



# 探索人工智能的未来

## Azure OpenAI 最新进展和应用场景

王芷  
微软生态伙伴事业部  
创新技术架构师



# ChatGPT/GPT-4 火爆出圈

×

...

## GPT-4通过律师资格考试：排名高于90%人类考生

原创 AIGC开放社区 AIGC开放社区

2023-03-21 07:41 发表于河北

收录于合集

#ChatGPT 197 #AIGC 209 #生成式AI 158

#大语言模型 96

专注AIGC领域的专业社区，关注GPT-3、百度文心一言等大语言模型（LLM）的发展和落地，以及国内LLM的发展和落地研究，欢迎关注！



AIGC开放社区

专注AIGC（生成式人工智能）领域的... >  
98篇原创内容

公众号

美国伊利诺伊理工大学-芝加哥肯特法学院在官网发布了消息，GPT-4通过了统一律师资格考试。该考试一共7项，GPT-4在民事诉讼、合同法、刑法、物权法、证据法5个学科考试中，得分率高于人类考生。

×

AI能力站 >

...

## 十大可能被取代的职业

媒体网站insider在与专家交谈后和研究后，得出10个工作被人工智能取代的风险最高



技术岗：程序员、工程师、数据分析师



媒体岗：广告、内容创作、写作、新闻



法律岗：法律或律师助理



市场研究分析师



教师



金融岗：金融分析师、个人财务顾问



交易员



平面设计师



会计师



客服人员

×

...

## 重磅！网梯睿学AI学习助手：基于ChatGPT的智能化学习新体验

原创 小梯 网梯科技 2023-03-21 18:48

发表于北京

当下，AI技术越来越成熟，其在各个领域的应用也日益广泛。而在这其中，自然语言处理技术无疑是应用最为广泛的，ChatGPT作为其中的佼佼者，更是备受关注和追捧。作为一种语言生成模型，ChatGPT凭借其强大的生成能力和多样的应用场景，迅速成为了广受欢迎的AI应用之一。

2023年3月，网梯睿学通过Microsoft Azure OpenAI直接引入ChatGPT的基础能力，重磅推出【睿学AI学习助手】，为广大学子提供全新的学习方式和学术支持！通过将网梯公司的专业技术和ChatGPT的API相结合，学习助手不仅能够解决用户的问题，还能够根据用户的需求提供个性化的服务和支持，帮助用户更快地掌握知

×

...

