

Balamurugan Balakreshnan/Jay Padhya

Cloud Solution Architect

babal@microsoft.com

AI Platform and Services

Microsoft corp.

LinkedIn: [Balamurugan Balakreshnan | LinkedIn](#)

Profile: [My Information \(balakreshnan.github.io\)](#)

Medium: <https://medium.com/@balabala76>

Open AI

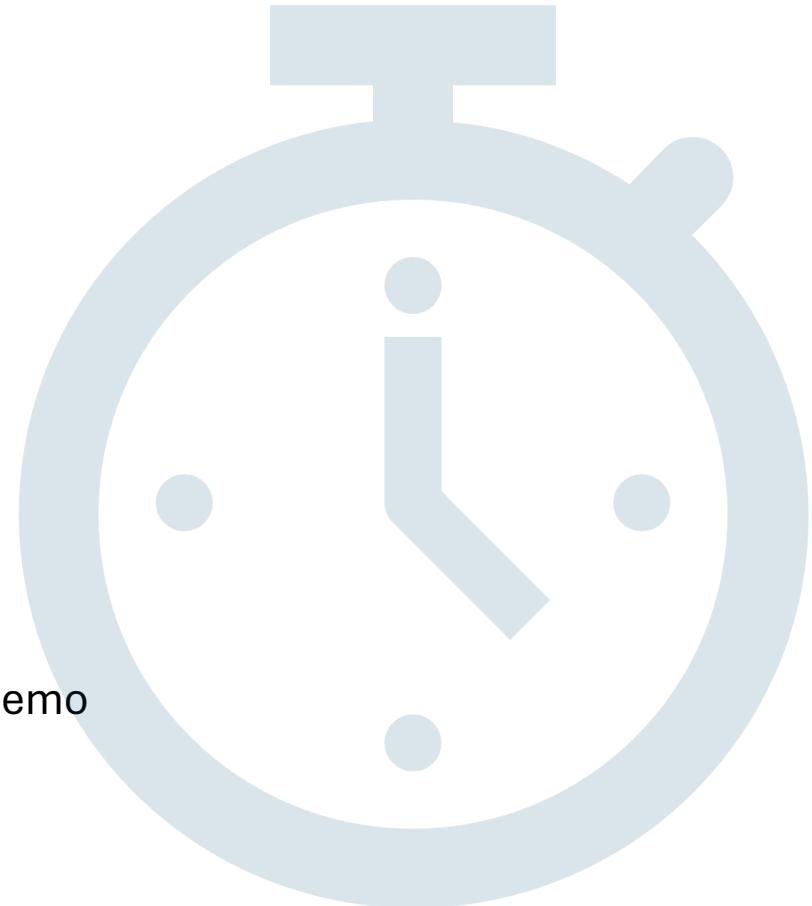
Large Language Models



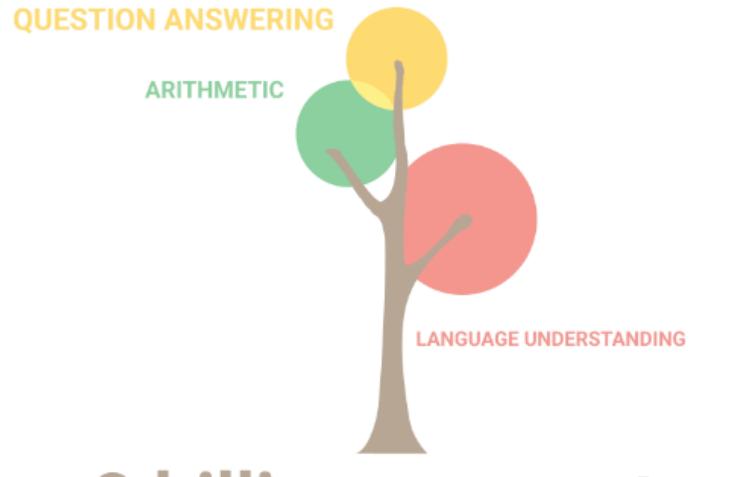


Agenda

- 1 Hour LLM and Azure Open AI
- 30 minutes - Build your application using LLM and Demo
- 1 hour – Build your application using LLM and Demo
- 15 min Q&A
- 5 min Closeout

