Azure OpenAI

TheTechPlatform
Where Tech Meets.

Azure OpenAI Series (Virtual)

Saturday, 16 Mar, 2024

Unleashing the Power of Artificial Intelligence in the Cloud– Part 2

Aroh Shukla

Regional Microsoft Cloud Architect

MVP Alumni, MCT

TheTechPlatform
Where Tech Meets.

Aroh Shukla

MVP Alumni, MCT

Global Speaker

- Passionate to **learn**.
- Passionate to **share knowledge**.
- Passionate to work on **Microsoft Technologies**

Aroh.Shukla@gmail.com

/arohshukla

@aaroh_bits

TheTechPlatform
Where Tech Meets.

AGENDA SLIDE

Part 1

01

Getting Started

How to get access to Azure OpenAI

02

Models

different models available in Azure OpenAI

03

Use Cases

Real life examples of Azure OpenAI

04

Pricing & Deployment

How Azure OpenAI pricing tiers and deploy ment

05

Work with text models

hands-on tutorial on how to work with text models in Azure OpenAI.

TheTechPlatform
Where Tech Meets.

Azure Open AI

Sessions Roadmap

Part 01

Exploring Azure OpenAI: a rewarding journey into integrating Azure services with OpenAI tech. Learn about ChatGPT, DALL-E2, and more to establish a strong AI foundation and unleash creativity.

Part 2

Gain basic Azure OpenAI insight: access, models, use cases, responsible AI, pricing, deployment, and text model utilization.

Part 3

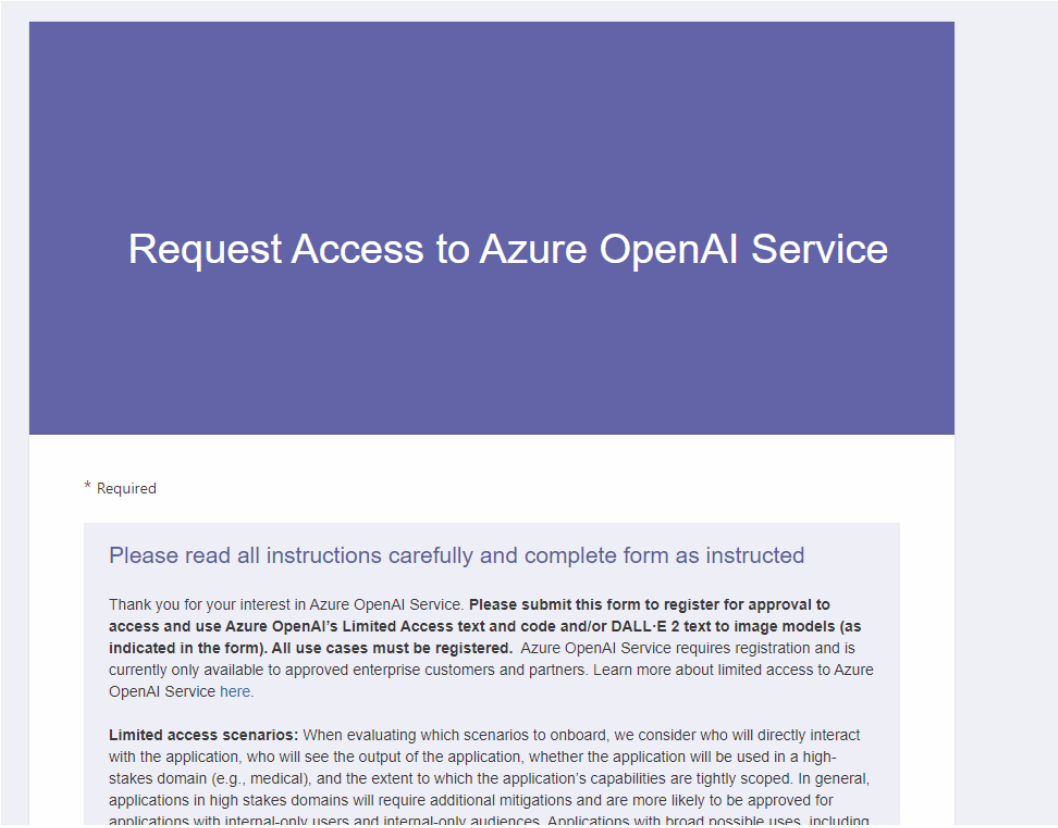
Part 3 enhances Part 2's learning: refine models, ensure security, employ content safety, demo applications. Relevant for those eager to leverage Azure OpenAI for personal AI solutions..