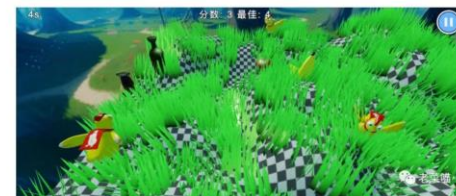
## ChatGPT做了一个二次元游戏！

原创 老菜喵 老菜喵 2023-04-07 20:16

发表于广东

### 1.0 游戏策划设计

继上次借助ChatGPT做了一个3D小游戏后，很多朋友问我，AI可以做大型项目么？还是仅限于简单的小游戏。



\*AI生成的3D小游戏

所以二喵准备接着用AI设计一款中型体量的卡牌游戏，发布到微信小游戏和海外平台。

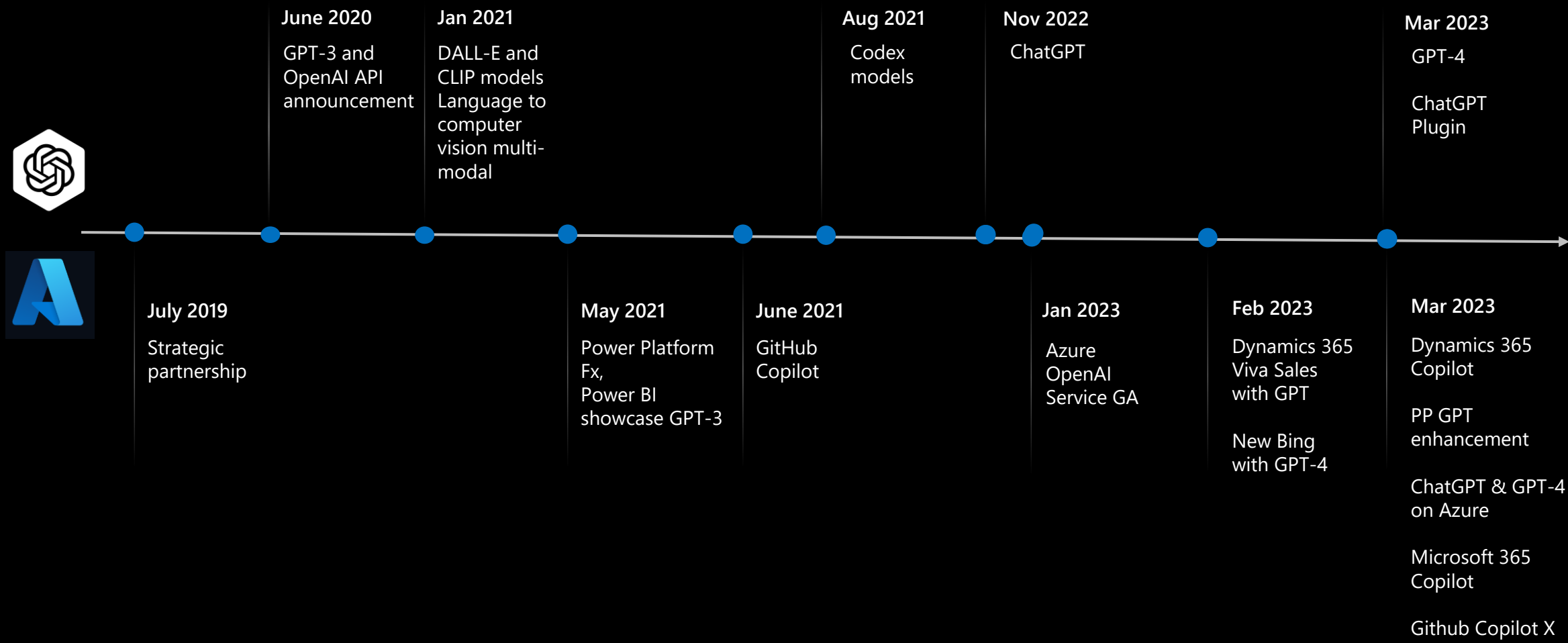
从以下几点着手，来探索AI能力的上限。

1. 策划能力 AI是否能提供完整且有趣的游戏策划案？

2. 美术能力 AI是否能提供游戏的美术素材？

# OpenAI 与 Microsoft 战略合作

不仅仅是100亿这么简单



# Copilot in Word

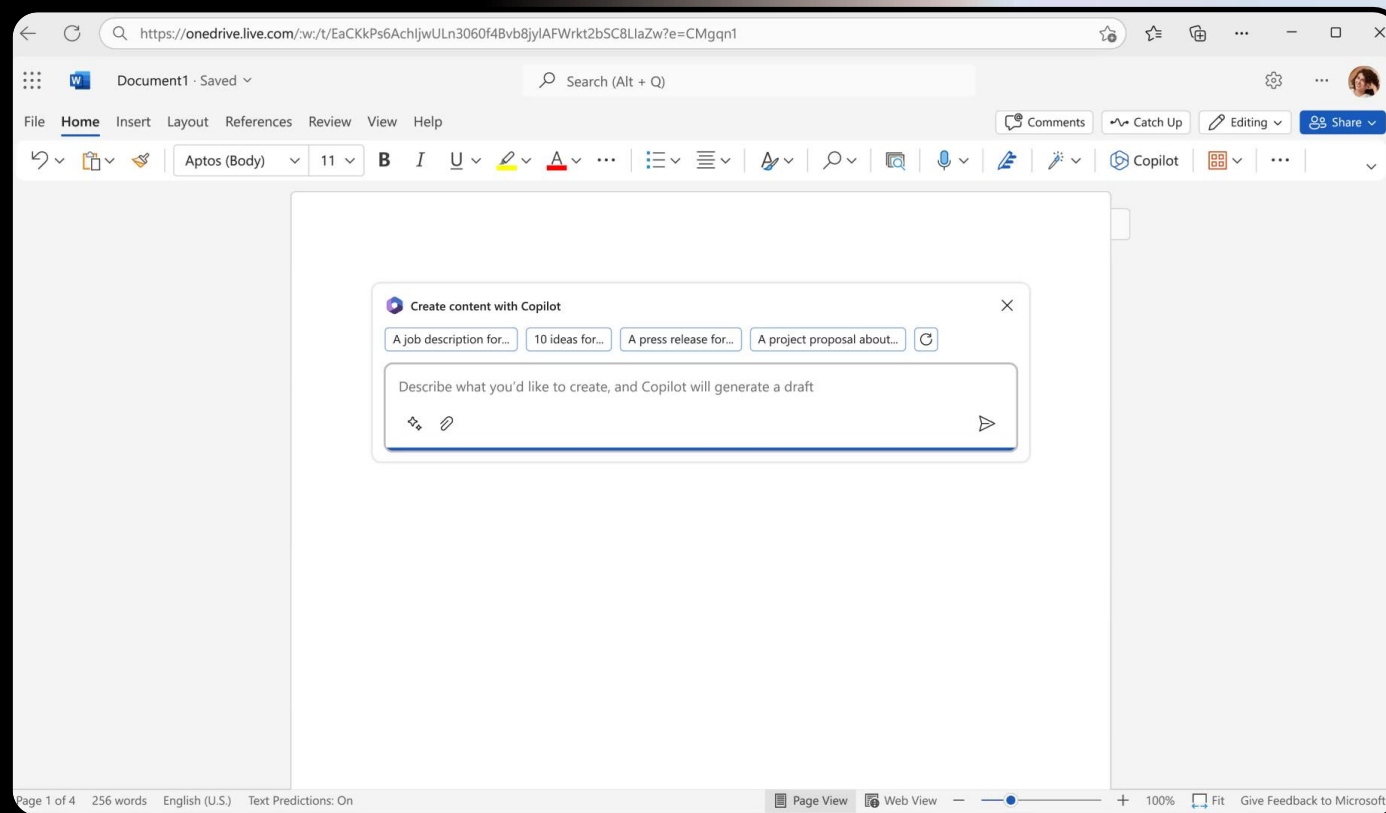
Copilot写作、编辑、总结、创造

您可以询问 Copilot:

根据 [Document] 和 [Excel] 中的数据起草一份两页的项目提案

使第三段更简洁，将文档的语气更改为更随意

根据此粗略大纲创建一页草稿。



# Copilot in PowerPoint

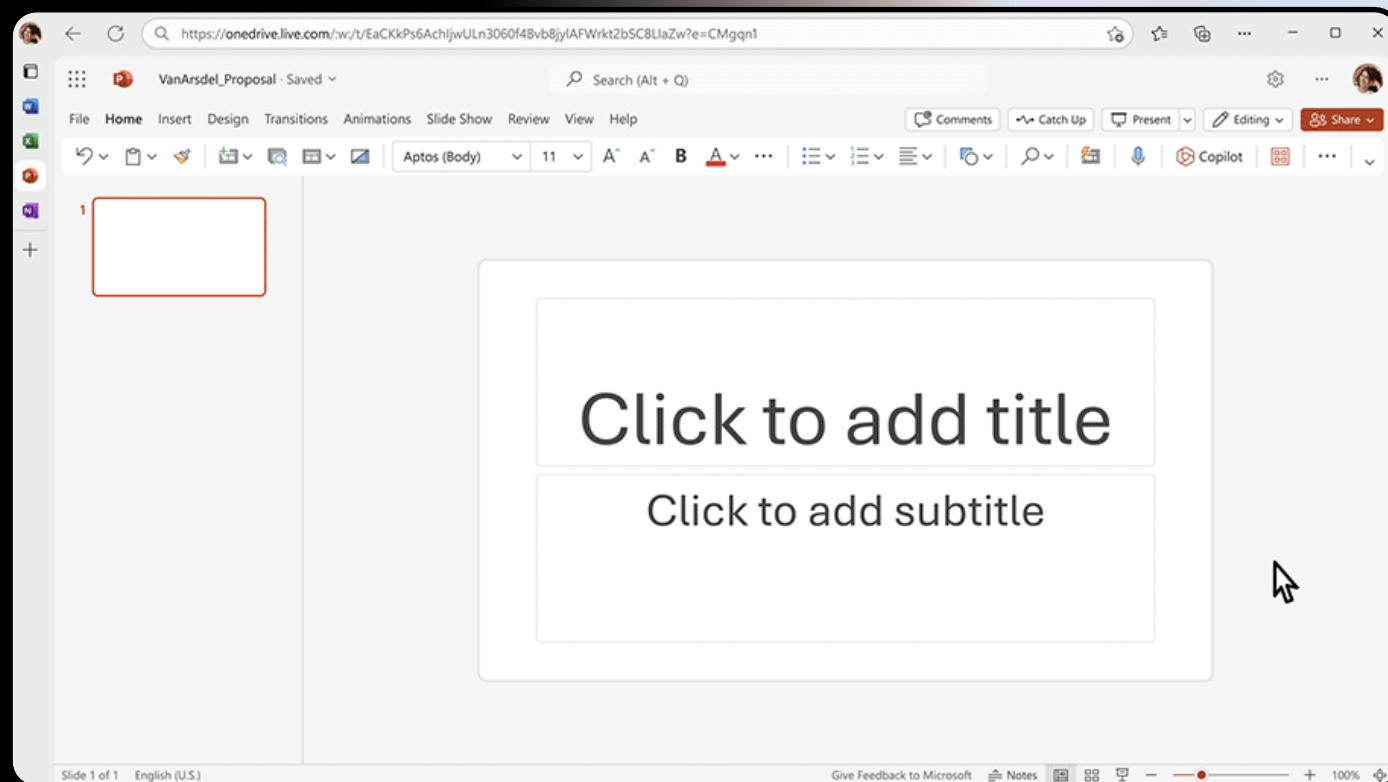
Copilot帮助您实现从idea到PPT的秒级转换

您可以询问 Copilot:

创建五张幻灯片的演示文稿基于给定的 Word 文档，并包含相关的照片

将此演示文稿合并为三张幻灯片摘要

将这三个项目符号重新格式化为三列，每列都有一张图片。



# Copilot in Excel

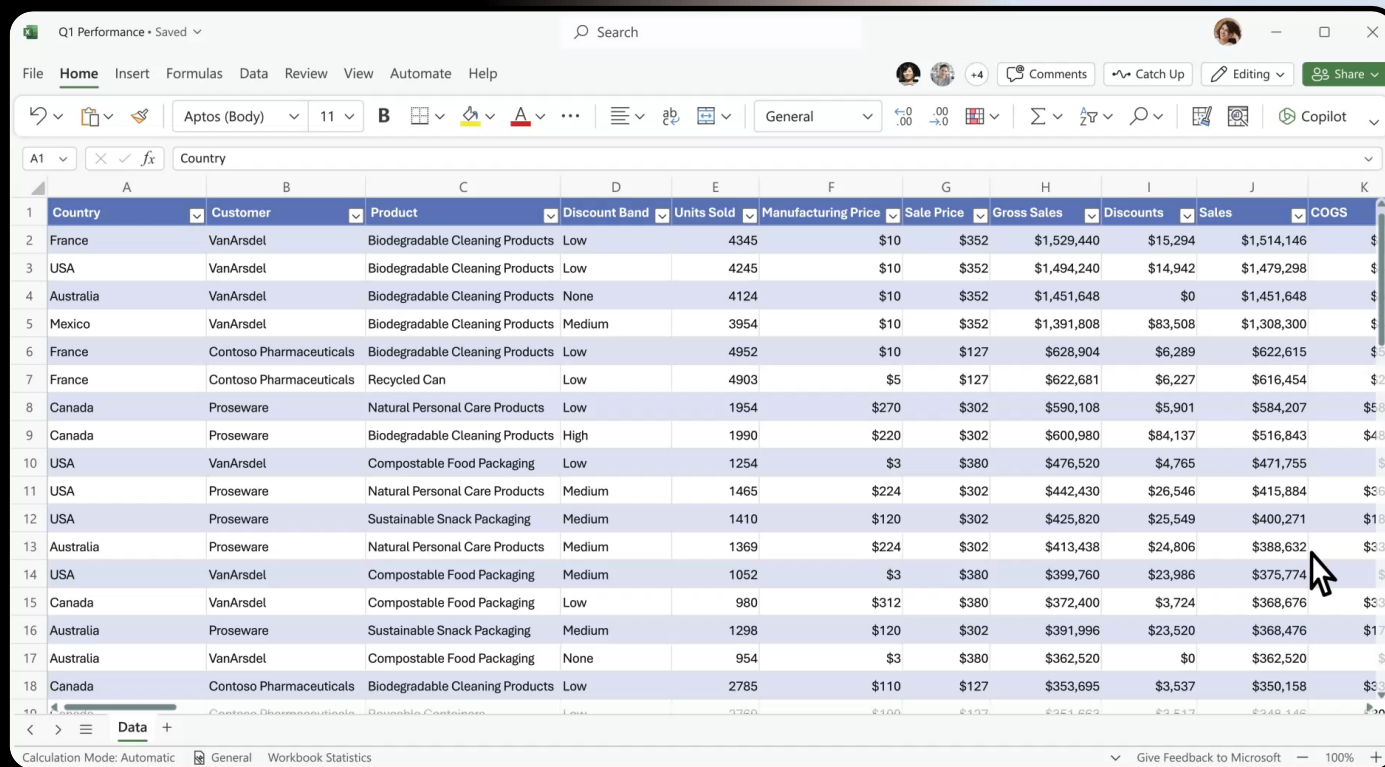
Copilot 与您并肩工作，帮助分析和探索您的数据

您可以询问 Copilot:

按类型和渠道细分销售，插入表格

预测 [variable] 的影响并生成图表以帮助可视化

模拟 [variable] 增长率的变化将如何影响我的毛利率



	A	B	C	D	E	F	G	H	I	J	K
1	Country	Customer	Product	Discount Band	Units Sold	Manufacturing Price	Sale Price	Gross Sales	Discounts	Sales	COGS
2	France	VanArsdel	Biodegradable Cleaning Products	Low	4345	\$10	\$352	\$1,529,440	\$15,294	\$1,514,146	\$
3	USA	VanArsdel	Biodegradable Cleaning Products	Low	4245	\$10	\$352	\$1,494,240	\$14,942	\$1,479,298	\$
4	Australia	VanArsdel	Biodegradable Cleaning Products	None	4124	\$10	\$352	\$1,451,648	\$0	\$1,451,648	\$
5	Mexico	VanArsdel	Biodegradable Cleaning Products	Medium	3954	\$10	\$352	\$1,391,808	\$83,508	\$1,308,300	\$
6	France	Contoso Pharmaceuticals	Biodegradable Cleaning Products	Low	4952	\$10	\$127	\$628,904	\$6,289	\$622,615	\$
7	France	Contoso Pharmaceuticals	Recycled Can	Low	4903	\$5	\$127	\$622,681	\$6,227	\$616,454	\$2
8	Canada	Proseware	Natural Personal Care Products	Low	1954	\$270	\$302	\$590,108	\$5,901	\$584,207	\$58
9	Canada	Proseware	Biodegradable Cleaning Products	High	1990	\$220	\$302	\$600,980	\$84,137	\$516,843	\$48
10	USA	VanArsdel	Compostable Food Packaging	Low	1254	\$3	\$380	\$476,520	\$4,765	\$471,755	\$
11	USA	Proseware	Natural Personal Care Products	Medium	1465	\$224	\$302	\$442,430	\$26,546	\$415,884	\$36
12	USA	Proseware	Sustainable Snack Packaging	Medium	1410	\$120	\$302	\$425,820	\$25,549	\$400,271	\$18
13	Australia	Proseware	Natural Personal Care Products	Medium	1369	\$224	\$302	\$413,438	\$24,806	\$388,632	\$33
14	USA	VanArsdel	Compostable Food Packaging	Medium	1052	\$3	\$380	\$399,760	\$23,986	\$375,774	\$
15	Canada	VanArsdel	Compostable Food Packaging	Low	980	\$312	\$380	\$372,400	\$3,724	\$368,676	\$33
16	Australia	Proseware	Sustainable Snack Packaging	Medium	1298	\$120	\$302	\$391,996	\$23,520	\$368,476	\$17
17	Australia	VanArsdel	Compostable Food Packaging	None	954	\$3	\$380	\$362,520	\$0	\$362,520	\$
18	Canada	Contoso Pharmaceuticals	Biodegradable Cleaning Products	Low	2785	\$110	\$127	\$353,695	\$3,537	\$350,158	\$3

# 企业级AI能力输出

## Applications

 Microsoft 365

 Microsoft Dynamics 365

Partner Solutions

## Application Platform

AI Builder



Power BI



Power Apps



Power Automate



Power Virtual Agents

## Scenario-Based Services

Applied AI Services



Bot Service



Cognitive Search



Form Recognizer



Video Indexer



Metrics Advisor



Immersive Reader

## Customizable AI Models

Cognitive Services



Vision



Speech



Language



Decision



OpenAI Service

## ML Platform



Azure Machine Learning



Business Users



Developers & Data Scientists

# Azure OpenAI

为企业安全保驾护航

微软不会使用客户数据来训练来训练全世界普遍使用的基础模型

您的数据受到最全面的企业合规性和安全控制措施的保护，确保不会被其他用户访问

训练和调用的数据将被加密存储于您的资源所在区域



# Azure OpenAI Service

GPT-3.5

GPT-4.0

DALL·E

ChatGPT



部署在 Azure 订阅中，由你保护，仅由你访问，并绑定到数据集和应用程序



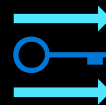
大型预训练 AI 模型，用于解锁新场景



使用数据和超参数微调的自定义 AI 模型



内置负责任的 AI，用于检测和减少有害使用



通过基于角色的访问控制（RBAC）和专用网络实现企业级安全性

# Azure OpenAI 服务与Open AI API的区别

安全性和可靠性 Vnets   RBAC   Auth   合规   数据安全   区域可用性   监测   SLA	
计费 专用资源以及基于事务的资源	
推理   微调	
安全系统和控制 敏感词分类器   允许/阻止列表	
GPT-3 模型 (语言): Ada, Babbage, Curie, Davinci	未来模型

Azure Open AI 服务	OpenAI APIs
✓	●
✓	●
✓	✓
✓	●
✓	✓

# API usage/rate limits

限制名称	限制值
每个区域的 OpenAI 资源	2
每个模型每分钟请求数*	Davinci 模型（002 及更高版本）：120 ChatGPT 模型（预览版）：300 GPT-4 模型（预览版）：12 所有其他模型：300
每个模型每分钟的标记数*	Davinci 模型（002 及更高版本）： 40,000 ChatGPT 模型：120,000 所有其他模型：120,000
最大微调模型部署*	2
能够将同一模型部署到多个部署	不允许
每个资源的训练作业总数	100
每个资源同时运行训练作业的最大数目	1
排队的最大训练作业数	20

\*all rate limits and availability  
[Learn more.](#)



# Region availability

Regions currently available at launch:

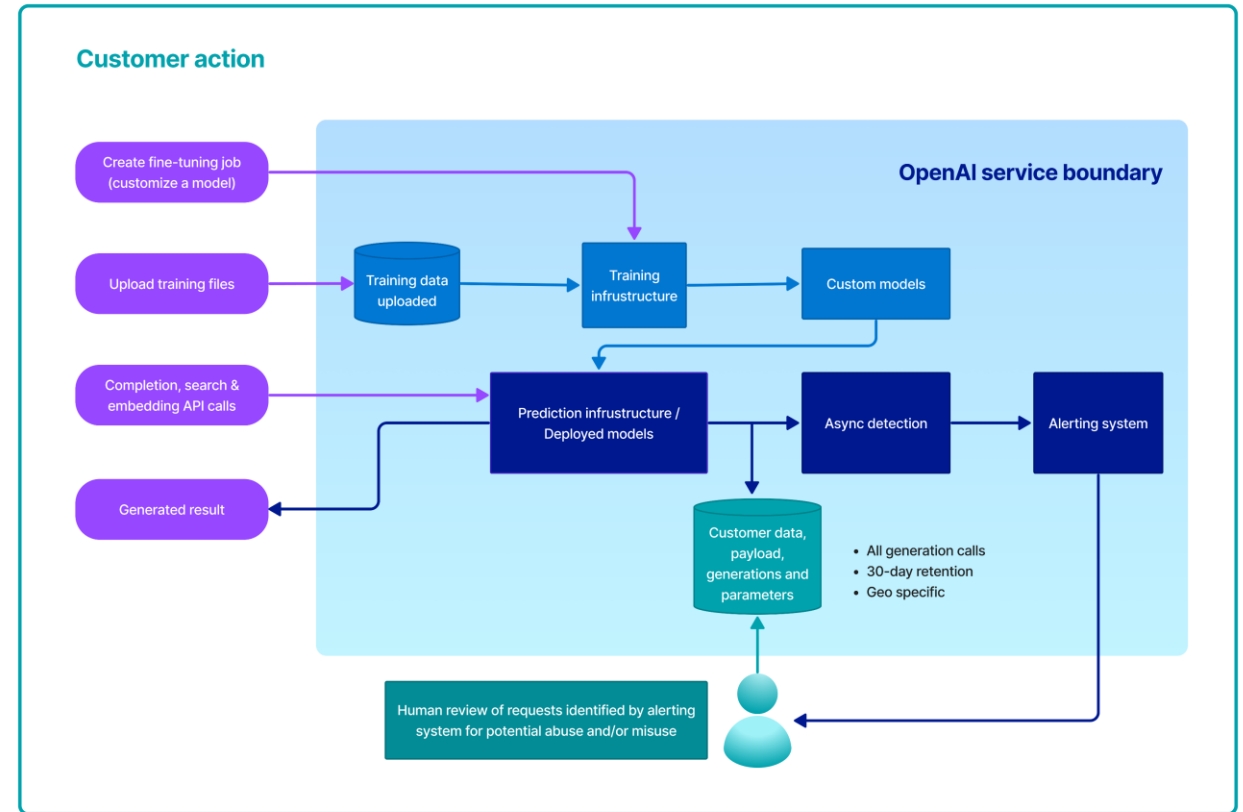
- East US
- South Central US
- West Europe

