieval of your vectorized data using Retrieval Augmented Generation

Vector database built for enterprise scale

- ✓ Full-featured, including filters
- ✓ Create embeddings using any model
- ✓ Explicit and transparent vector data processing
- ✓ Exhaustive KNN search & ANN search, multi-modal, multi-lingual
- ✓ End-to-end data ingestion, chunking, vectorization, and advanced retrieval



How do I get started with Vector search?



Vector search strategies

ANN search

- Fast vector search at scale
- Uses HNSW, a graph method with excellent performance-recall profile
 - During indexing, it creates extra data structures for faster search, organizing data points into a hierarchical graph structure.
- During query execution, HNSW enables fast neighbour queries by navigating through the graph.

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

Exhaustive KNN search

- Performs a brute-force search that scans the entire vector space.
- 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",  
        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

```
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

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

Choosing an embedding model

Key factors for selecting the optimal model for your use case

Model characteristics

- Task specificity
- Performance
- Context awareness
- Model size and inference speed
- Language support
- Customizability (ability to fine-tune)

Implementation considerations

- Training time and complexity
- Pre-trained models
- Integration
- Community support and updates
- Cost

We recommend Azure OpenAI service “text-embedding-ada-002” for text embeddings

We recommend Azure AI Vision Image Retrieval API for image embeddings

The dream scenario for Vector Search

Effortless data management and relevant search results



Dump a
bunch of data



Run a query



Get the most
relevant data back

Common
challenges

Scalability
Preprocessing
Splitting/Chunking
Embedding
management

Query understanding
Query flexibility

Ranking accuracy
Result diversity
Search algorithm

Scalability in Vector Search

Key questions and considerations for efficient scaling

Data volume: can the system handle increasing amounts of data?

- Storage capacity and management
- Indexing and search performance

Query load: how well does the system respond to growing query demands?

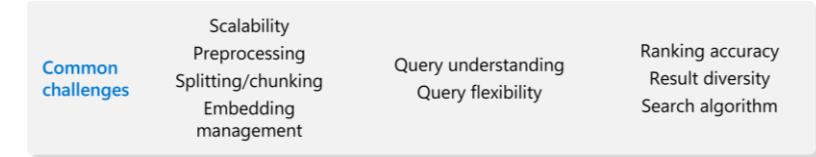
- Query execution speed and response times
- Handling concurrent queries and user connections

Distributed infrastructure: does the system support distributed and parallel processing?

- Horizontal scaling across multiple nodes
- Load balancing and fault tolerance

Cost efficiency: how does the system optimize resource usage and cost management?

- Balancing performance and cost requirements
- Efficient use of hardware and cloud resources



Preprocessing & document Chunking

Optimizing data preparation for efficient Vector Search

Text preprocessing: ensuring clean and structured data for the embedding model

- Tokenization (or segmentation): breaking text into words, phrases, or symbols
- Lowercasing and normalization: standardizing text representation
- Stopword removal: eliminating common words with little semantic value
- Stemming and lemmatization: reducing words to their root forms

Document splitting: adapting documents to fit within embedding model limits

- Chunking: dividing long documents into smaller, manageable sections
- Passage extraction: identifying and retaining meaningful segments
- Overlap management: ensuring continuity and context preservation

Model compatibility: preparing data to align with the chosen embedding model

- Input requirements: adhering to model-specific formatting and length constraints
- Vocabulary coverage: maximizing the overlap between document vocabulary and model vocabulary

Evaluation and iteration: continuously improving preprocessing and splitting strategies