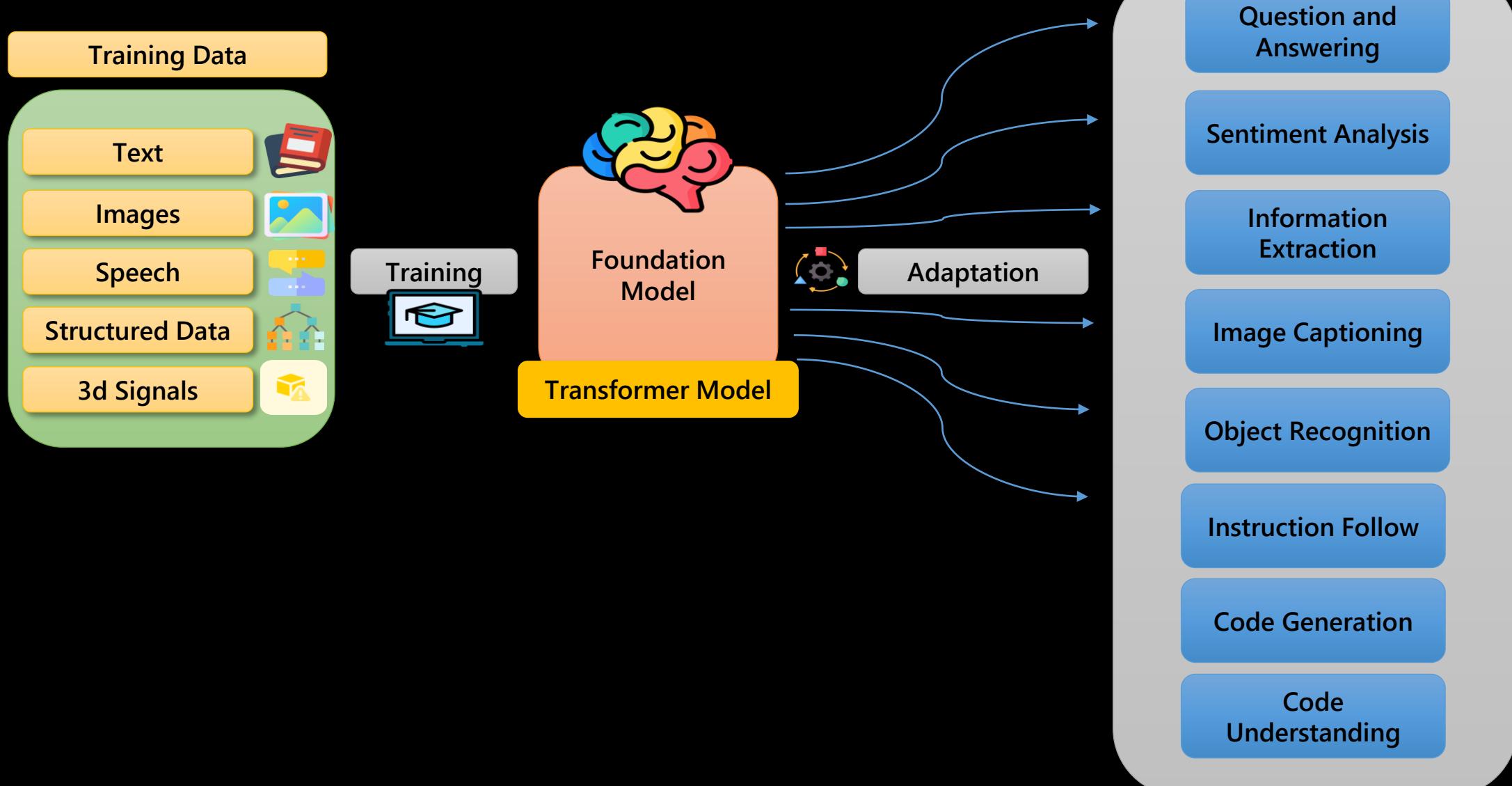


Evolution of Text and Large Language Model Tasks



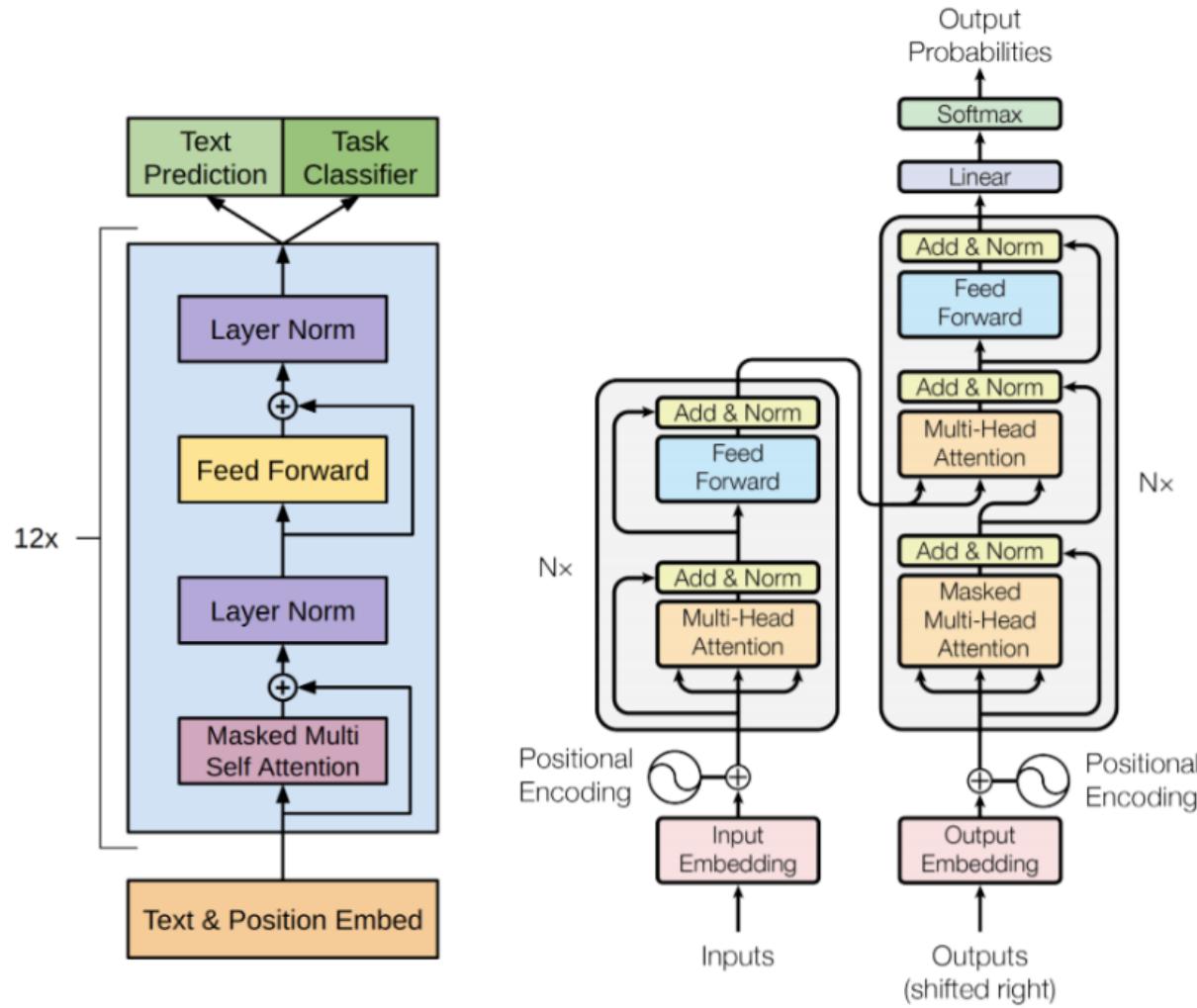
8 billion parameters

Foundation Models

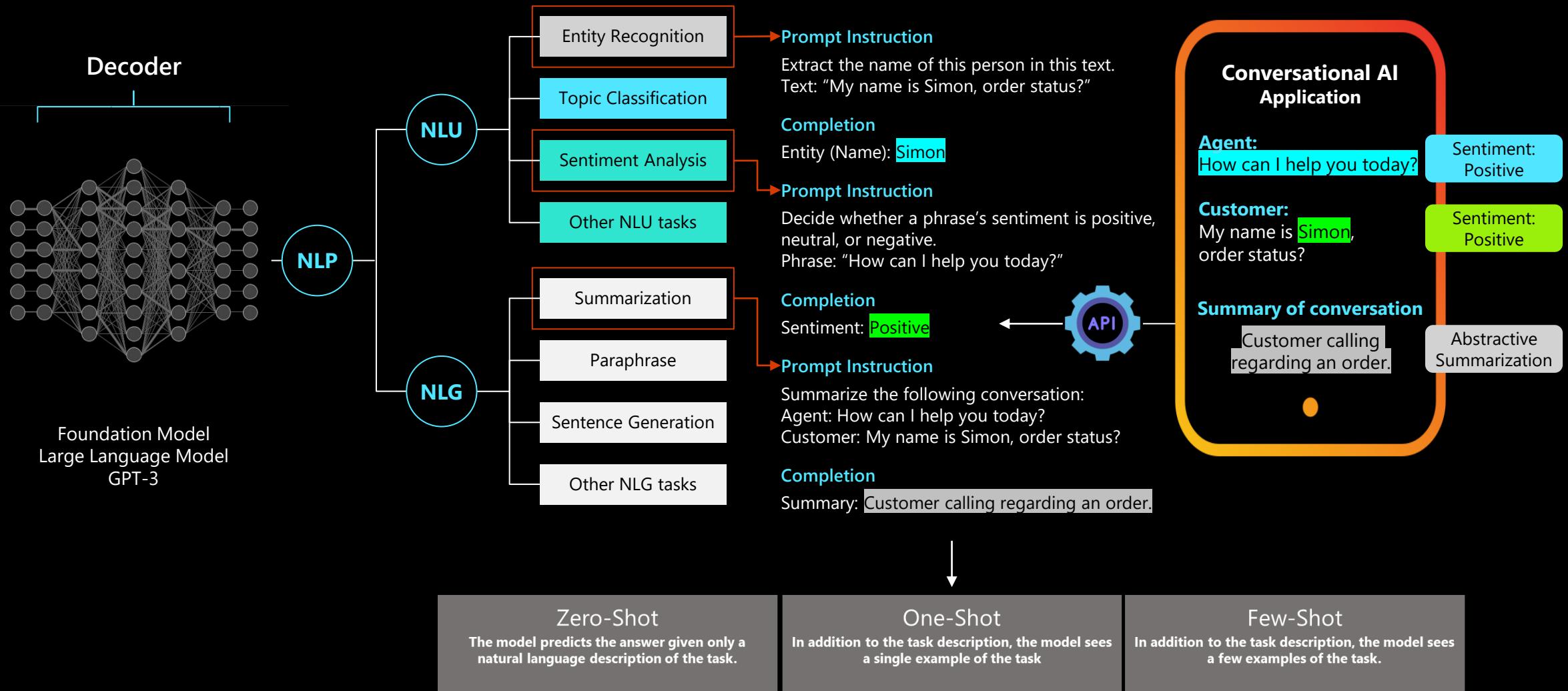




LLM Architecture

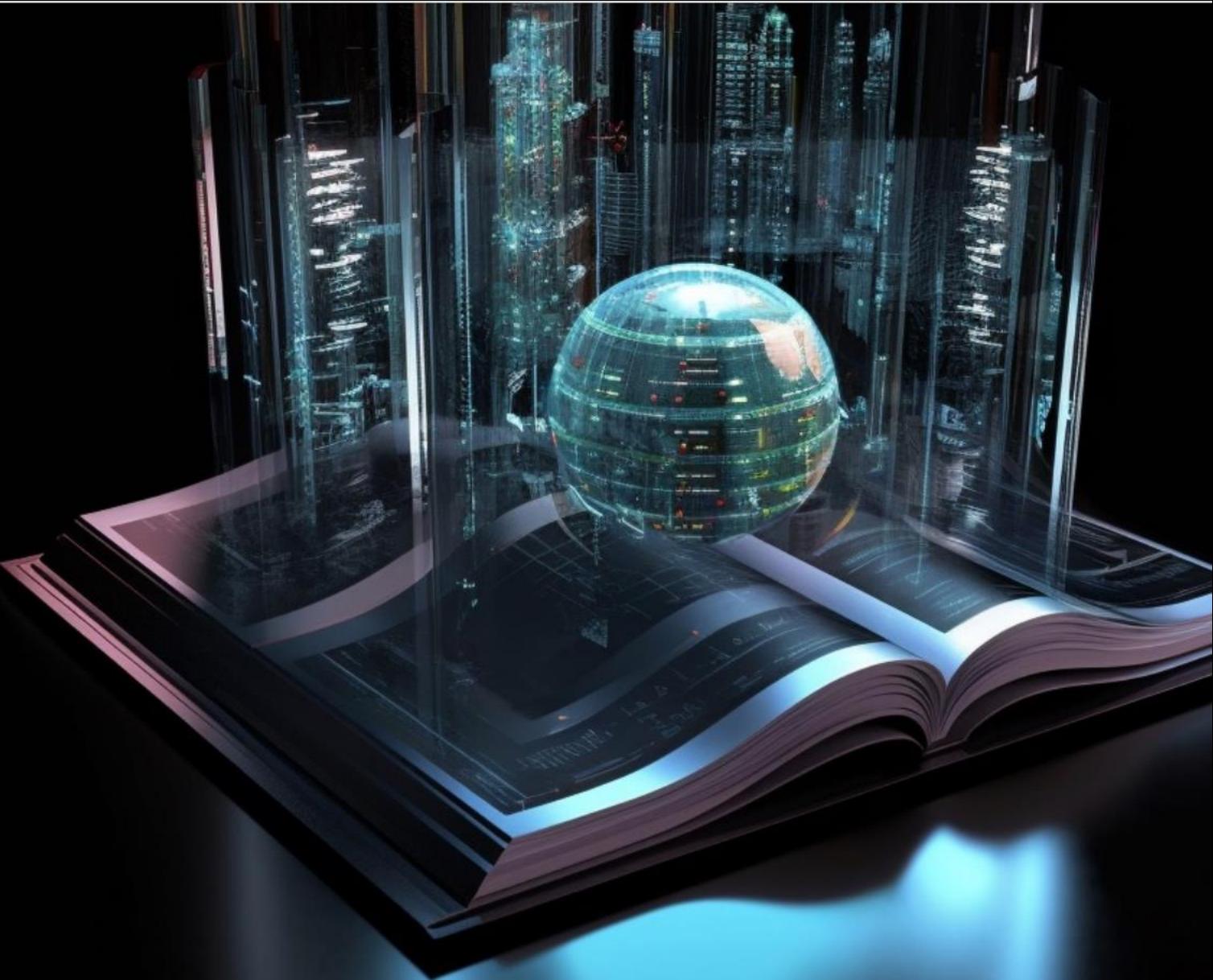


Model use out of the box—prompting

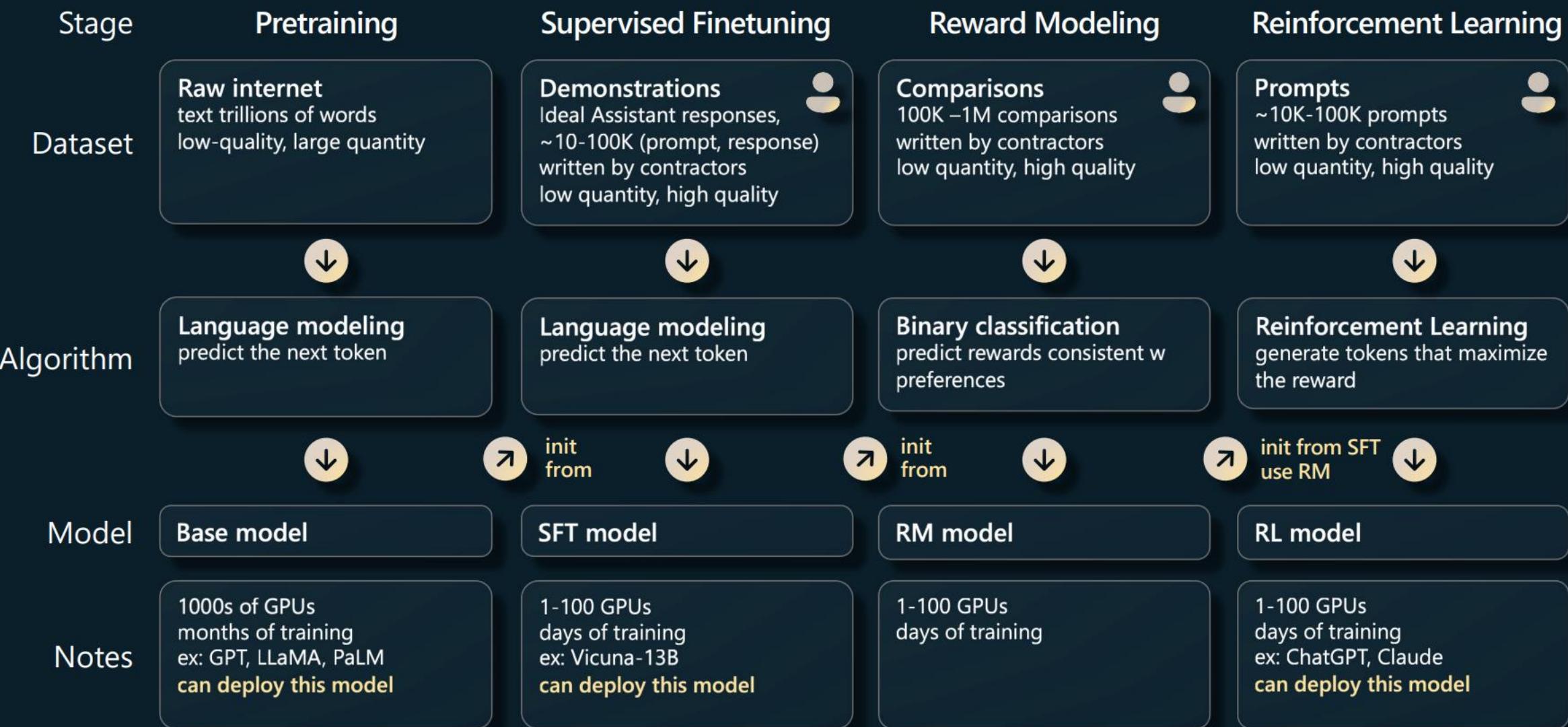


How to train your (Chat)GPT Assistant

An emerging recipe



GPT Assistant training pipeline



Data collection

Download a large amount of publicly available data



Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

Table 1: Pre-training data. Data mixtures used for pre-training, for each subset we list the sampling proportion, number of epochs performed on the subset when training on 1.4T tokens, and disk size. The pre-training runs on 1T tokens have the same sampling proportion.

[Training data mixture used in Meta's LLaMA model]

Open datasets: RedPajama, Pile

2 example models

GPT-3 (2020)

50,257 vocabulary size
2048 context length
175B parameters
Trained on 300B tokens

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

Training: (rough order of magnitude to have in mind)

- $O(1,000 - 10,000)$ V100 GPUs
- $O(1)$ month of training