TheTechPlatform
Where Tech Meets.

How to access Azure OpenAI Service?

- 1. Access is granted **upon request.**
- 2. Submission of a **form is required**.
- 3. Microsoft verifies **company information** before providing access to your **Azure Subscription**.

- [Request Access to Azure OpenAI Service URL]



Request Access to Azure OpenAI Service

* Required

Please read all instructions carefully and complete form as instructed

Thank you for your interest in Azure OpenAI Service. **Please submit this form to register for approval to access and use Azure OpenAI's Limited Access text and code and/or DALL·E 2 text to image models (as indicated in the form). All use cases must be registered.** Azure OpenAI Service requires registration and is currently only available to approved enterprise customers and partners. Learn more about limited access to Azure OpenAI Service here.

Limited access scenarios: When evaluating which scenarios to onboard, we consider who will directly interact with the application, who will see the output of the application, whether the application will be used in a high-stakes domain (e.g., medical), and the extent to which the application's capabilities are tightly scoped. In general, applications in high stakes domains will require additional mitigations and are more likely to be approved for applications with internal-only users and internal-only audiences. Applications with broad possible uses, including

TheTechPlatform
Where Tech Meets.

DEMO 1: How to request the Azure OpenAI

Azure OpenAI Service

- Provides **REST API access** to OpenAI language models
- The only cloud offering OpenAI models
- Improved **reliability**
- **Security controls**
- **Great SLA – 99.9%**

TheTechPlatform
Where Tech Meets.

Azure OpenAI Models

- The following **model families** are available in Azure OpenAI service:

GPT-3 (incl. ChatGPT)

GPT-4

Codex

Embedding

Azure OpenAI Models

- Each model family has its **own set of capabilities**
- When calling the Azure OpenAI REST API we need to specify the exact model **we want to use in the model family**
- Using the following format:

```
{capability}-{family}-{version}
```

GPT-3

- A family of models that **can understand** and **generate natural language**
- Each model in GPT-3 has its own tradeoff between **capability** and **performance**
- Models are named in alphabetical order. Goes from the **fastest** to the **most capable**

GPT-3

- GPT-3 models:

`text-ada-001`	The **fastest model.** Good for **parsing text** and **basic classification tasks**
`text-babbage-001`	Can be used for **semantic search** and **simple classification**
`text-curie-001`	Can be used **for translation**, **complex classification, text sentiment, summarization**
`text-davinci-003`	The most capable model. Use for identifying complex intent and summarization
`gpt-35-turbo`	ChatGPT. Conversational model capable of complex interactions in a conversation-in / message-out format. Has its own API

TheTechPlatform
Where Tech Meets.

GPT-4

- Improve over **GPT-3**
- Capable of **solving difficult problems**
- Better accuracy than GPT-3
- Optimized for **chats**

GPT-4

- In order to get access:
  - Have OpenAI service access
  - ~~Apply to join the waiting list~~
- Link is in the resources of this lecture
- We won’t use it in this course

TheTechPlatform
Where Tech Meets.

GPT-4

- GPT-4 models:

`gpt-4`	Supports up to **8,192** input tokens
`gpt-4-32k`	Supports up to **32,768** input tokens

TheTechPlatform
Where Tech Meets.