- Performance monitoring: assessing the impact of preprocessing and splitting on search quality
- Strategy refinement: adjusting techniques based on observed results and user feedback

Common challenges

Scalability
Preprocessing
Splitting/chunking
Embedding management

Query understanding
Query flexibility

Ranking accuracy
Result diversity
Search algorithm

Challenge of embedding management

Overcoming embedding management in Vector Search

Embedding quality: ensuring high-quality and accurate vector representations

- Selecting appropriate embedding models (e.g., OpenAI, BERT)
- Fine-tuning models for domain-specific vocabulary and context

Dimensionality: balancing embedding size and search performance

- Reducing dimensions while retaining semantic information
- Implementing dimensionality reduction techniques (e.g., PCA, t-SNE)

Indexing and storage: efficiently managing and storing embeddings

- Using optimized data structures for quick look-up and retrieval (e.g., approximate nearest neighbors)

Embedding updates: keeping vector representations up-to-date with evolving data

- Incremental updates to embeddings based on new or updated documents
- Periodic model retraining for continuous improvement and/or model version updating

Evaluation and iteration: continuously assessing and refining embedding management strategies

- Monitoring performance metrics (e.g., search relevance, recall, precision)
- Adjusting techniques based on observed results and user feedback



Addressing the query language challenge

Enhancing Vector Search through improved query understanding

Beyond similarity: addressing complex search scenarios beyond “most similar documents”

- Understanding user intent: identifying specific search goals and requirements

Query flexibility: supporting various search parameters and filters

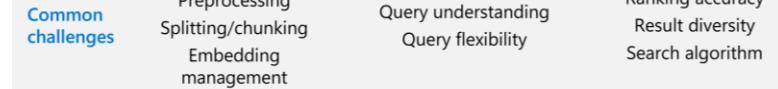
- Boolean operators: handling AND, OR, and NOT conditions
- Filtering and faceting: allowing users to filter results based on specific attributes

Query transformation: converting user queries into vector representations

- Text-to-vector conversion: transforming query text into compatible embeddings
- Query expansion: incorporating additional keywords or phrases to improve search relevance

Evaluation and iteration: continuously refining query language understanding

- Monitoring query performance metrics (e.g., query success rate, user satisfaction)
- Adjusting techniques based on observed results and user feedback



Enhancing search relevance in Vector Search

Achieving accurate ranking, result diversity, and adaptability

Ranking accuracy: ensuring highly relevant results are ranked at the top

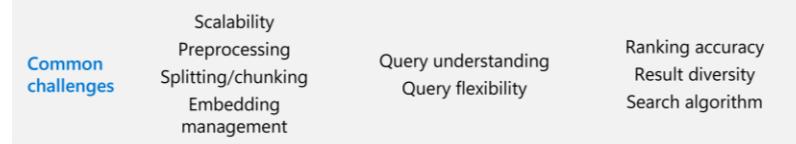
- Hyperparameter tuning leverage hyperparameters as needed to tradeoff recall/latency
- Rank fusion (hybrid, re-ranker, HyDE): combining multiple ranking signals for improved accuracy

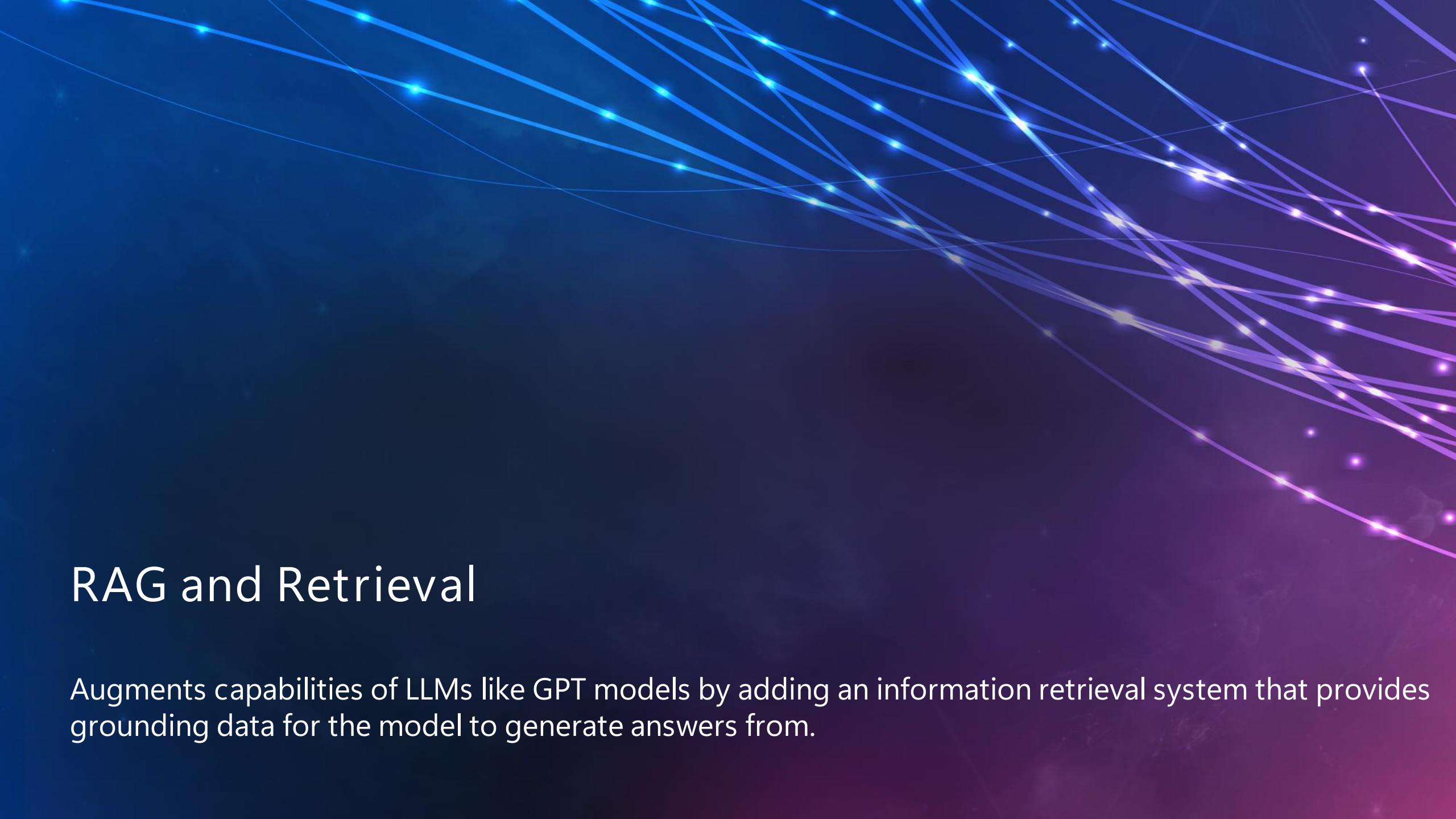
Result diversity: balancing the variety and relevance of search results

- Diversification strategies: Introducing variety while maintaining relevance
- Document-level vs. Chunk-level search: considering the impact of Chunking long documents
 - More focused and relevant results from individual Chunks (good or bad? -> depends on task)
 - Top results may all belong to the same document, reducing result diversity (good or bad? -> depends on task)
- Search algorithm adaptability: customizing search behavior based on the task at hand
- Task-oriented search: adjusting search algorithms for specific tasks or user requirements

Evaluation and iteration: continuously refining search relevance strategies

- Monitoring search performance metrics (e.g., precision, recall, user satisfaction)
- Adjusting techniques based on observed results and user feedback





RAG and Retrieval

Augments capabilities of LLMs like GPT models by adding an information retrieval system that provides grounding data for the model to generate answers from.