- $O(1-10)$ \$M

LLaMA (2023)

32,000 vocabulary size
2048 context length
65B parameters
Trained on 1-1.4T tokens

params	dimension	n_{heads}	n_{layers}	learning rate	batch size	n_{tokens}
6.7B	4096	32	32	$3.0e^{-4}$	4M	1.0T
13.0B	5120	40	40	$3.0e^{-4}$	4M	1.0T
32.5B	6656	52	60	$1.5e^{-4}$	4M	1.4T
65.2B	8192	64	80	$1.5e^{-4}$	4M	1.4T

Table 2: Model sizes, architectures, and optimization hyper-parameters.

Training for 65B model:

- 2,048 A100 GPUs
- 21 days of training
- \$5M

[Language Models are Few-Shot Learners, OpenAI 2020]
[LLaMA: Open and Efficient Foundation Language Models, Meta AI 2023]

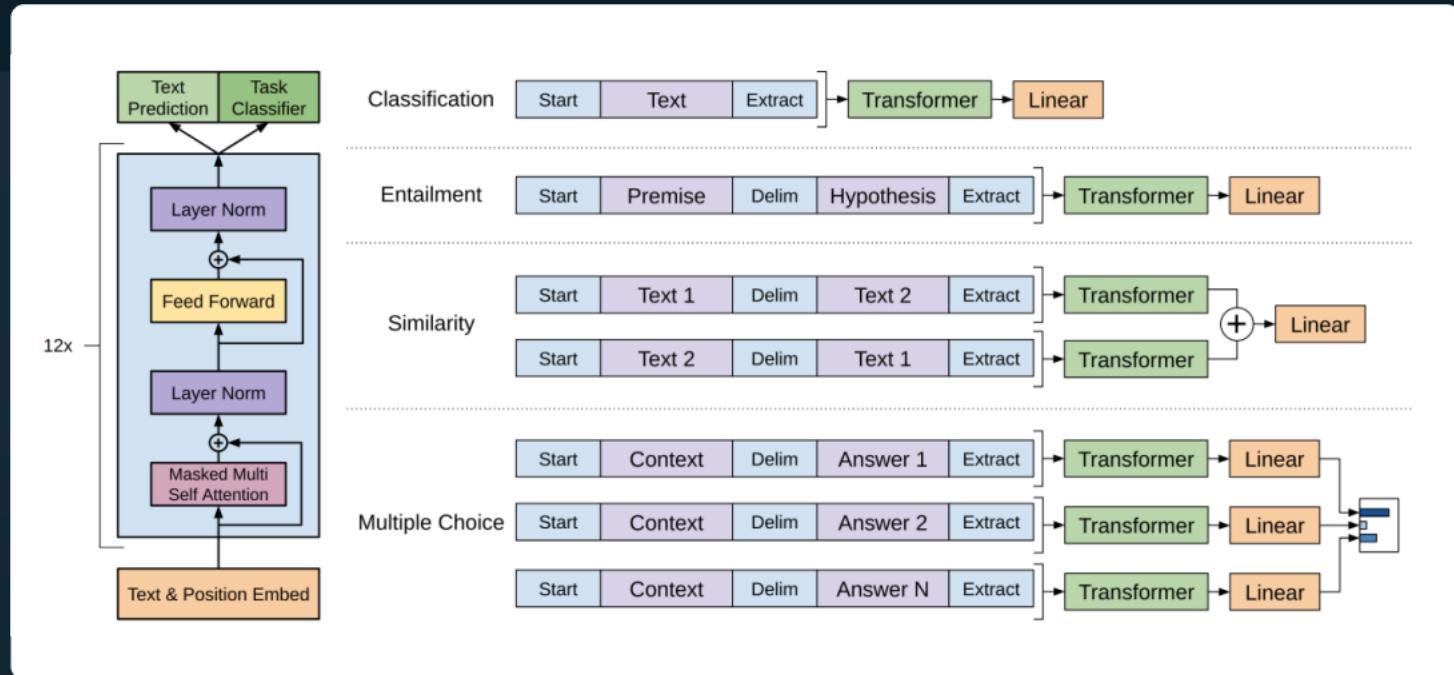
Base models learn powerful, general representations

Step 1:

Model “pretraining” on large unsupervised dataset

Step 2:

model “finetuning” on small supervised dataset



Improving Language Understanding by Generative Pre-Training, Radford et al. 2018 (GPT-1)

RL Training

Blue are the prompt tokens, identical across rows

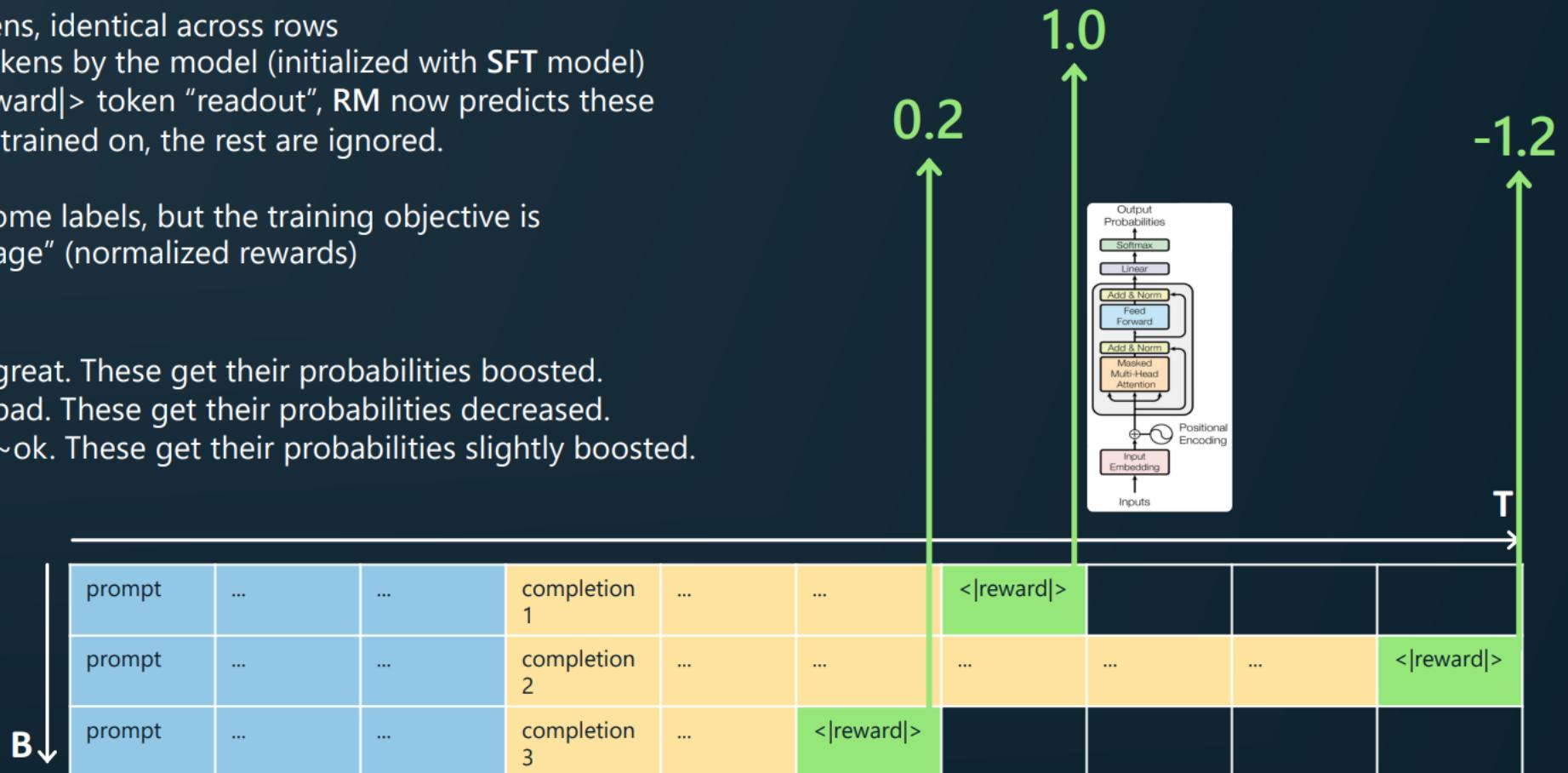
Yellow are completion tokens by the model (initialized with **SFT** model)

Green is the special `<|reward|>` token "readout", RM now predicts these
Only the **yellow** cells are trained on, the rest are ignored.