Codex

- Based on GPT-3
- Specializes in **understanding** and **generating code**
- Trained on **billions of lines of public code from GitHub**

Codex

- Works best in Python ☺
- Supports also:

C#	Ruby
JavaScript	Swift
Go	TypeScript
Perl	SQL
PHP	Shell

Codex

- Codex models:

`code-Cushman-001`	**Fast**, good for simple tasks
`code-davinci-002`	The **most capable model**, can perform any code-related task. Great understanding of code segments

Embeddings

- A special format of data representation
- Used by **machine learning models** and algorithms
- We won’t work with embeddings in these session
- Microsoft Tutorial

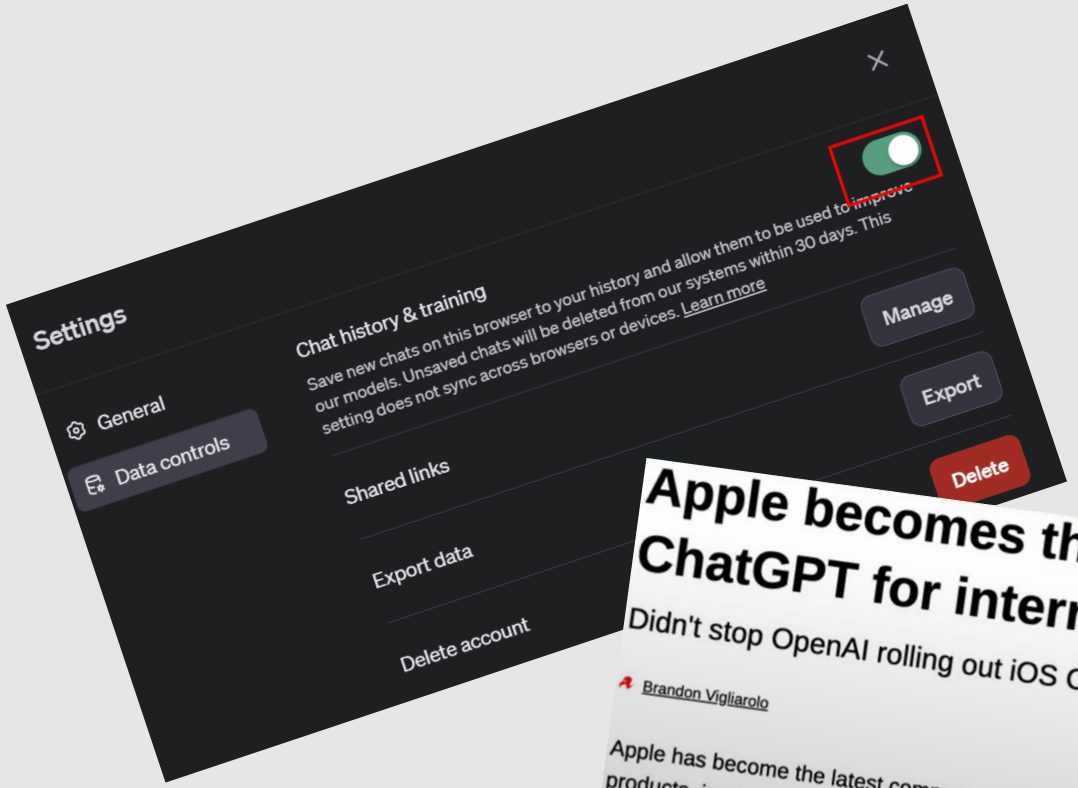
DEMO 2: Azure OpenAI Playground

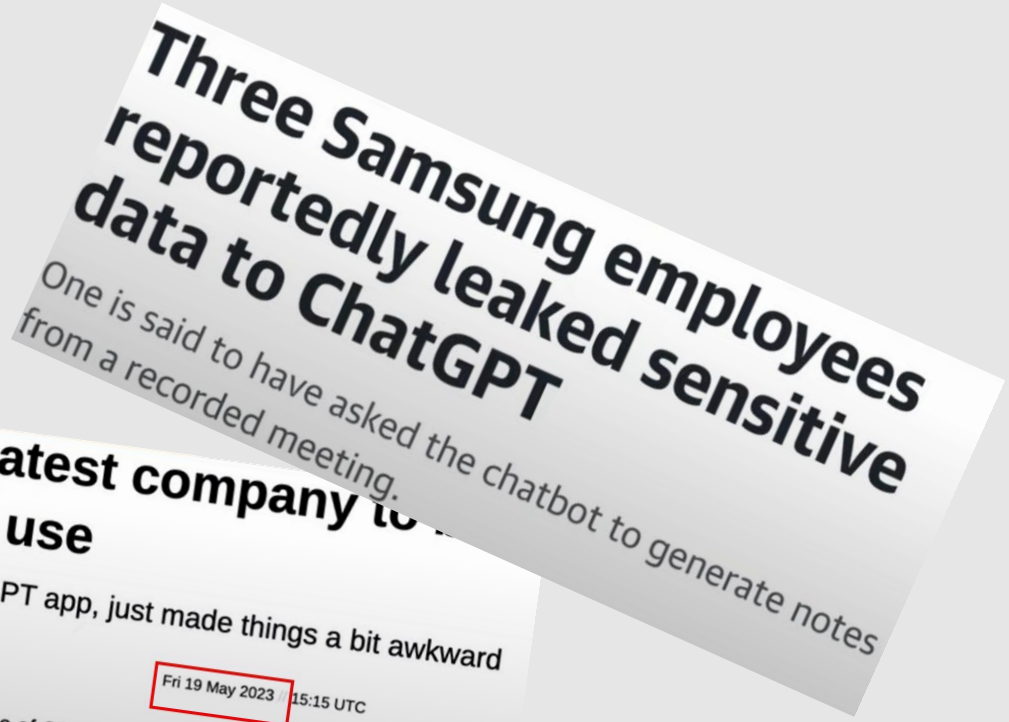
Bing Chat vs Bing Enterprise vs M365 Copilot

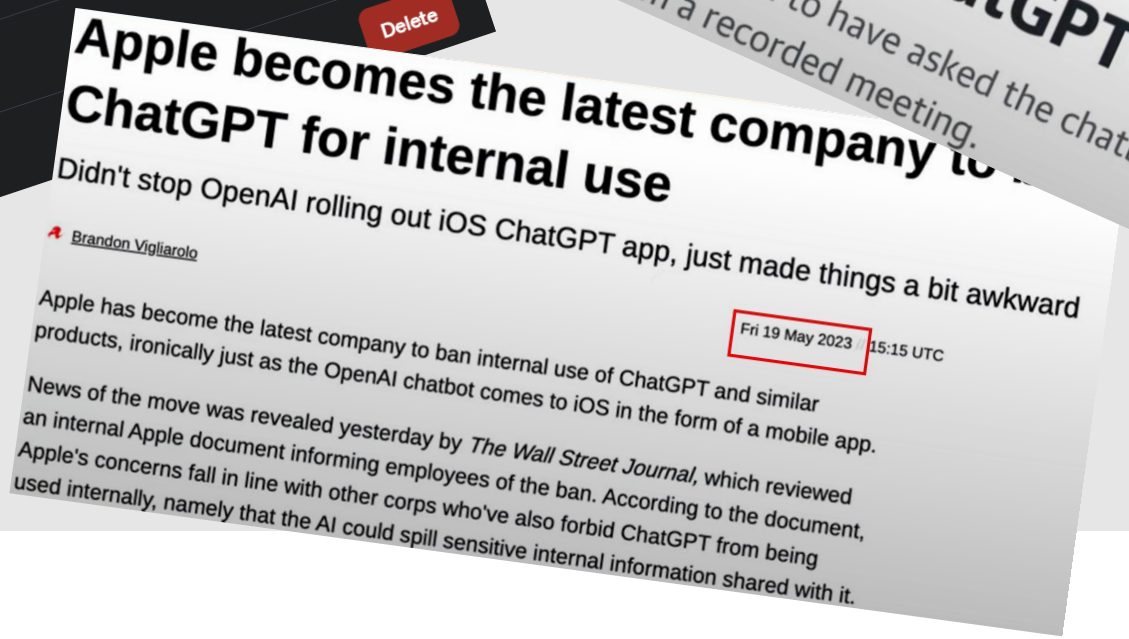
Extract rich insights from documents and summarizing them