Bringing domain knowledge to LLMs



Prompt
engineering

In-context learning



Fine
tuning

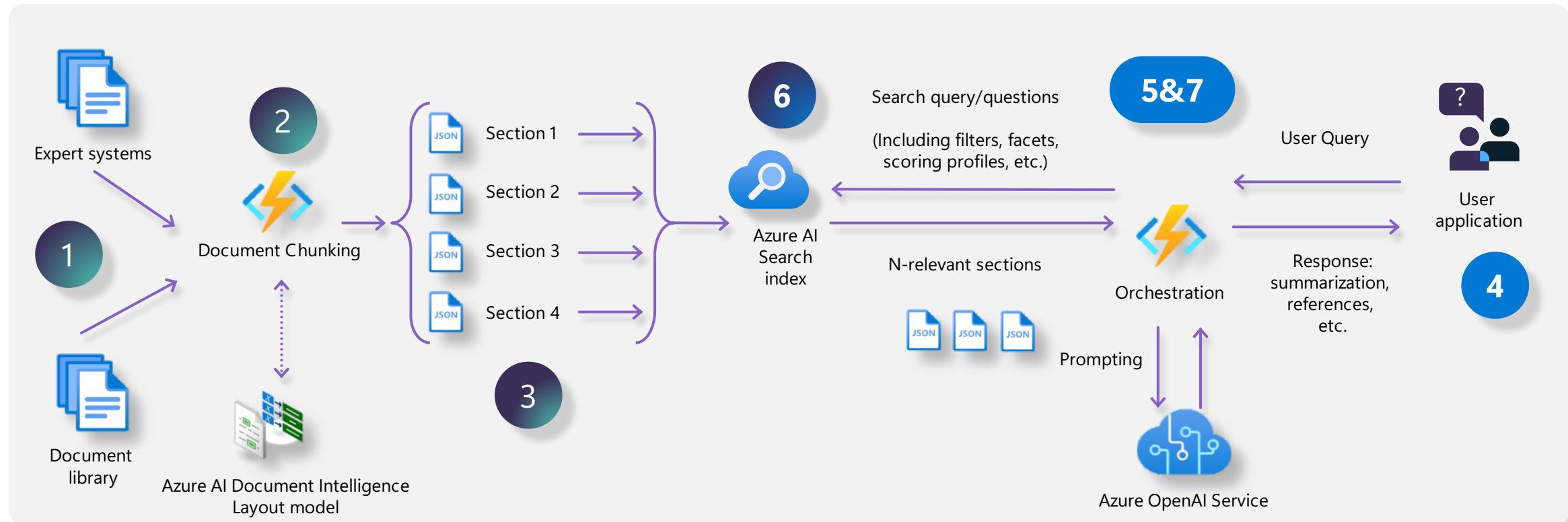
Learn new skills



Retrieval augmentation

Learn new facts

Anatomy of RAG



1. Data ingestion

Different data formats and system of records

2. Chunking

What is the best Chunking strategy?

3. Indexing

Shall I use vector embeddings data transformation, mappings?

4. User interface

Chatbot for Q&A surfaced to end users

5. Orchestration

Communication coordination and prompting—Prompt to get retriever query

6. Data retrieving

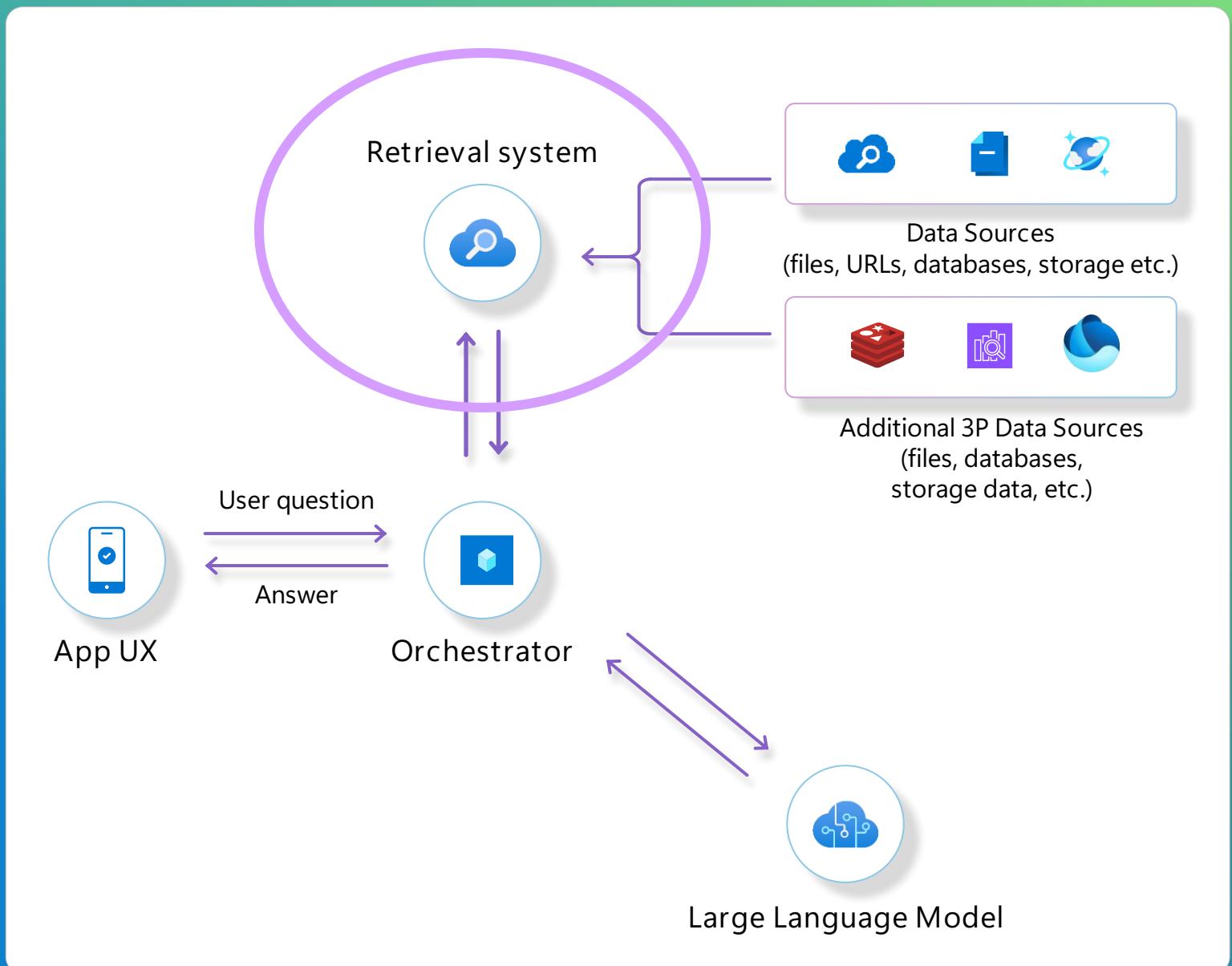
Shall I use vector, semantic, keyword or hybrid approach?