The sampled tokens become labels, but the training objective is weighted by the "advantage" (normalized rewards)

In this example:

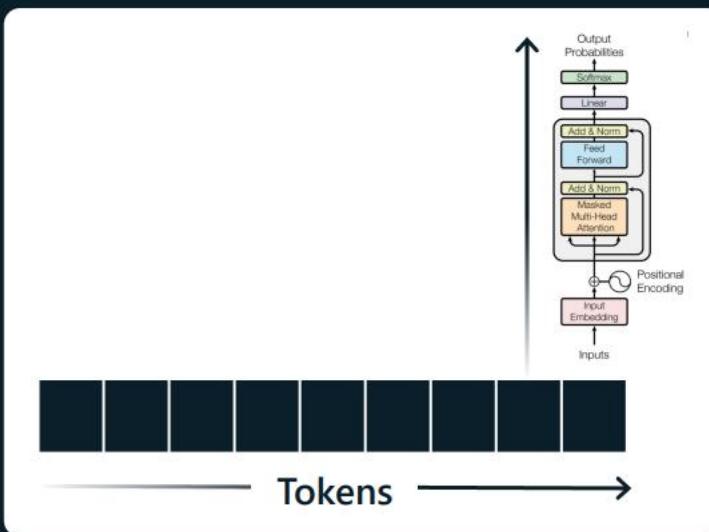
- Row #1 tokens were great. These get their probabilities boosted.
- Row #2 tokens were bad. These get their probabilities decreased.
- Row #3 tokens were ~ok. These get their probabilities slightly boosted.



[Proximal Policy Optimization Algorithms, Schulman et al. 2017]

"California's population is 53 times that of Alaska."

Human text generation vs. LLM text generation



- All of the internal monologue is stripped away in the text LLMs train on
- They spend the ~same amount of compute on every token
- => **LLMs don't reproduce this behavior by default!**
- They don't know what they don't know, they imitate the next token
- They don't know what they are good at or not, they imitate the next token
- They don't reflect. They don't sanity check. They don't correct their mistakes along the way
- They don't have a separate "inner monologue stream in their head"
- They do have very large fact-based knowledge across a vast number of areas
- They do have a large and ~perfect "working memory" (their context window)

Use cases

Models may be biased
Models may fabricate (“hallucinate”) information
Models may have reasoning errors
Models may struggle in classes of applications, e.g. spelling related tasks
Models have knowledge cutoffs (e.g. September 2021)
Models are susceptible to prompt injection, “jailbreak” attacks, data poisoning attacks,...

Recommendations:

- Use in low-stakes applications, combine with human oversight
- Source of inspiration, suggestions
- Copilots over autonomous agents



*Ensure that artificial
general intelligence (AGI)
benefits humanity*



*Empower every person and
organization on the planet
to achieve more*

GPT-3

Text

GPT 3.5 and GPT 4.0

Conversation

Codex

Code

DALL·E 2

Images



Generative AI

GPT-3/4

Prompt:

Can you summarize this personal finance article in 2-sentences for my tax-filing customers:

<https://www.fidelity.com/learning-center/personal-finance/tax-pitfalls>

Response:

Tax pitfalls can be costly, so it's important to be aware of them. Some common tax pitfalls include failing to file or pay taxes on time, not reporting all income, claiming too many deductions, and not keeping accurate records.

Fidelity offers a list of common tax pitfalls that individuals should be aware of in order to avoid costly mistakes.

Codex

Prompt:

```
Table customers,  
columns = [CustomerId,  
FirstName, LastName,  
Company, Address, City,  
State, Country,  
PostalCode]
```

Create a SQL query for all customers in Texas named Jane

```
query =
```

Response:

```
SELECT *  
FROM customers  
WHERE State = 'TX' AND  
FirstName = 'Jane'
```

DALL·E

Prompt:

a ball of fire with vibrant colors to show the speed of innovation at our media and entertainment company

Response:



ChatGPT

Prompt:

I'm having trouble getting my Xbox to turn on.

Response:

There are a few things you can try to troubleshoot this issue, First, ... Next, check the power cord ...

Prompt:

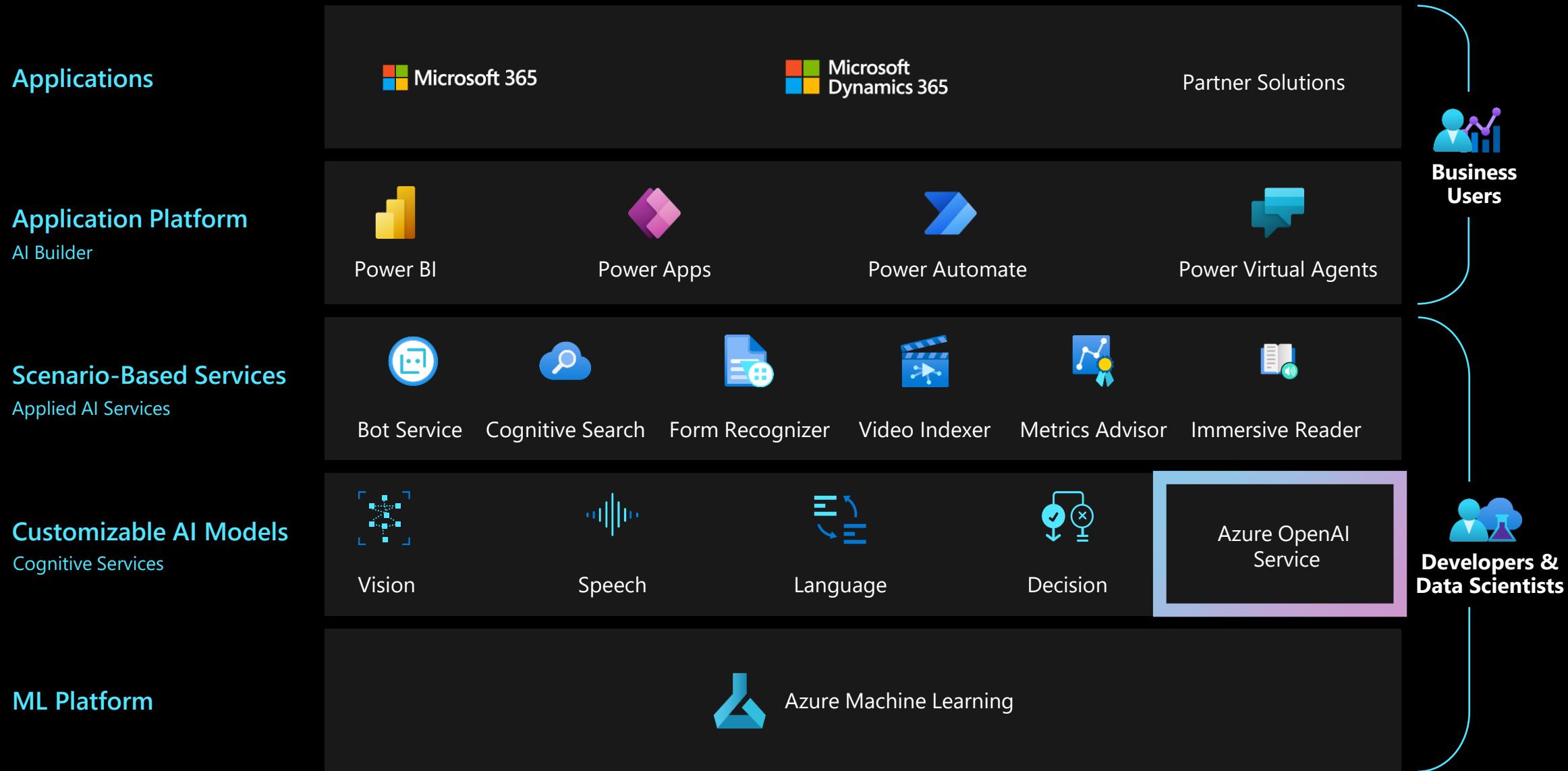
Thanks! That worked. What games do you recommend for my 14-year-old?

Response:

Here are a few games that you might consider:

(bulleted list) ...