To see the list of available regions, please visit:

[Product available by region](#)

REA	NORWAY	QATAR	SWEDEN	SWITZERLAND	UNITED ARAB EMIRATES	UNITED KINGDOM	UNITED STATES					
Products	tral	Sweden Central	Switzerland North	UAE North	UK South	UK West	Central US	East US	East US 2	North Central US	South Central US	
<a href="#">Azure Cognitive Services</a>		✓	✓	✓	✓		✓	✓	✓	✓	✓	
<a href="#">Anomaly Detector</a>		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
<a href="#">Computer Vision</a>		✓	✓	✓	✓	⬇	✓	✓	✓	✓	✓	
<a href="#">Content Moderator</a>		✓	✓	✓	✓		✓	✓	✓	✓	✓	
<a href="#">Custom Vision</a>					✓		⬇	✓	✓	✓	✓	
<a href="#">Face API</a>		✓	✓	✓	✓		✓	✓	✓		✓	
<a href="#">Cognitive Service for Language</a>			✓	✓	✓		✓	✓	✓	✓	✓	
<a href="#">Language Understanding (LUIS)</a>			✓	✓	✓		✓	✓	✓	✓	✓	
<a href="#">Personalizer</a>		✓	✓	✓	✓		✓	✓	✓	✓	✓	
<a href="#">QnA Maker</a>			✓	✓	✓	✓	✓	✓	✓	✓	✓	
<a href="#">Speaker recognition</a>												
<a href="#">Speech Services</a>		✓	✓	✓	✓		✓	✓	✓	✓	✓	
<a href="#">Translator</a>												
<a href="#">Azure OpenAI Service</a>								✓			✓	

# How Azure OpenAI processes data



# Working with the ChatGPT model

## Previous GPT-3 models

Previous models were text-in and text-out

(i.e., they accepted a prompt string and returned a completion to append to the prompt).

---

Answer questions from the context below.

Context:

A neutron star is the collapsed core of a massive supergiant star, which had a total mass of between 10 and 25 solar masses, possibly more if the star was especially metal-rich.

Q: What is a neutron star?

A:

## The ChatGPT model

The ChatGPT model is conversation-in and message-out.

(i.e., it expects a prompt string that is formatted in a specific chat-like transcript format and returns a completion that represents a model-written message in the chat)

---

<|im\_start|>system

Assistant is an AI Chatbot designed to answer questions from the context provided below.

Context:

A neutron star is the collapsed core of a massive supergiant star, which had a total mass of between 10 and 25 solar masses, possibly more if the star was especially metal-rich.

<|im\_end|>

<|im\_start|>user

What is a neutron star?

<|im\_end|>

<|im\_start|>assistant



# Understanding the ChatGPT prompt format

## The system message

The system message is included at the beginning of the prompt between the `<|im_start|>system` and `<|im_end|>` tokens.

This message is used to prime the model and you can include a variety of information in the system message including:

- A brief description of the assistant
- The personality of the assistant
- Instructions for the assistant
- Data or information needed for the model

## User and assistant messages

After the system message, you can include a series of messages between the *user* and the *assistant*. Each message should begin with the `<|im_start|>` token followed by the role (*user* or *assistant*) and end with the `<|im_end|>` token.

To trigger a response from the model, the prompt should end with `<|im_start|>assistant` token indicating that it's the assistant's turn to respond.

## Example prompt

`<|im_start|>system`

You are an Xbox customer support agent whose primary goal is to help users with issues they are experiencing with their Xbox devices. You are friendly and concise. You only provide factual answers to queries, and do not provide answers that are not related to Xbox.

`<|im_end|>`

`<|im_start|>user`

Why won't my Xbox turn on?

`<|im_end|>`

`<|im_start|>assistant`

There could be a few reasons why your Xbox isn't turning on....

`<|im_end|>`

`<|im_start|>user`

I confirmed the power cord is plugged in but it's still not working

`<|im_end|>`

`<|im_start|>assistant`

# ChatGPT benefits



## Conversational

The conversational nature of the model makes it easier to interact with so you can more easily get the most out of the model.

## Multi-turn

The conversational nature of ChatGPT makes it easy to follow up on the model's response. This gives users an easy mechanism to ask suggest edits, ask for clarification, etc.

## Creative

The ChatGPT model excels at creative tasks like content writing and storytelling.

# ChatGPT limitations



## Hallucinations

While the ChatGPT model has proven to have extensive knowledge, it can still be wrong at times. It's important to understand this limitation and apply mitigations for your scenario.

## Non-conversational tasks

The ChatGPT model was optimized for conversational tasks. This means it might not perform as well on structured tasks like entity extraction, classification, etc. For more structured use cases, we recommend comparing ChatGPT with other models such as *text-davinci-003*.