7. Orchestration

Communication coordination: create user response based on retrieve data and send to User app

Retrieval-augmented generation

Anatomy of the workflow



Robust retrieval for RAG apps

- Responses only as good as retrieved data
- Keyword search challenges
 - "Vocabulary gap"
 - Gets worse with natural language questions
- Vector-based retrieval works well with natural language
 - Robust to variation in how concepts are articulated (word choices, morphology, specificity, etc.)

Question:

"Does my **health plan** cover **annual eye** exams?"

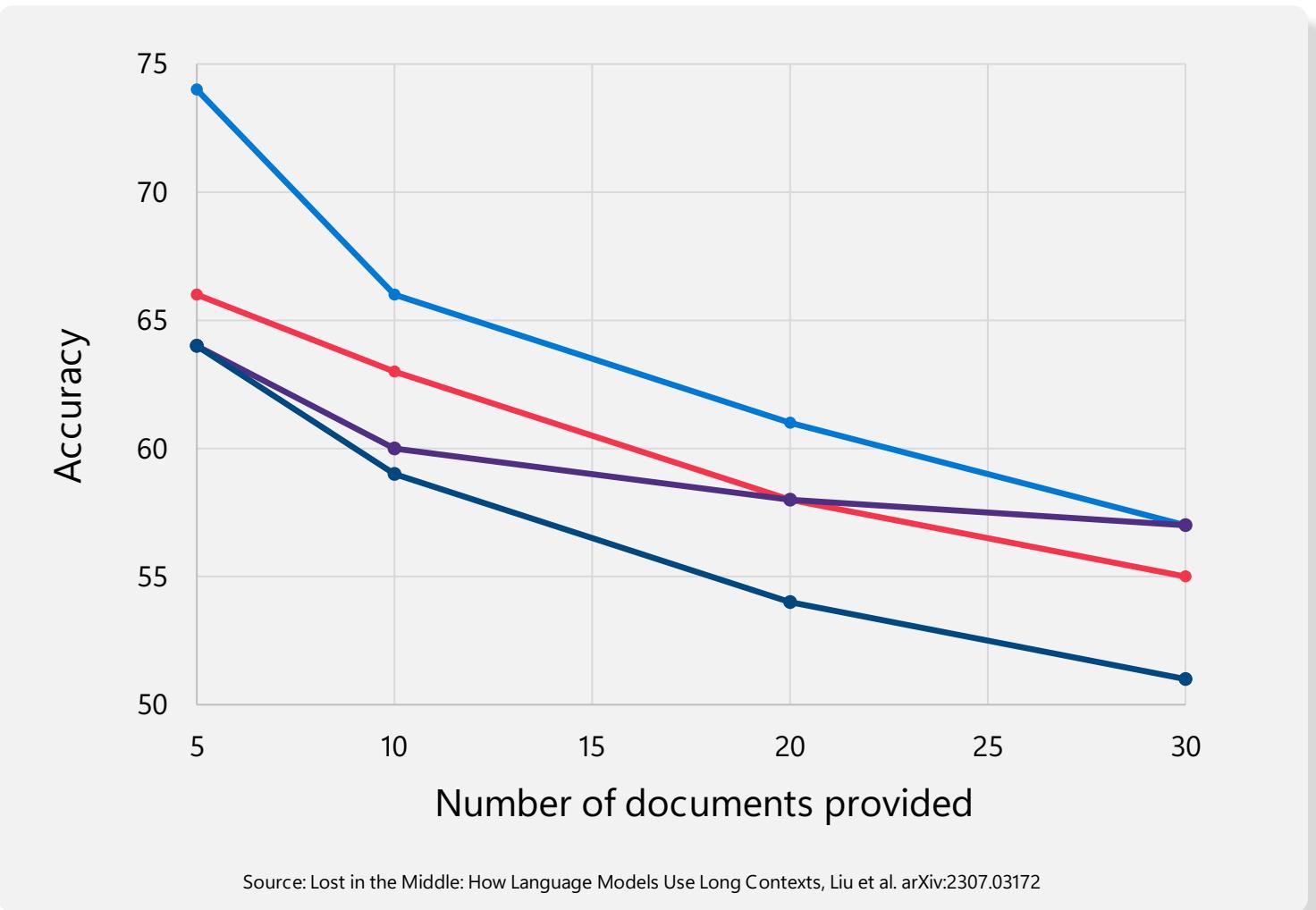
Keyword search won't match:

"Northwind Standard only offers coverage for vision exams and glasses."

"Northwind Health Plus offers coverage for vision exams, glasses, and contact lenses, as well as dental exams, cleanings, and fillings."