OpenAI ChatGPT

Serious Concern in your data



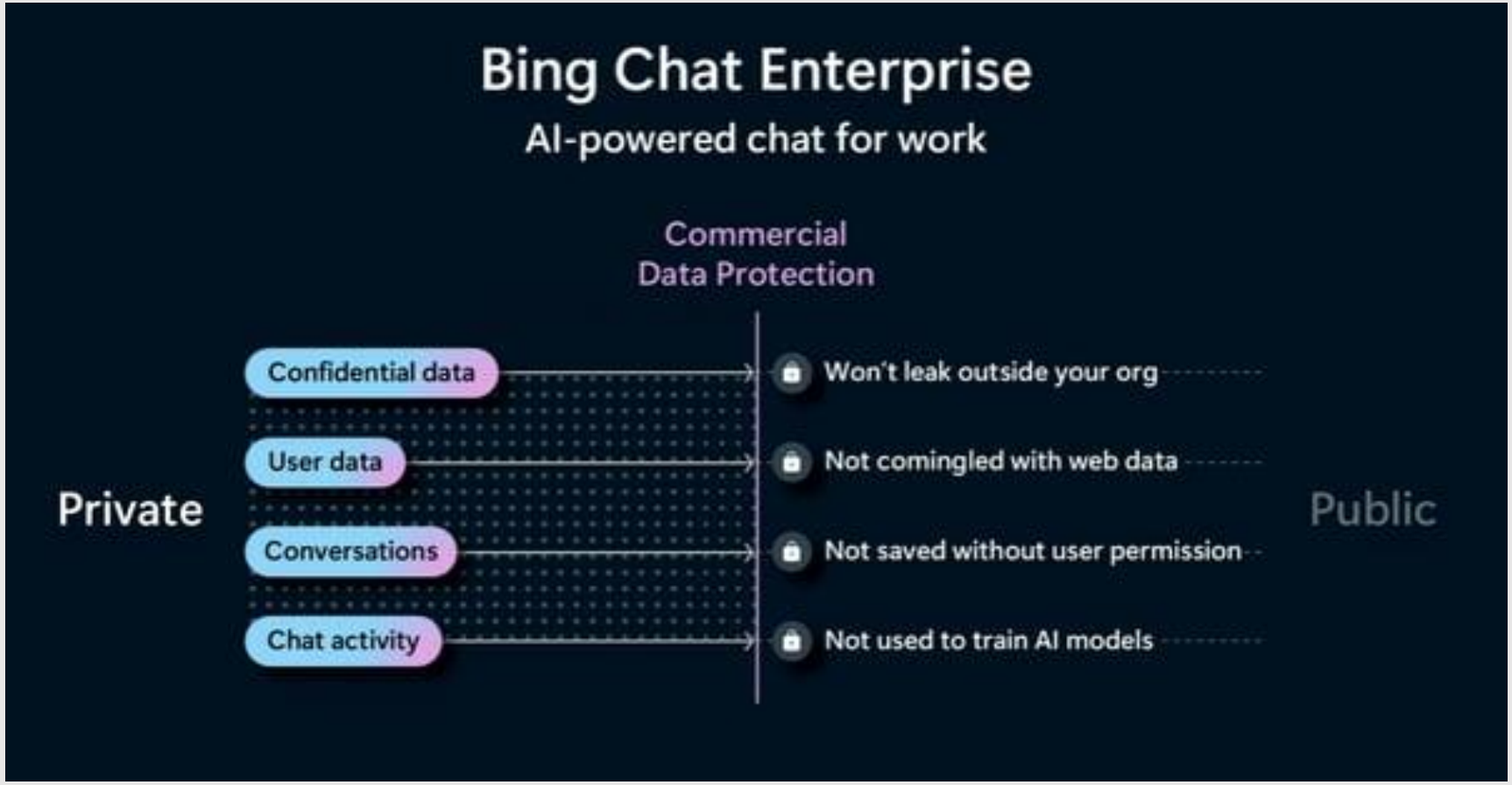




TheTechPlatform
Where Tech Meets.

Bing Chat vs Bing Enterprise vs M365 Copilot

Extract rich insights from documents and summarizing them



TheTechPlatform
Where Tech Meets.