# Tokens

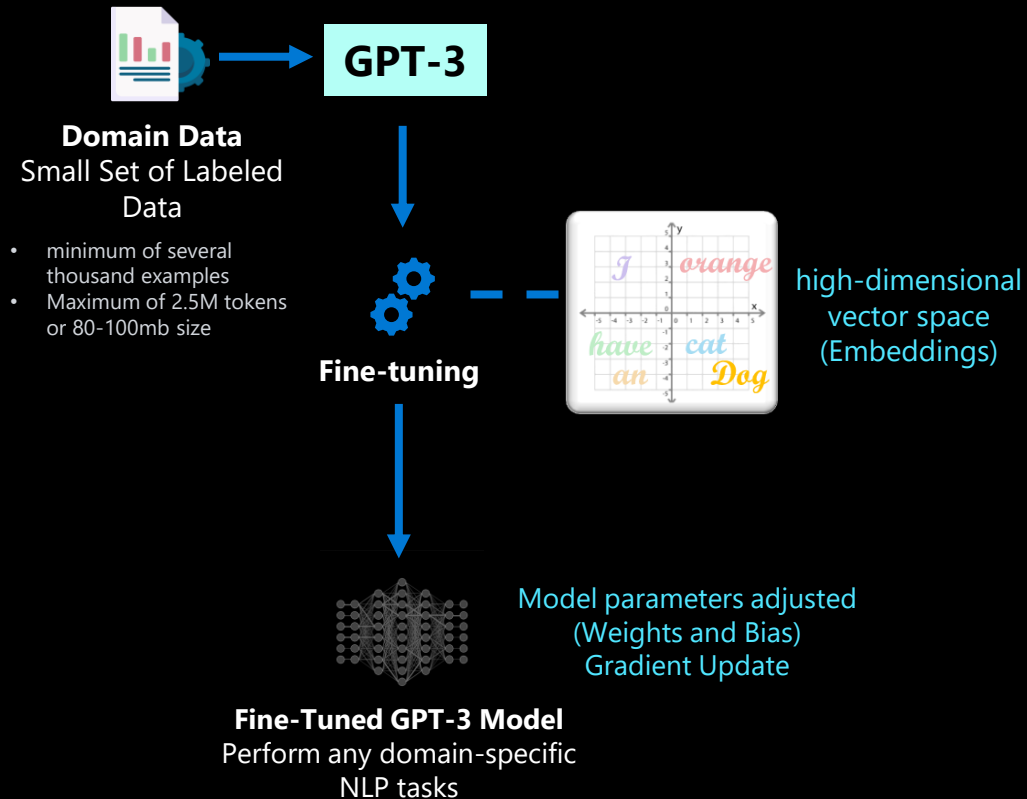
You can think of tokens as pieces of words used for natural language processing. For English text, 1 token is approximately 4 characters or 0.75 words.

---

中文1 token $\sim$ =0.5汉字

# Model Adaptation with specific domain data

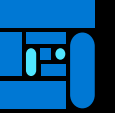
## Fine-Tuning



**Fine-tuning** results is a new model being generated with updated weights and biases.

This is in contrast to **few-shot learning** in which model weights and biases are not updated.

# When Fine-Tuning is needed



If model is making untrue statements ("hallucinations"), then mitigate the hallucinations

Accuracy of results of the model does not meet customer requirements

Fine-tuning lets you get more out of the models available through the API by providing:

- Higher quality results than prompt design
- Ability to train on more examples than can fit in a prompt
- Lower latency requests

Fine-tuning improves over few-shot learning by training on many more examples than can fit in the prompt, letting you achieve better results on a wide number of tasks.



# Best practices of Fine-Tuning



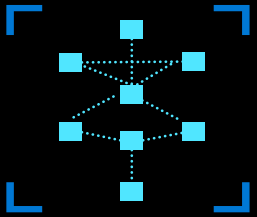
Fine-tuning data set must be in JSON format

A set of training examples that each consist of a single input ("prompt") and its associated output ("completion")

For classification task, the prompt is the problem statement, completion is the target class

For text generation task, the prompt is the instruction/question/request, and completion is the text ground truth

# Embeddings



An embedding is a special format of data representation that can be easily utilized by machine learning models and algorithms.

The embedding is an information dense representation of the semantic meaning of a piece of text.

Each embedding is a vector of floating-point numbers, such that the distance between two embeddings in the vector space is correlated with semantic similarity between two inputs in the original format.

For example, if two texts are similar, then their vector representations should also be similar.

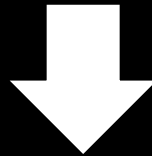
# Embeddings make it possible to map content to a “semantic space”

A neutron star is the collapsed core of a massive supergiant star



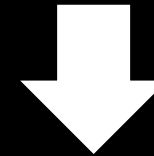
[ 15 34 24 13 ...]

A star shines for most of its active life due to thermonuclear fusion.



[16 22 89 26 ...]

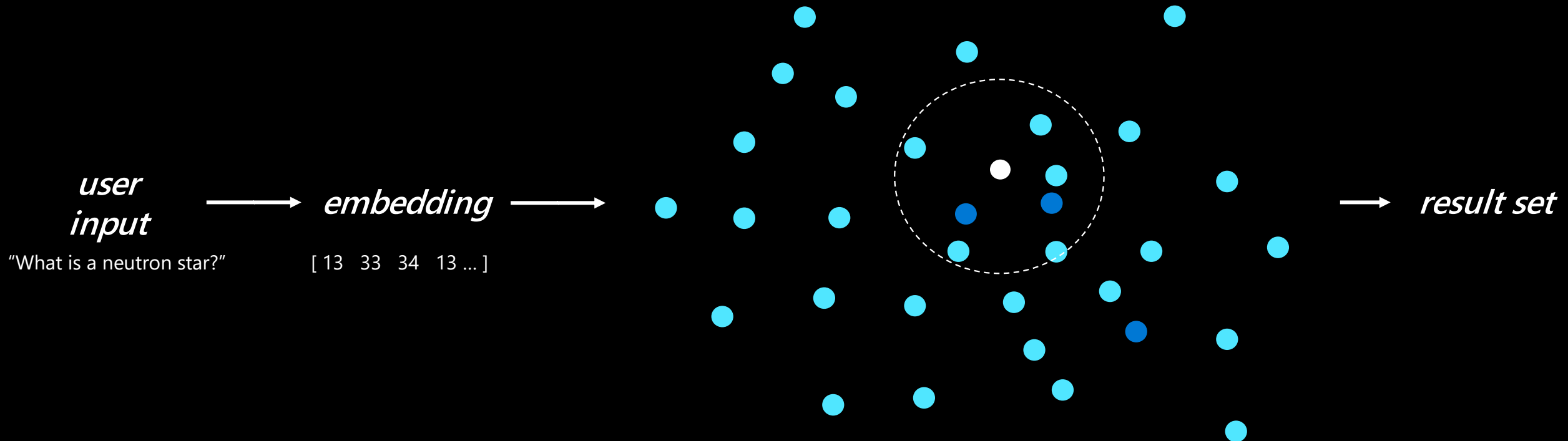
The presence of a black hole can be inferred through its interaction with other matter