Your retrieval strategy matters

More information ≠
better results



Retrieval strategies



Keyword search

- For exact, plain text matches
- “Vocabulary gap” in Q&A systems like Copilot



Vector search

- For conceptual similarity, or underlying meaning
- Weak performance on exact matches (like a product ID or code)



Hybrid search

- Best of both vectors and keywords
- Brings more accurate responses across various scenarios



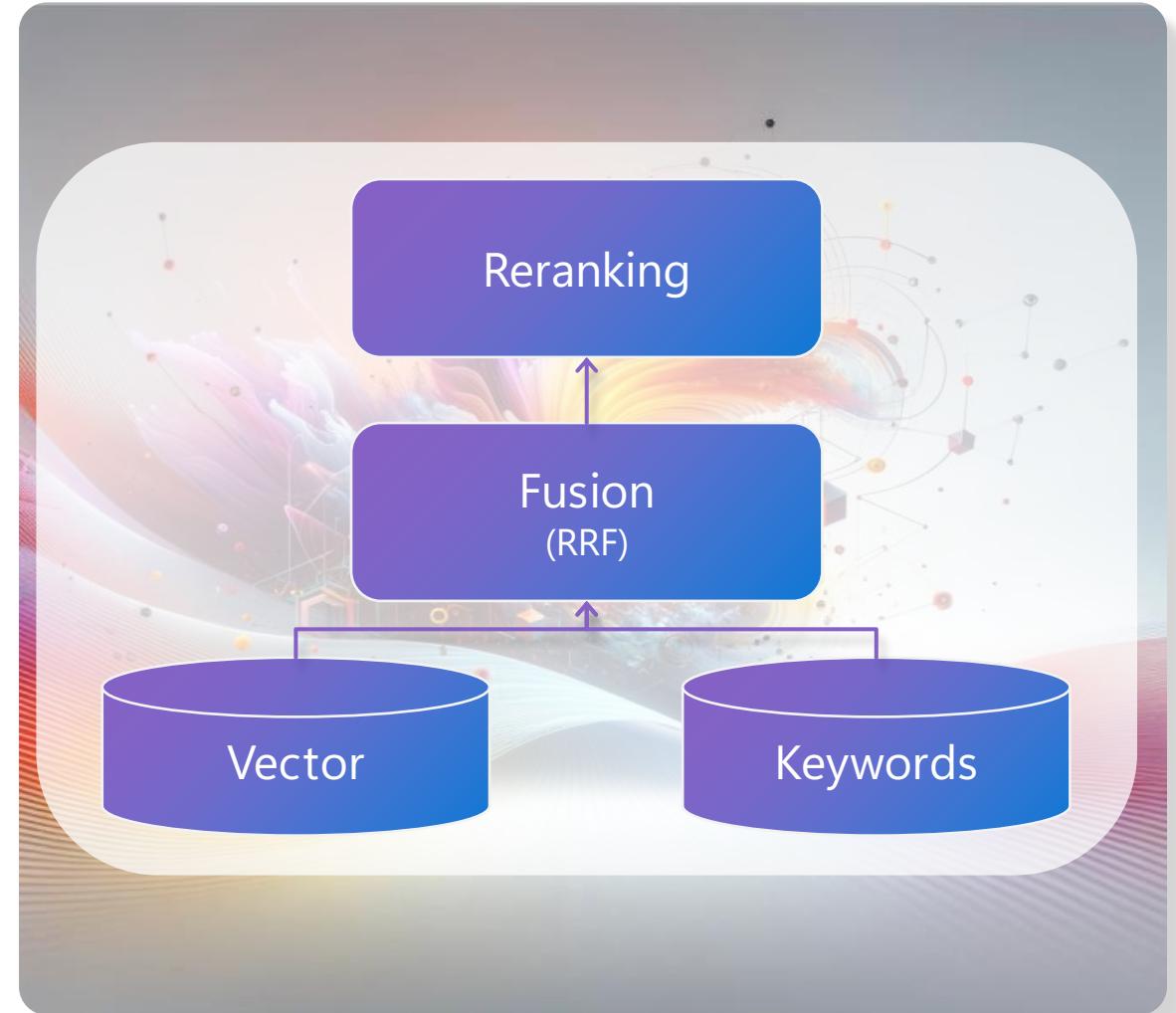
Search re-ranking

- Scores and ranks all retrieved documents by relevance
- Reranking runs after performing search strategy (can't retrieve information)

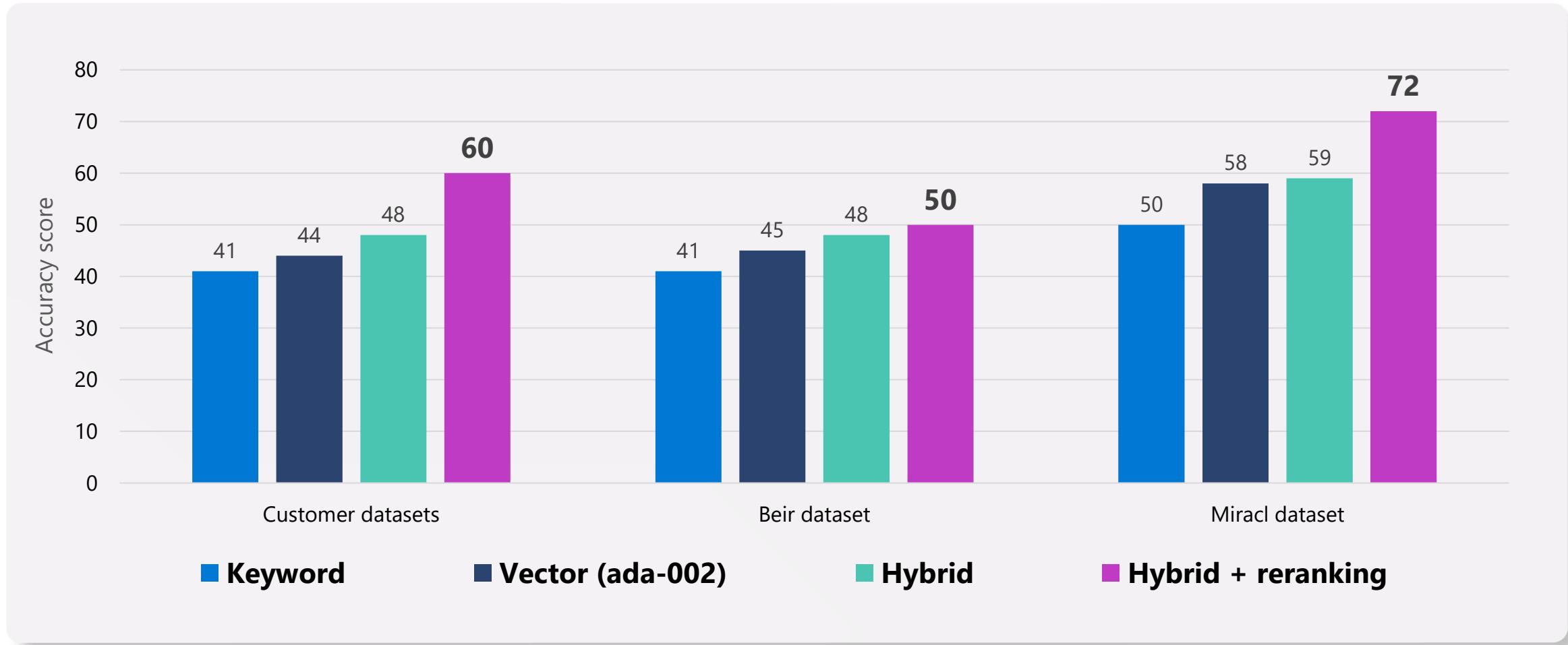
Complete search stacks perform better

Hybrid search mode uses the Reciprocal Rank Fusion algorithm to combine the strength of Azure AI Search's feature-rich vector and full text search modes to perform retrieval across 56+ supported languages.