Azure AI



Azure OpenAI Service

Top capabilities



Content generation



Summarization



Code generation



Semantic search

Customer use cases



Azure AI

AI you can trust

Your Azure OpenAI Service instance
is isolated from every other customer

Your data is not used to train
the foundation AI models

Your data is protected by
the most comprehensive enterprise
compliance and security controls

| Microsoft Azure Cloud

Runs on trust

Your data is your data

Data is stored encrypted in your Azure subscription

Your data from any fine-tuning
is not used to train the
foundation AI models

Azure OpenAI Service provisioned in your Azure subscription

Model fine tuning stays in your Azure subscription
and never moves into the foundation AI models

Your data is protected by
the most comprehensive enterprise
compliance and security controls

Encrypted with Customer Managed Keys

Private Virtual Networks, Role Based Access Control

Soc2, ISO, HIPPA, CSA STAR Compliant

Privacy and Security in Azure OpenAI Service

<https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy>

- Authorized Microsoft employees will have access to prompts and generated text data that triggered automated systems for the purpose of investigating and verifying potential abuse (*1)

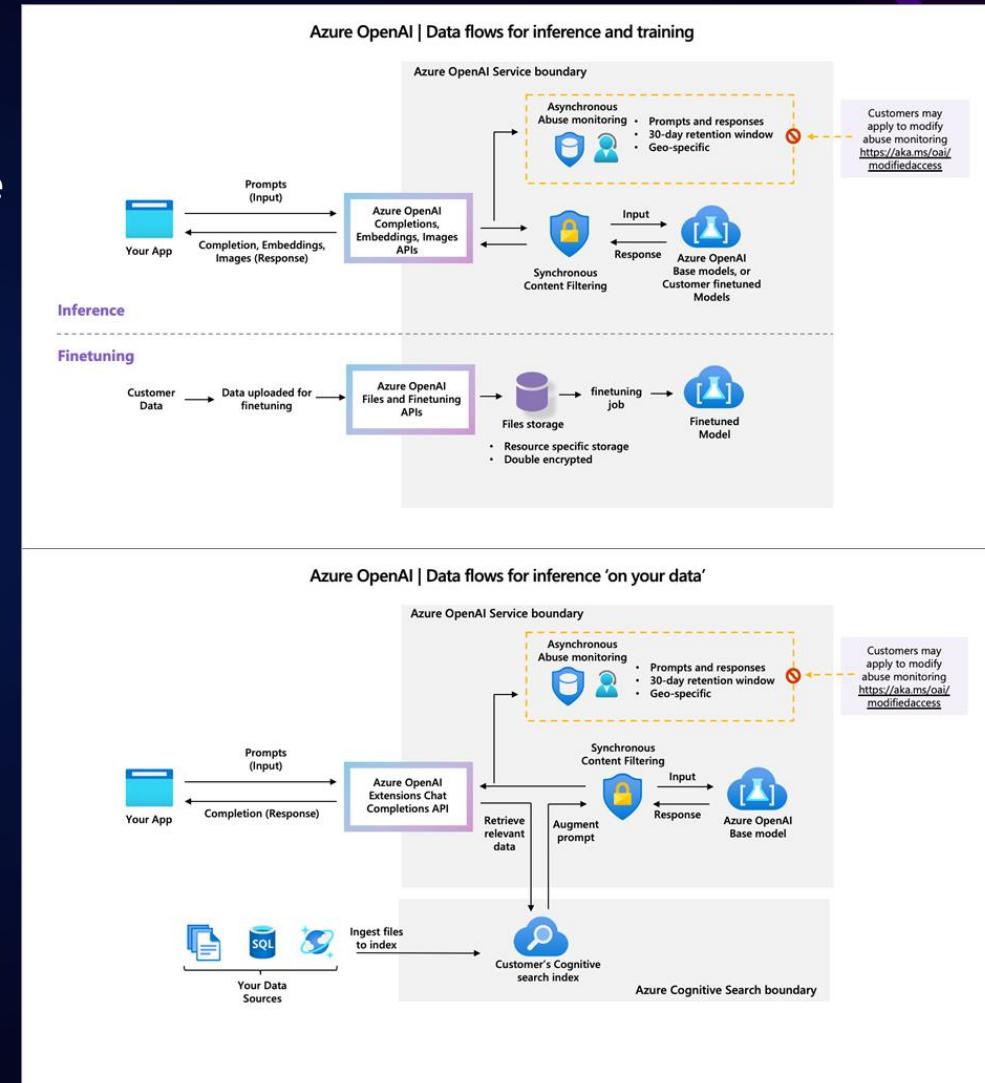
You can request the ability to opt out of both data logging and the human review process. This allows trusted users with low-risk scenarios to have the data and privacy controls they need

- Training data uploaded for fine tuning purposes is stored in Azure Storage and encrypted at rest with Microsoft-managed keys

Customer encryption keys can also be used

- The training data is only used to fine-tune your model, and Microsoft will NOT train the model. or will not be used to improve

Same terms as other Azure AI services



(*1) For customers deploying Azure OpenAI services in the EU, approved Microsoft employees will be located in the EU.

Update

Azure OpenAI Service

GPT-3

Codex

DALL·E

ChatGPT

GPT-4



Deploy on your
own data

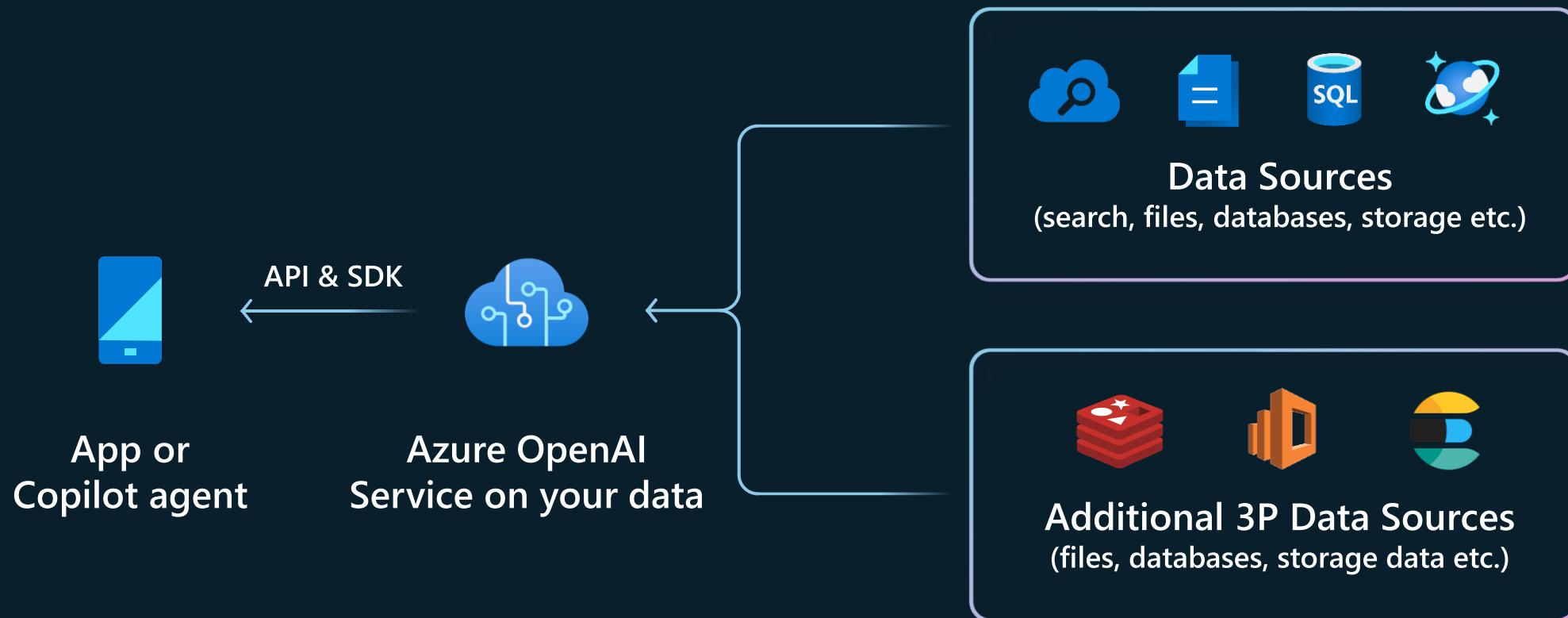


Provisioned
Throughput
Model



Plugins for
Azure OpenAI
Service

Azure OpenAI Service on your data





Demo

Just Announced

Azure AI Studio



Build and
train your
own
models



Ground
Azure
OpenAI
Service and
OSS models
using your
data



Built-in
vector
indexing



Retrieval
augmented
generation
made easy



Create
prompt
flows



AI safety
built-in

Microsoft Copilots powered by Azure OpenAI

Microsoft 365 Copilot

Works alongside
you in the apps you
use every day

[Learn more](#)

Dynamics 365 Copilot

Turbocharge your
workforce with a
copilot for every
job role

[Learn more](#)

Copilot in Power Platform

Imagine it, describe
it, and Power
Platform builds it

[Learn more](#)

Microsoft Security Copilot

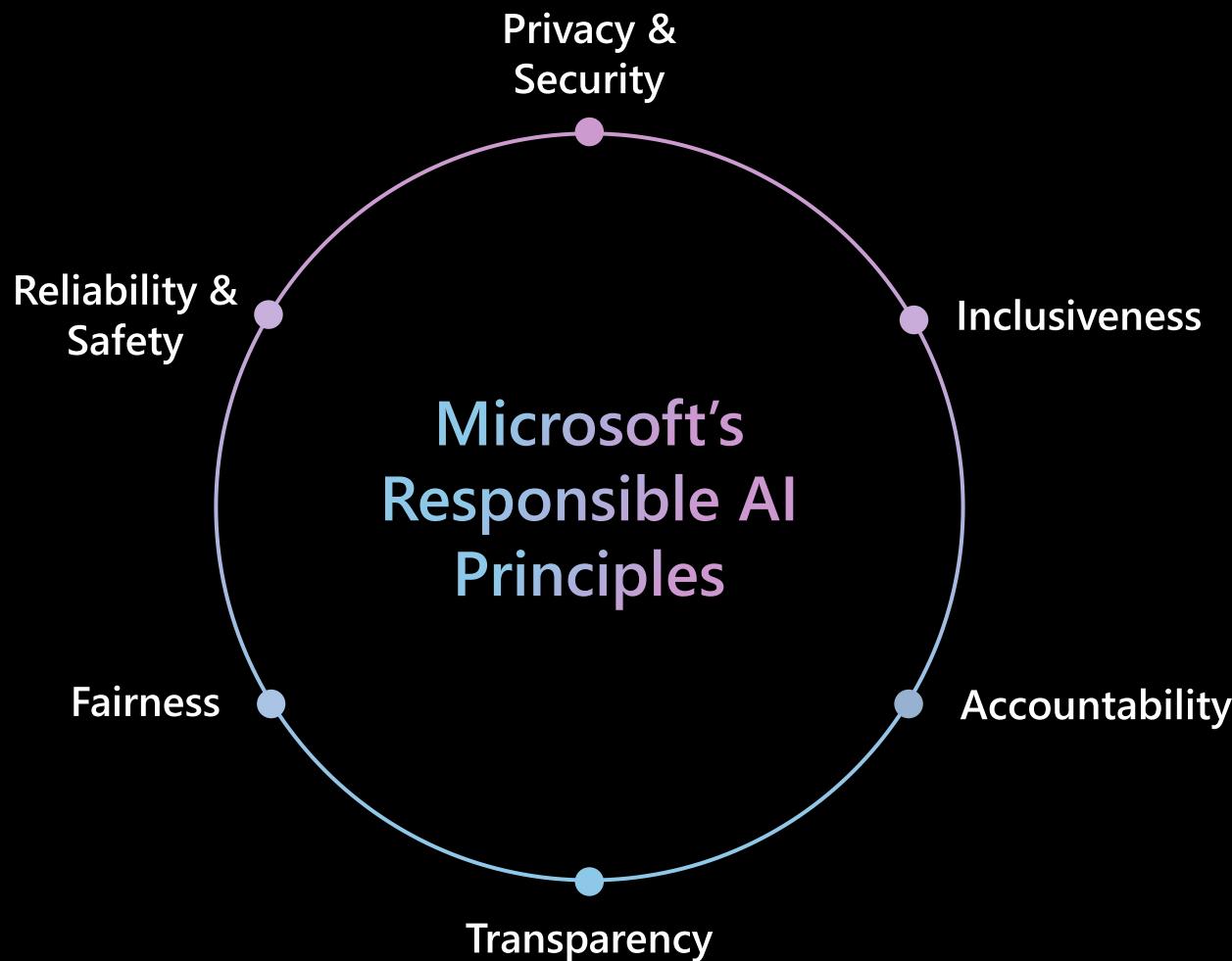
Defend at machine
speed with
Microsoft Security
Copilot

[Learn more](#)

GitHub Copilot

Increase developer
productivity to
accelerate
innovation

[Learn more](#)



Building blocks to enact principles



Tools and processes



Training and practices



Rules



Governance

Azure OpenAI Service customer momentum



servicenow.