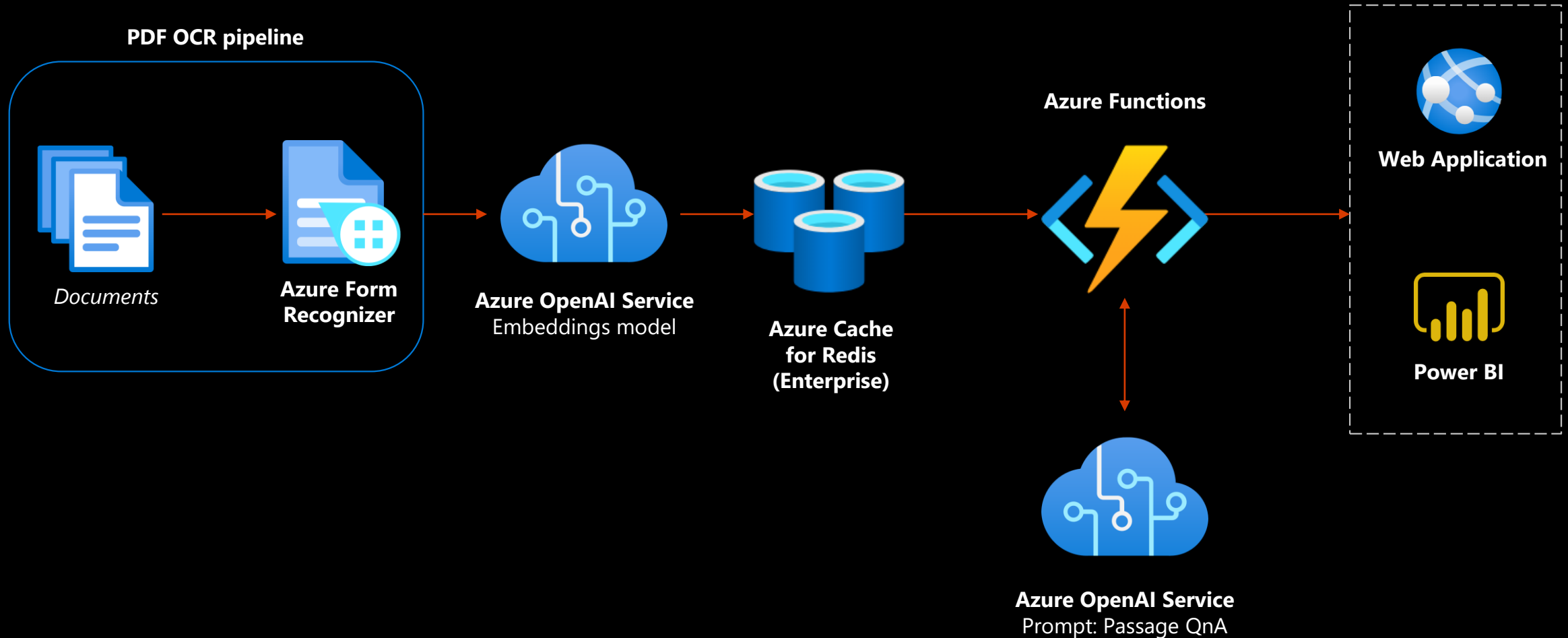
[ 20 13 31 89 ...]

# Similarity Search with embeddings

Once you encode your content as embeddings, you can then get an embedding from the user input and use that to find the most semantically similar content.