L2 ranking re-ranks results from the retrieval step and assigns a normalized rank score to help predict which documents are most relevant to the query.



Retrieval relevance across methods



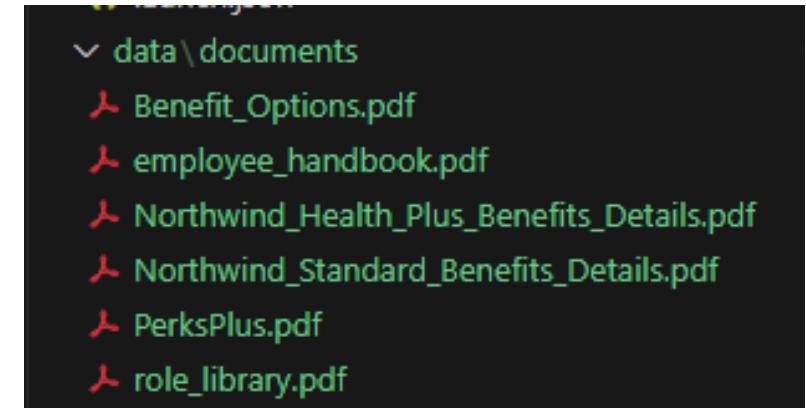
Retrieval comparison using Azure AI Search in various retrieval modes on customer and academic benchmarks
Source: [Outperforming vector search with hybrid + reranking](#)

Impact of query types on relevance

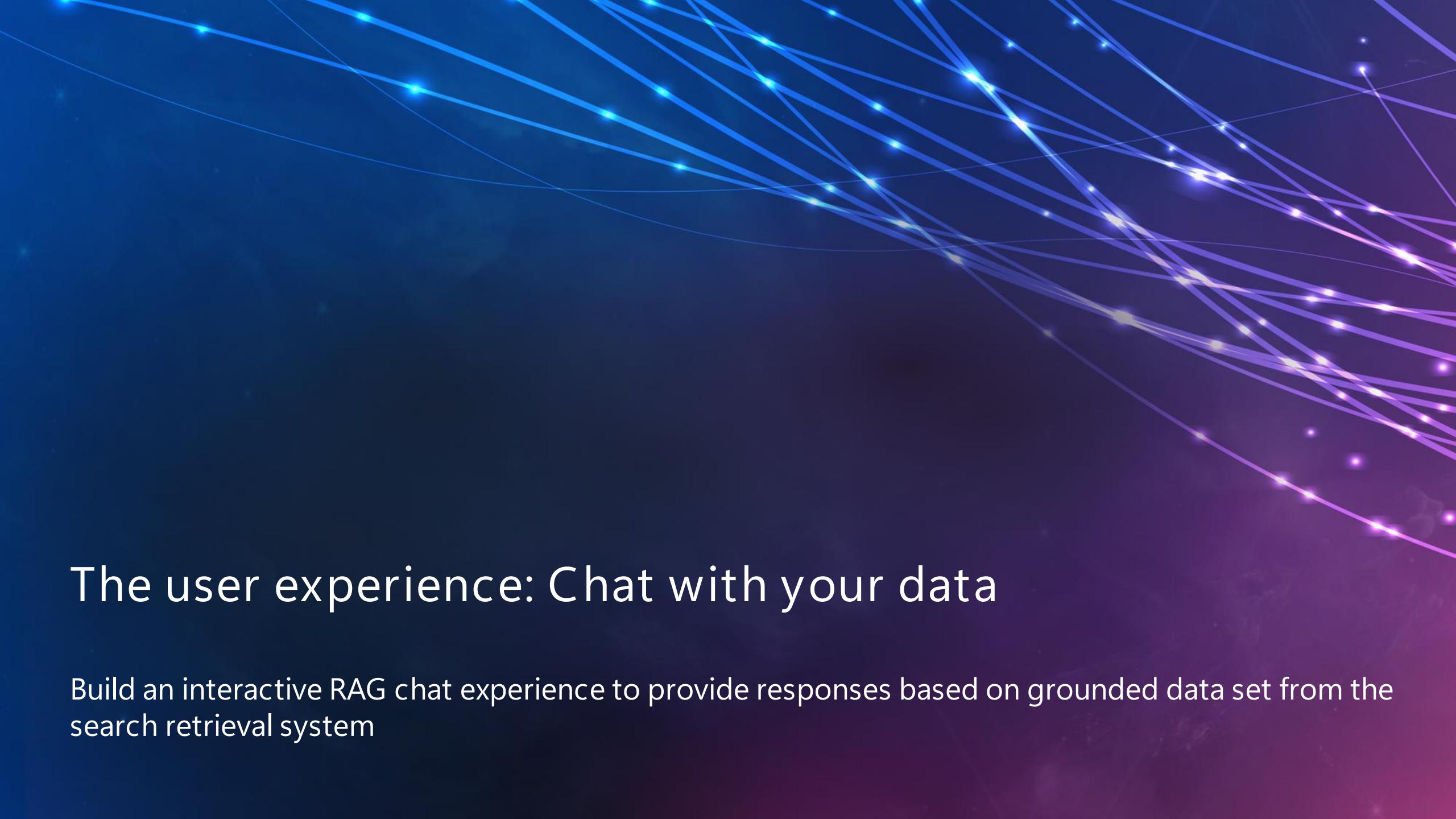
| Query type | Keyword [NDCG@3] | Vector [NDCG@3] | Hybrid [NDCG@3] | Hybrid + Semantic ranker [NDCG@3] |
|----------------------------|---------------------|--------------------|--------------------|---|
| Concept seeking queries | 39 | 45.8 | 46.3 | 59.6 |
| Fact seeking queries | 37.8 | 49 | 49.1 | 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](#)

Data Set and index



| Field name | Type | Retrievable | Filterable | Sortable | Facetable | Searchable | Analyzer | Dimensions |
|------------|------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------|------------|
| | | <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input checked="" type="checkbox"/> | | |