Leverage Bing Chat Enterprise

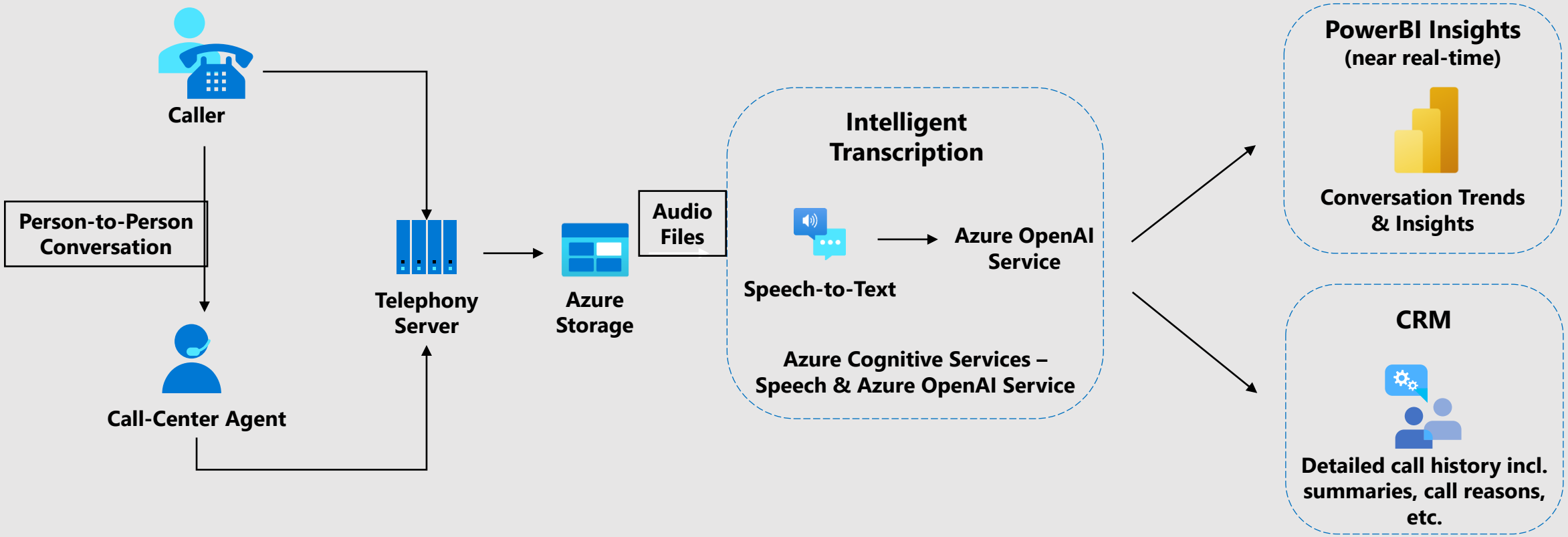
Bing Enterprise built ChatGPT

Drafting email messaging with Bing Enterprise Chat

Creating Team message with Bing Enterprise Chat

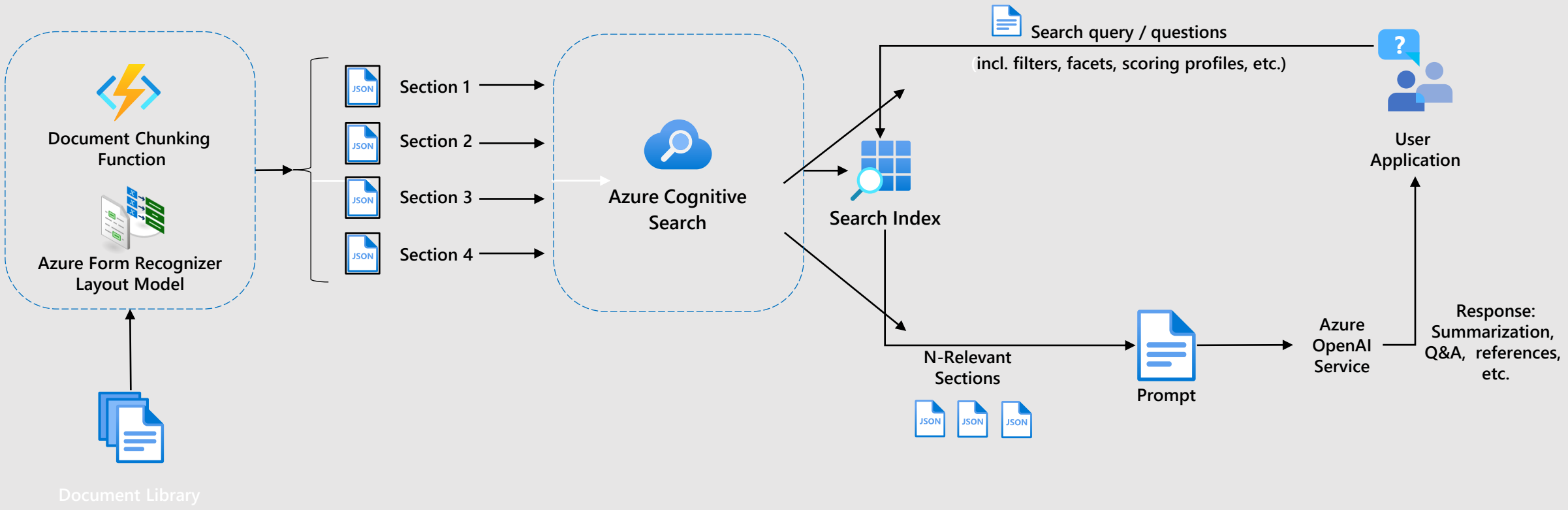
Summarizing PDF with Bing Enterprise Chat

Contact Center Analytics using Speech API & Azure OpenAI Service



TheTechPlatform
Where Tech Meets.

AI-Powered Q&A over Enterprise Data Sources



DEMO 3: Bing Chat Enterprise

Limits and Quotas

Limit Name	Limit Value
OpenAI resources per region per Azure subscription	30
Default DALL-E quota limits	2 concurrent requests
Maximum prompt tokens per request	Varies per model. For more information, see Azure OpenAI Service models
Max fine-tuned model deployments	2
Total number of training jobs per resource	100
Max simultaneous running training jobs per resource	1
Max training jobs queued	20
Max Files per resource	30
Total size of all files per resource	1 GB
Max training job time (job will fail if exceeded)	720 hours
Max training job size (tokens in training file) x (# of epochs)	2 Billion
Max size of all files per upload (Azure OpenAI on your data)	16 MB

Reference URL: Quotas and limits reference

Pricing

Pricing details:

Language models

Models	Context	Prompt (Per 1,000 tokens)	Completion (Per 1,000 tokens)
GPT-3.5-Turbo	4K	$0.0015	$0.002
GPT-3.5-Turbo	16K	$0.003	$0.004