DataRobot

Icertis

Providence

snelstart

Albusters

ALERT
INNOVATION

Adobe

AON



KPMG

AUTODESK

BMW GROUP



CARMAX

clevertar

Coca-Cola

cognizant
intuition engineered

Databook

DocuSign

Downer

ELSEVIER
Health

EY

Farmlands
co-operative

GENIE AI

JIFAD

accenture

Inpris

inworld

Kepro®
Making healthcare work.

klaviyo

arvato
BERTELSMANN

nationale
nederlanden

ODP

pwc

R T L +



avanade

sogeti
Part of Capgemini

soul
machines™

STRABAG

Trelent

Typeface

CIPIO.ai

WordLift

zammo

An example of Prompt Engineering- Topic Classifier

Classify customer's question. Classify between category 1 to 3.

Detailed guidelines for how to choose:

choose 1 if the question is about auto insurance.

choose 2 if the question is about home flood insurance.

choose 3 if the question is not relevant to insurance.

Instructions
High level and detailed

Customer question: Hi there, do you know how to choose flood insurance?

Classified topic: 2

Customer question: Hi there, I have a question on my auto insurance.

Classified topic: 1

Customer question: Hi there, do you know how to apply for financial aid?

Classified topic: 3

Examples
Order of examples matter

Customer question: {insert new question here}

Classified topic:

Task and Prompting
answer

Two Types of Prompt Engineering

Zero-shot

The model predicts the answer given only a natural language description of the task.

Classify customer's question. Classify between category 1 to 3.

Detailed guidelines for how to choose:

choose 1 if the question is about auto insurance.

choose 2 if the question is about home flood insurance.

choose 3 if the question is not relevant to insurance.

Customer question: {insert new question here}

Classified topic:

Few-shot

In addition to the task description, the model sees a few examples of the task.

Classify customer's question. Classify between category 1 to 3.

Detailed guidelines for how to choose:

choose 1 if the question is about auto insurance.

choose 2 if the question is about home flood insurance.

choose 3 if the question is not relevant to insurance.

Customer question: Hi there, do you know how to choose flood insurance?
Classified topic: 2

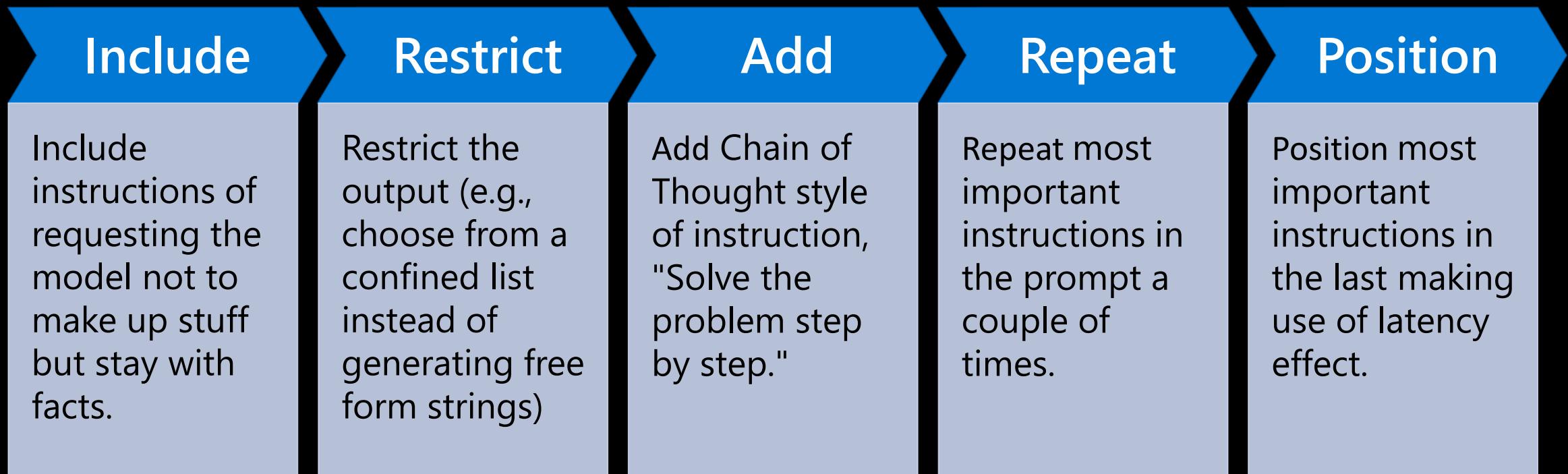
Customer question: Hi there, I have a question on my auto insurance.
Classified topic: 1

Customer question: Hi there, do you know how to apply for financial aid?
Classified topic: 3

Customer question: {insert new question here}

Classified topic:

Reduce Hallucination using Prompt Engineering



What is Fine-Tuning?

Fine-tuning is a way of utilizing transfer learning. Specifically, fine-tuning is a process that takes a model that has already been trained and tune it using a labeled dataset for a specific task.

Fine-tuning results in a new model being generated with updated weights and biases. This contrasts with few-shot learning in which model weights and biases are not updated.

To fine-tune a model, you'll need a set of training examples that each consist of a single input ("prompt") and its associated output ("completion").

When to use Fine-Tuning?



For some use cases, fine-tuned less capable models (Ada, Babbage, Curie) performs well but more cost effective and faster than Davinci.



Latency is important for use case. Because no need to include examples in the inputs to fine-tuned models, tokens on few-shot examples are removed, this enables lower-latency requests and saves inferencing costs.

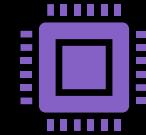


Customer has a lot of high-quality data and ground truth labels (at least a couple hundred for each use case), and customer committed to maintain high-quality data set overtime.



Use case is very domain specific, so model needs to learn from a lot of examples to know how to provide correct answers, and the number of examples needed don't fit in prompt due to token limit (less likely with 32K token limit).

General Recommendations



Direct use of foundation models with prompt engineering is preferred. For prompt engineering, try out a few different prompts and pick the best one.

Utilizing similarity search to create dynamic prompts can be beneficial in cases where the total content is too extensive, as it allows for the incorporation of the most pertinent information for each query.

For fine-tuning, only use the highest quality dataset. Monitor validation metrics and adjust parameters to avoid overfitting/underfitting.

Experiment with prompt engineering on advanced models and fine-tune less capable models, compare the performance metrics, operational latency and cost.

Azure OpenAI Service In Manufacturing & Mobility Overview

Azure OpenAI Service in Manufacturing & Mobility

Key Areas for Azure OpenAI Service to add value in our industry

These are new with Azure OpenAI Service—Generative AI

Create

Generative AI systems generates text, images, audio, writing samples, and other artifacts that can be built with computer-controlled systems like 3D printers

Identify

AI systems identify things like people in pictures, words in speech or handwriting and—most importantly—what's real vs. what's fake

These are improved with Azure OpenAI Service

Industry customer scenario

Demand forecasting	Machine translation
Automated quality inspection	Knowledge management
Predictive maintenance	Automated content generation
Remote asset monitoring	Analytics
Connected smart product	Sentiment analysis
Document process automation	Generating reports
Contact center	Developer productivity

Microsoft in Manufacturing & Mobility

Transform your workforce



Engage customers in new ways



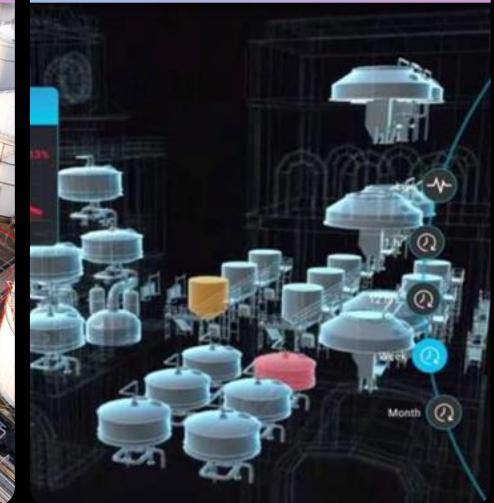
Build more agile factory/plants



Create more resilient supply chains



Unlock innovation and deliver new services



Sustainability | Security | Compliance

M&M Azure OpenAI Service Scenarios

Demand Forecasting



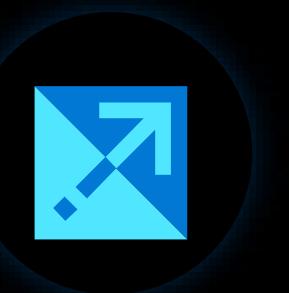
Issues & Challenges

- *Disconnected processes*
- *No feedback loop due to limited visibility of key data from the supply chain*
- *Departments operating in silos and restricted data flow*



Azure OpenAI Service Value Add

- *Allow Planners to use natural language to query forecast predictions*
- *Convert unstructured NLP feeds into structured data points*

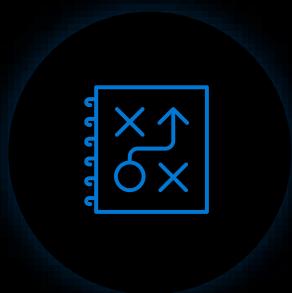


Potential Outcomes

- *Improve time to value* *through increased forecast speed and reduction in project time*
- *Reduce inventory carrying costs* *by predicting necessary stock levels at any given time*
- *Grow revenue* *by increasing in-stock percentage through stronger forecast accuracy*

M&M Azure OpenAI Service Scenarios

Automated Quality Inspection



Issues & Challenges

- **Latent anomaly detection:** Computation and data inference are often performed by remote applications
- **Lack of data insights:** IoT data is siloed and isn't leveraged for strategic decision-making.
- **Velocity of developing for the edge:** Edge technology requires significant coding and testing in a live environment.