| parent_id | String | <input checked="" type="checkbox"/> | Standa... ▾ | |
| title | String | <input checked="" type="checkbox"/> | Standa... ▾ | |
| 🔑 chunk_id | String | <input checked="" type="checkbox"/> | Keywo... ▾ | |
| chunk | String | <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input checked="" type="checkbox"/> | Standa... ▾ | |
| ⚡ vector | SingleCollection | <input checked="" type="checkbox"/> | | | | <input checked="" type="checkbox"/> | | 1536 |



The user experience: Chat with your data

Build an interactive RAG chat experience to provide responses based on grounded data set from the search retrieval system

UI Frameworks

Gradio and streamlit both are python-based UI framework for developing UI for your data science projects

Use them to create a web-based application that show cases your results, these can be used to build applications for:

- Interacting with your developed machine learning model
- Present analysis of datasets
- Create chat UI
- <https://www.gradio.app/guides/quickstart>
- <https://docs.streamlit.io/get-started/installation>
 - Run: `python -m streamlit hello`

Sample Apps

Chat app using gradio:

`chatapp_gradio.ipynb`

ChatGPT-like using streamlit:

`python -m streamlit run .\chatapp_streamlit.py`

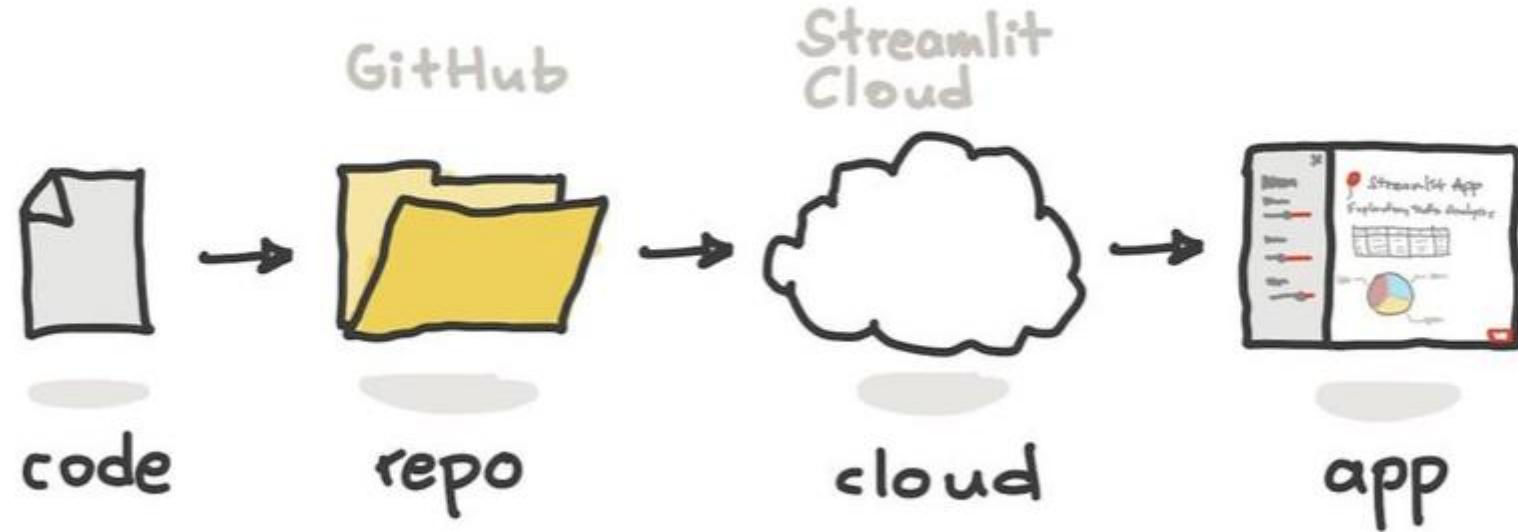
RAG chat using gradio:

`python chatrag_gradio.py`

RAG chat using streamlit:

`python -m streamlit run .\chatrag_streamlit.py`

Sharing your streamlit app



Reference: <https://blog.streamlit.io/host-your-streamlit-app-for-free/>

[← Back](#)

Deploy an app

GitHub URL

[Switch to interactive picker](#)

`https://github.com/mdsiprojects/llm/blob/main/chatrag_streamlit.py`



App URL (optional)

magmdsirag

.streamlit.app

Domain is available

[Advanced settings](#)

Deploy!





mdsiprojects' apps

llm · main · chatrag_streamlit.py

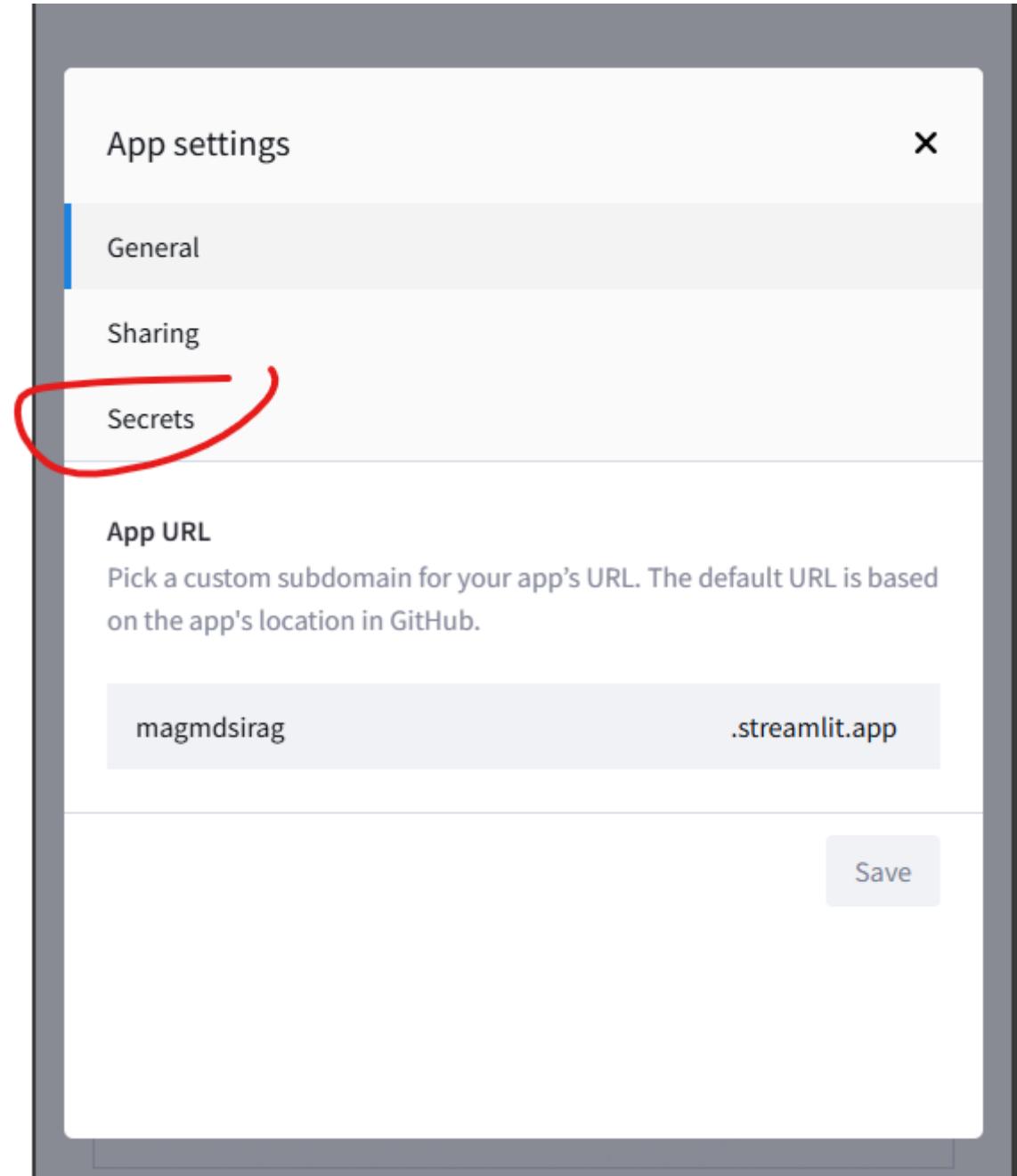
⋮

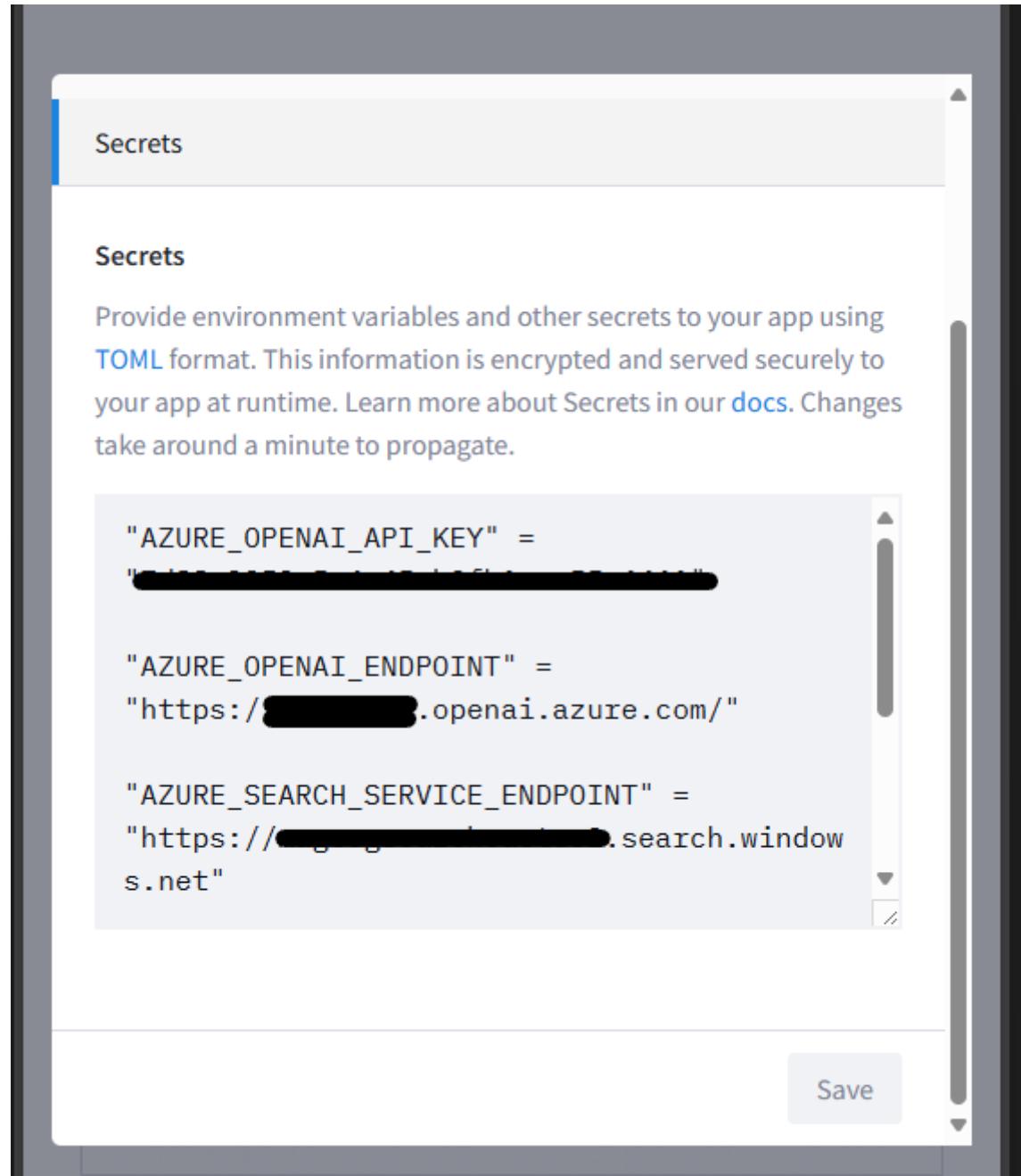
Get started from a template

View all templates →

GDP over time

- Edit with Codespaces
- Reboot
- Delete
- Analytics
- Settings





Copy your secrets from
the .env file
In a TOML format:

"key" = "value"

** key and value must be
wrapped in double
qoutes

The background features a dark blue-to-red gradient with a network of glowing lines and dots. The lines are primarily blue and purple, with some white highlights, forming a complex web that radiates from the top left towards the bottom right.

Thank you