GPT-4	8K	$0.03	$0.06
GPT-4	32K	$0.06	$0.12

Base models

Models	Usage per 1,000 tokens
Babbage-002	$0.0004
Davinci-002	$0.002

Reference: [Pricing Azure Details]

TheTechPlatform
Where Tech Meets.

DEMO 4: Pricing

TheTechPlatform
Where Tech Meets.

Using **Text Models**

- In order to use text models we need to deploy them
- After deployment we can test them and use the API
- Deployment is done in Azure OpenAI Studio

Azure OpenAI Studio

- A visual tool for working with OpenAI models
- Allows:

Model deployment

Testing with Playground

Fine tuning

Setting up Content Filters

TheTechPlatform
Where Tech Meets.

Content Filters

- AI can be abused to respond in harmful ways
- Azure OpenAI includes a built-in service to guard against that
- Uses the Azure AI Content Safety engine
  - We'll discuss it later
- You can define the content filter levels you want for your models
- Use them later in the API

TheTechPlatform
Where Tech Meets.

Quotas

- When creating a new deployment you're assigned a Quota
- Sets the maximum Tokens-Per-Minute (TPM) you can consume
- Per model, per region
- The goal is to avoid loading the OpenAI API
- You can adjust the quota to distribute it between models

Work With Your Data

- Enable GPT models to **access organizational data**
- For example: GPT will be able to **access organizational SQL Server** to pull data and include it in the response