Azure OpenAI Service Value Add

- Provide NLP queries on the vision anomalies and causes
- Use Synthetic Imaging to augment baseline ML training image volumes to improve ML accuracy and data set sizes

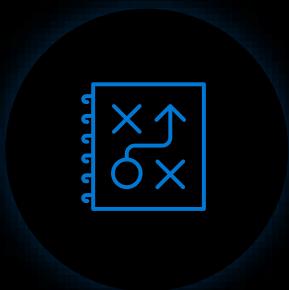


Potential Outcomes

- **Improve quality assurance pass rates and product yield by** Automating inspection with higher accuracy and redirecting human labor to value adding tasks
- **Increase customer satisfaction and retention by** Reducing delivery of defective products
- **Reduce operational costs by** catching anomalies faster, reducing amount of inventory that needs to be scrapped

M&M Azure OpenAI Service Scenarios

Predictive Maintenance



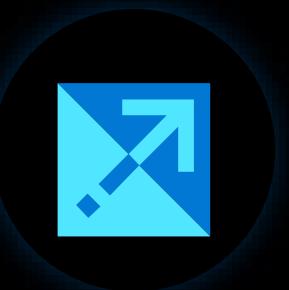
Issues & Challenges

- Data is not being captured from the factory floor
- Large upfront investment
- Security risks on data capture with legacy equipment
- Complexity of getting started



Azure OpenAI Service Value Add

- Converting unstructured data to structured SQL data points
- Improve Knowledge Mining

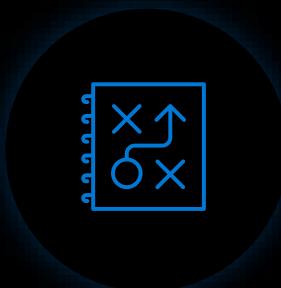


Potential Outcomes

- **Increase machinery uptime** *by* Predicting problem areas and solving before they become breaks
- **Reducing maintenance costs and repair time** *by* Proactively addressing problem areas before it becomes a larger failure
- **Reducing inventory carrying costs** *by* Accurately understanding spare part inventory needs at any given time

M&M Azure OpenAI Service Scenarios

Remote Asset Monitoring



Issues & Challenges

- Need to **improve equipment effectiveness and asset productivity** by capturing and analyzing all connected shopfloor data
- Need to **avoid unplanned downtime** due to unexpected machine failures
- Deliver real-time platforms that enable factories to **closely monitor daily operation and perform real-time diagnostics remotely**



Azure OpenAI Service Value Add

- **Use NLP to mine knowledge sources**
- **Convert unstructured data to structure data**
- **Convert NLP issues to images for analysis/comparison**

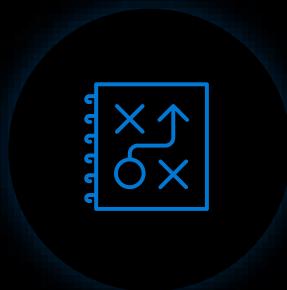


Potential Outcomes

- Machine learning models **generate real-time recommendations** across physical processes
- **Higher overall equipment utilization** and asset productivity
- Improved employee productivity with freeing up of time for performing more value-added tasks
- **High app performance and instant efficiency insights** worldwide with low latency

M&M Azure OpenAI Service Scenarios

Connected Smart Product



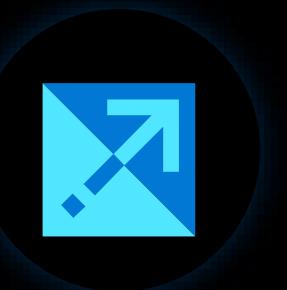
Issues & Challenges

- Product **lifecycles continue to grow in complexity**, which limits visibility into processes and hinders continual improvements
- Ubiquitous connectivity, dynamic customer demands, and volatile procurement behavior are causing long-term **changes in the way products are designed** and managed
- Need ways to keep pace with rapid product design changes to **improve customer satisfaction and retention**



Azure OpenAI Service Value Add

- Convert unstructured NLP feeds into structured data points
- Allow designers to generate product design based on consumer sentiment and feedback
- Optimize material usage/requirements in the design process

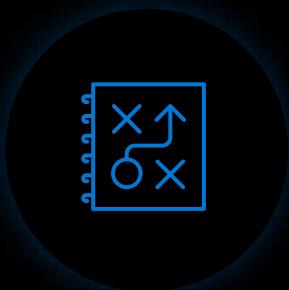


Potential Outcomes

- **Shorten the time to market** by updating features of digitally connected products through the cloud
- **Become nimbler in competitive responsiveness** by Applying digital transformation on traditional product innovation processes
- **Prevent bottlenecks and production challenges** through Improvement and validation of designs by production engineers
- **Improve operational efficiencies** by Migrating on-premise infrastructures with poor performance and limited scale

M&M Azure OpenAI Service Scenarios

Contact Center



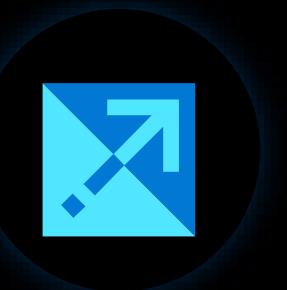
Issues & Challenges

- General staffing shortage
- Increase cost of workforce
- Increase customer expectation



Azure OpenAI Service Value Add

- Speech API & GPT-3 can extract rich insights from call transcripts.
- E2E Call Center analysis: classification, sentiment, summarization and email generation.

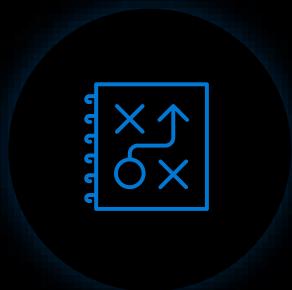


Potential Outcomes

- Automatically generate responses to customer inquiries.
- Summary of customer support conversation logs
- Improve customer satisfaction
- Improve time to resolution of customer inquiries

M&M Azure OpenAI Service Scenarios

Automated content generation



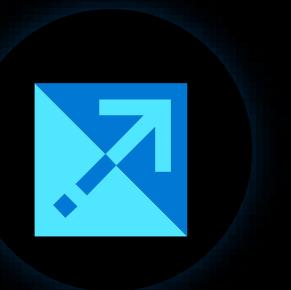
Issues & Challenges

- Too much data to properly curate within a given SLA
- Time consuming tasks to summarize product reviews/feedback
- Difficult to segment, personalize, and filter reviews/feedback



Azure OpenAI Service Value Add

Generative AI can be used to automatically generate content on a variety of topics, saving time and money for content creators.



Potential Outcomes

- Generate more relevant contents with summarization
- Improve SEO

Azure OpenAI Service

Top 4 capabilities



Content generation

Call center analytics: automatically generate responses to customer inquiries

Generate personalized UI for your website



Summarization

Call center analytics: summary of customer support conversation logs

Subject matter expert document: summarization (e.g., Financial reporting, analyst articles)

Social media trends summarization



Code generation

Convert natural language to SQL (or vice versa) for telemetry data

Convert natural language to query proprietary data models

Code documentation



Semantic search

Search reviews for a specific product/service

Information discovery and knowledge mining

Examples of multiple model use cases

End to end call center analytics: classification, sentiment, entity extraction, summarization and email generation

Customer 360: hyper-personalisation using timely summarization of customer queries & trends, search, and content generation

Business process automation: search through structured & unstructured documentation, generate code to query data models, content generation

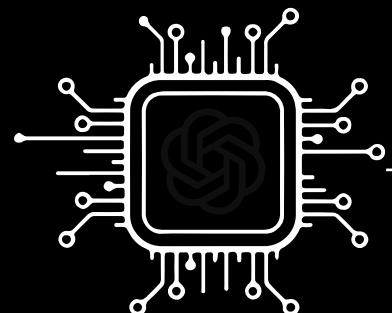


GitHub Copilot

Your AI pair programmer

```
19     screenshot = ImageGrab.grab()  
20     # Convert to text  
21     text = image_to_string(screenshot)  
22     # Parse text for email addresses  
23     emails = re.findall(r'[\w\.-]+@[\\w\.-]+\.', text)  
24     return emails  
25  
26 def validate(addresses) :  
27     |  
28
```

OpenAI Codex Model



Public code and text
on the internet

GitHub



GitHub
Copilot Service

Provide editor context

Provide suggestions

Improve suggestions

Private code

```
JS fetch_pic.js    push_to

1 const fetchNASAPictureOfTheDay = () => {
2   return fetch('https://api.nasa.gov/mars/rovers/curiosity/images/latest?sol=1000')
3     .then(response => response.json())
4     .then(json => {
5       return json;
6     });
7 }
8
9
10
11
12 }
```