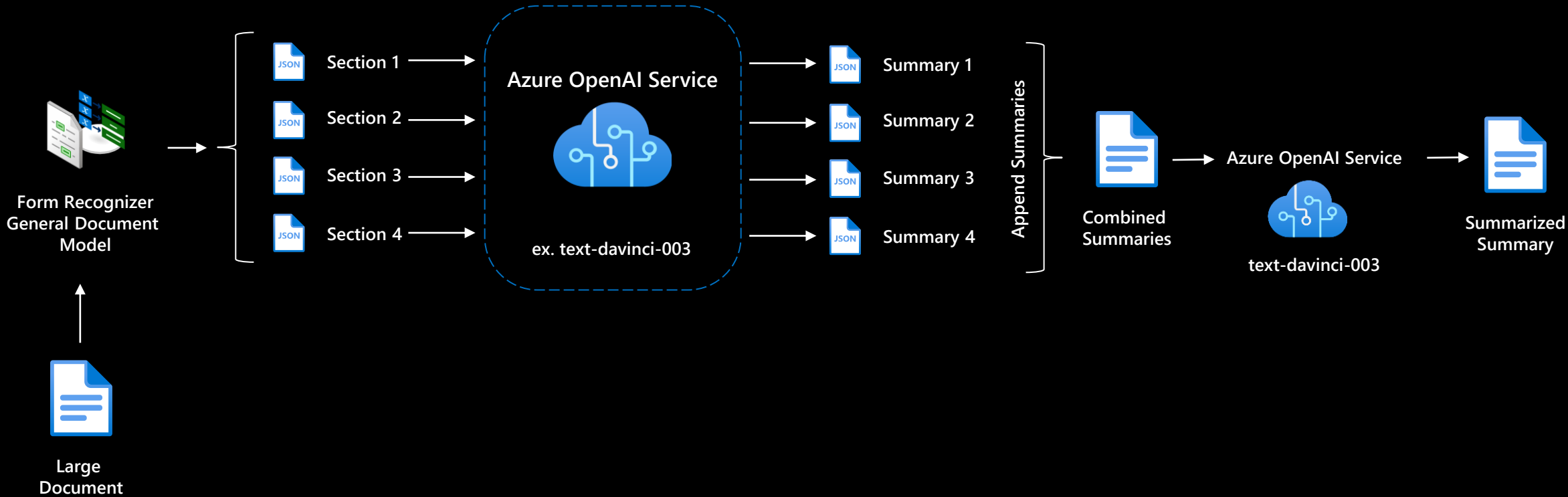


# Architecture design to build the PoC

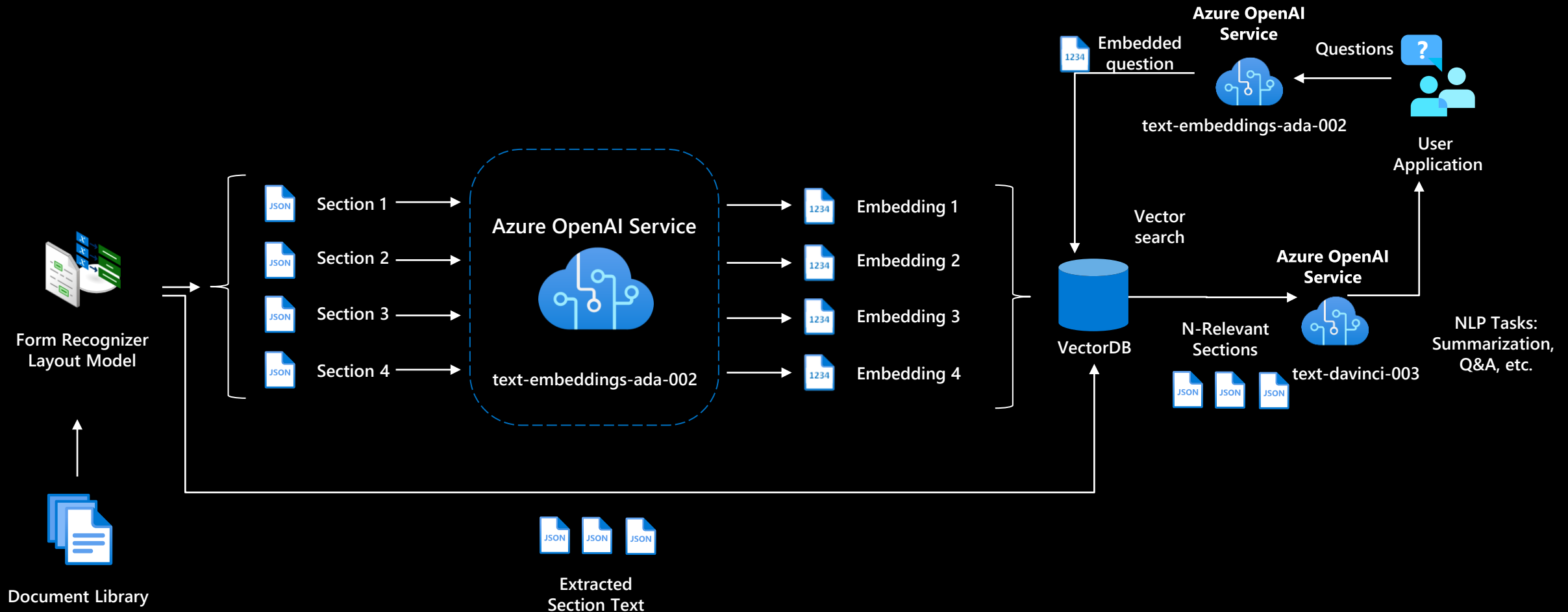




# Large Document Summarization

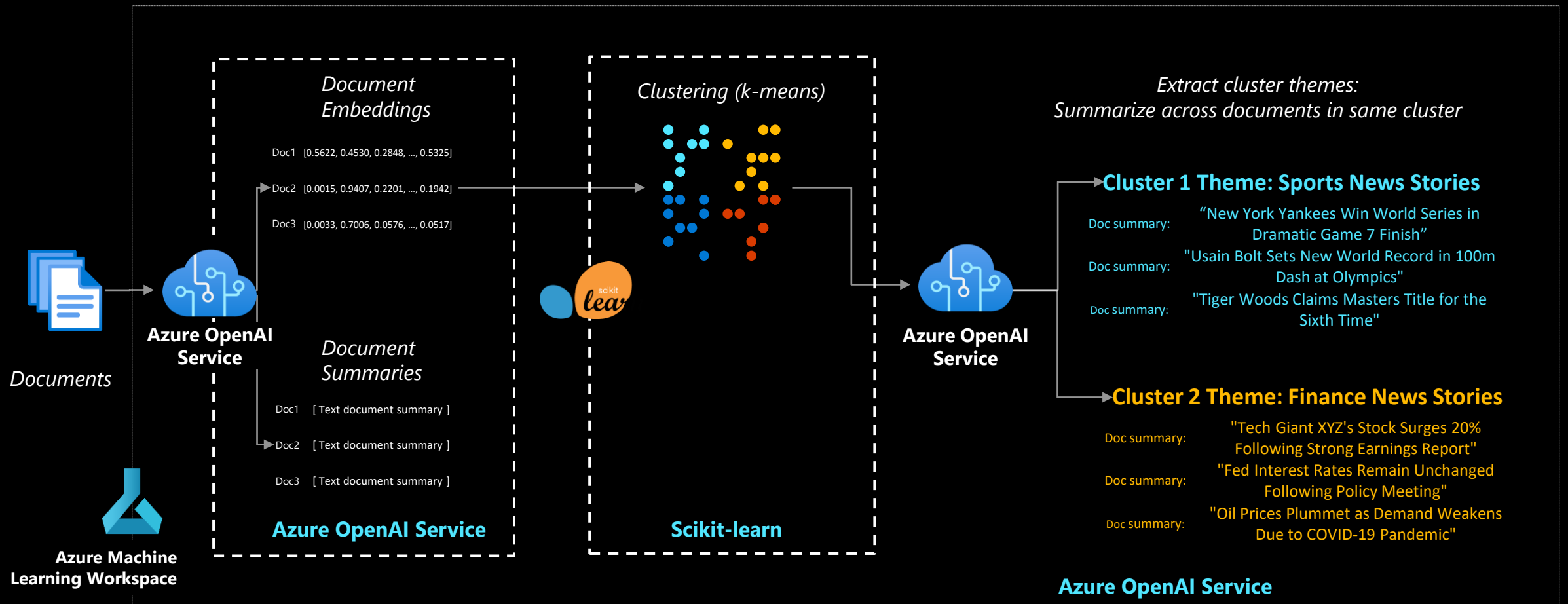


# Q&A with Semantic Answering over Document Library



# Document Clustering, and Cluster Theme Generation

Summarize and extract document similarity embeddings,  
mine for similar clusters, generate cluster themes



# 企业级应用场景



客户服务

基于特定任务或知识的  
问答机器人

泛化的闲聊机器人

Chat your Customers



销售市场

舆情监控  
社交媒体趋势总结

客制化的产品推荐

Chat your Web



内容生成

专业主题文档生成  
(财务报告, 分析师文章)

代码生成和代码注释

Chat your Products



知识管理

企业内部的知识挖掘和  
检索

针对特定知识体系的  
培训资料和问答系统

Chat your Docs



辅助决策

自然语言交互的  
自服务数据查询

决策者的数据顾问

Chat your Data

## 合作伙伴案例

问答式企业知识搜索 (Chat your Web) : 针对特定文章做内容检索、摘要生成、自然语言答复, 做企业内部的New Bing

智能文档处理 (Chat your Doc) : 针对纸质化, 图片等非结构化数据进行文字识别和关键信息抽取, 并通过RPA系统进行自动录入

智能数据Copilot (Chat your Data): 针对企业现有数据提供基于自然语言实现零代码的数据自服务查询