Copilot



Copilot for Every Microsoft Cloud Experience

Microsoft 365 Copilot

Works alongside you in the office apps you use every day

Dynamics 365 Copilot

Turbocharge your workforce with a copilot for every job role

Copilot in Power Platform

Imagine it, describe it, and Power Platform builds it

Microsoft Security Copilot

Defend at machine speed with Microsoft Security Copilot

GitHub Copilot

Increase developer productivity to accelerate innovation



Microsoft

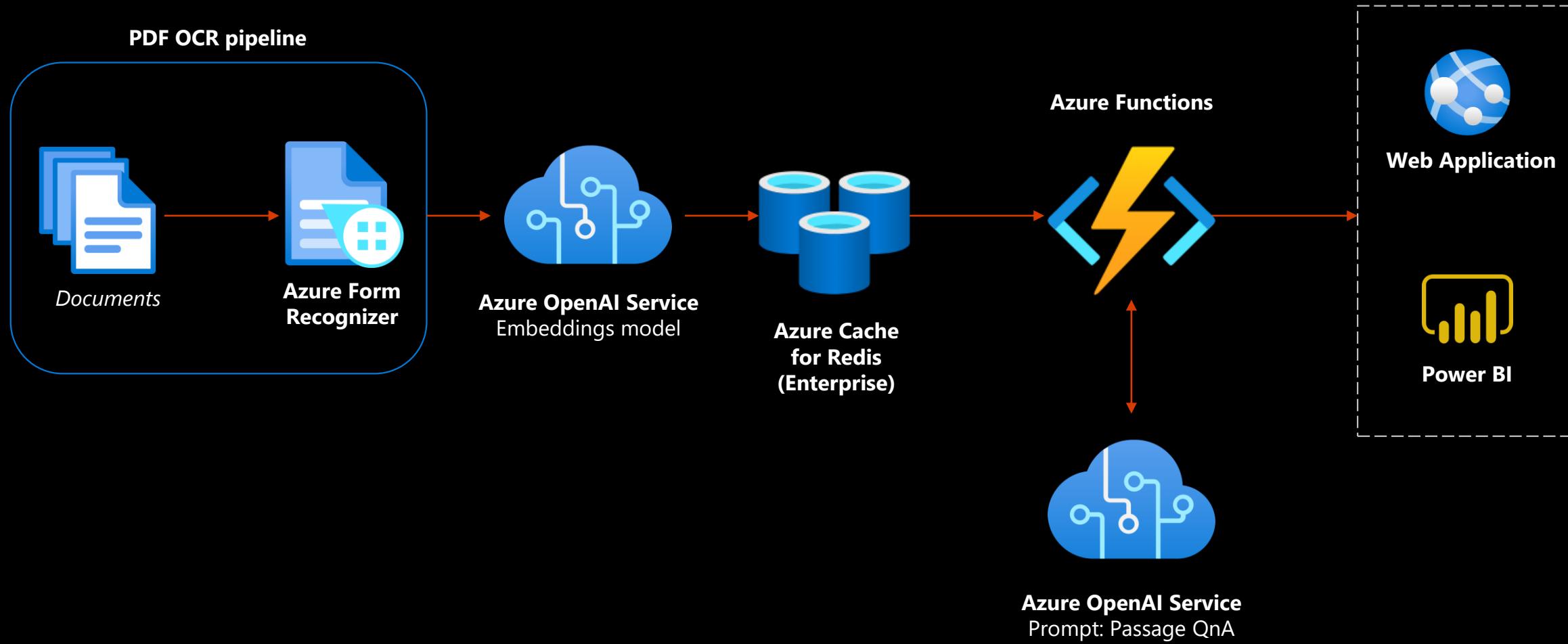


OpenAI



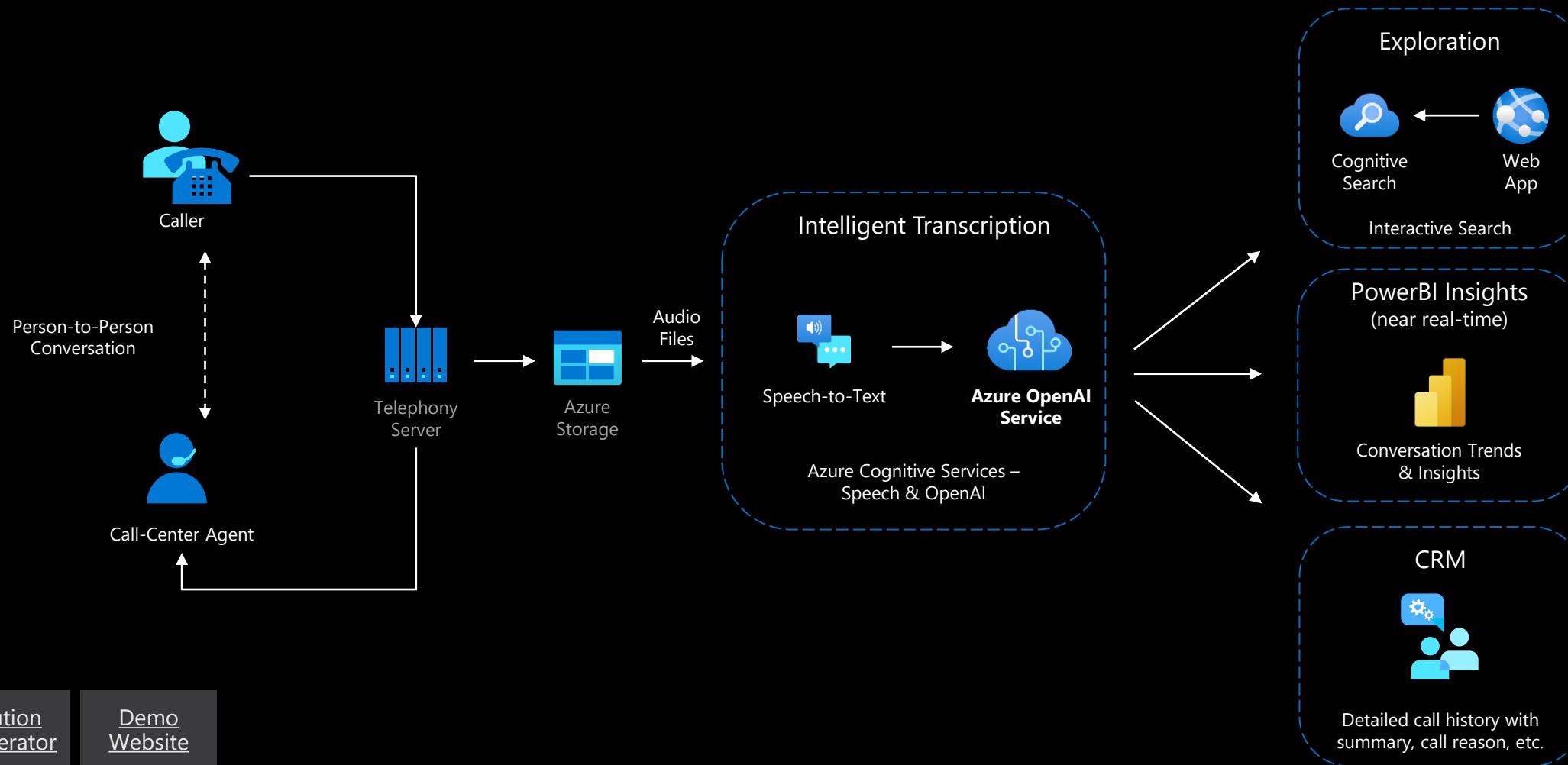
Thank you

Architecture design to build the PoC



Contact Center Analytics using Speech API & OpenAI

Extract rich insights from call transcripts

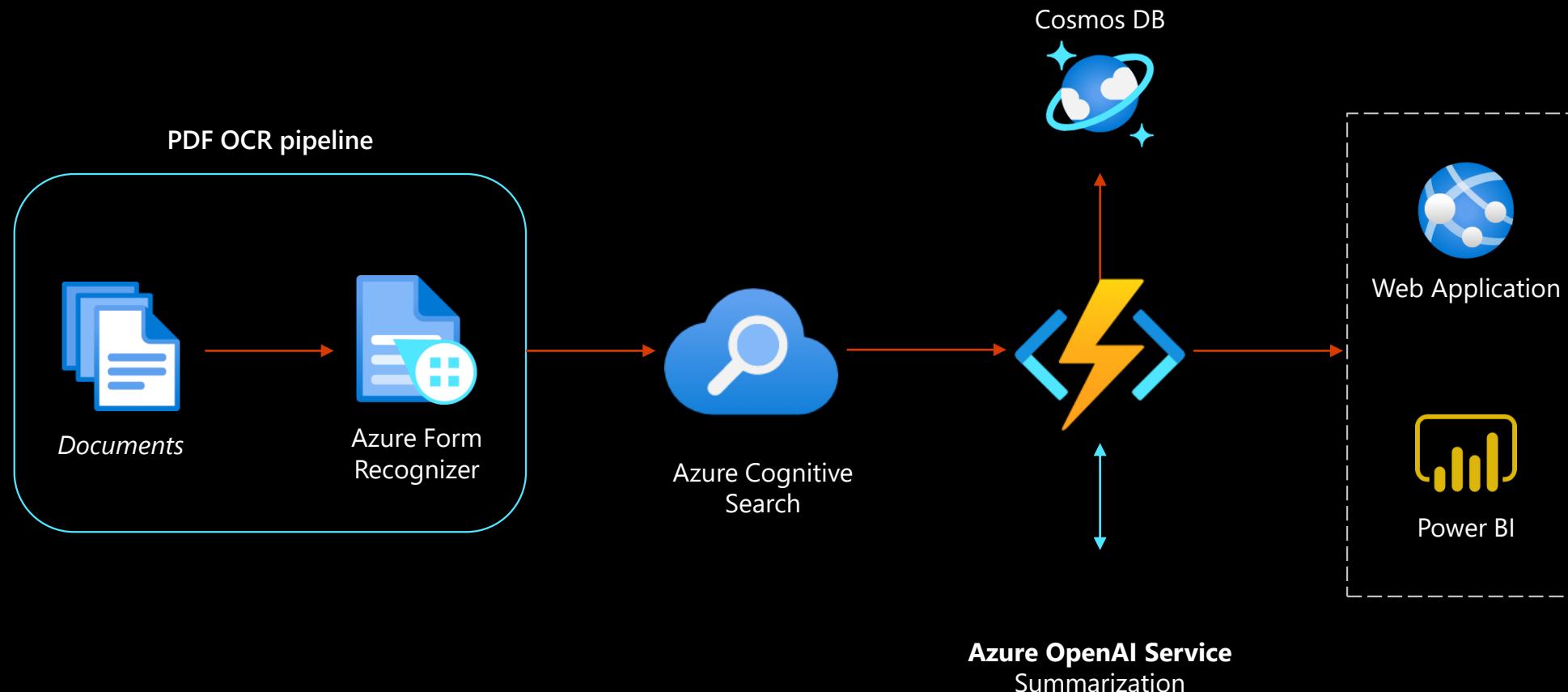


[Solution Accelerator](#)

[Demo Website](#)

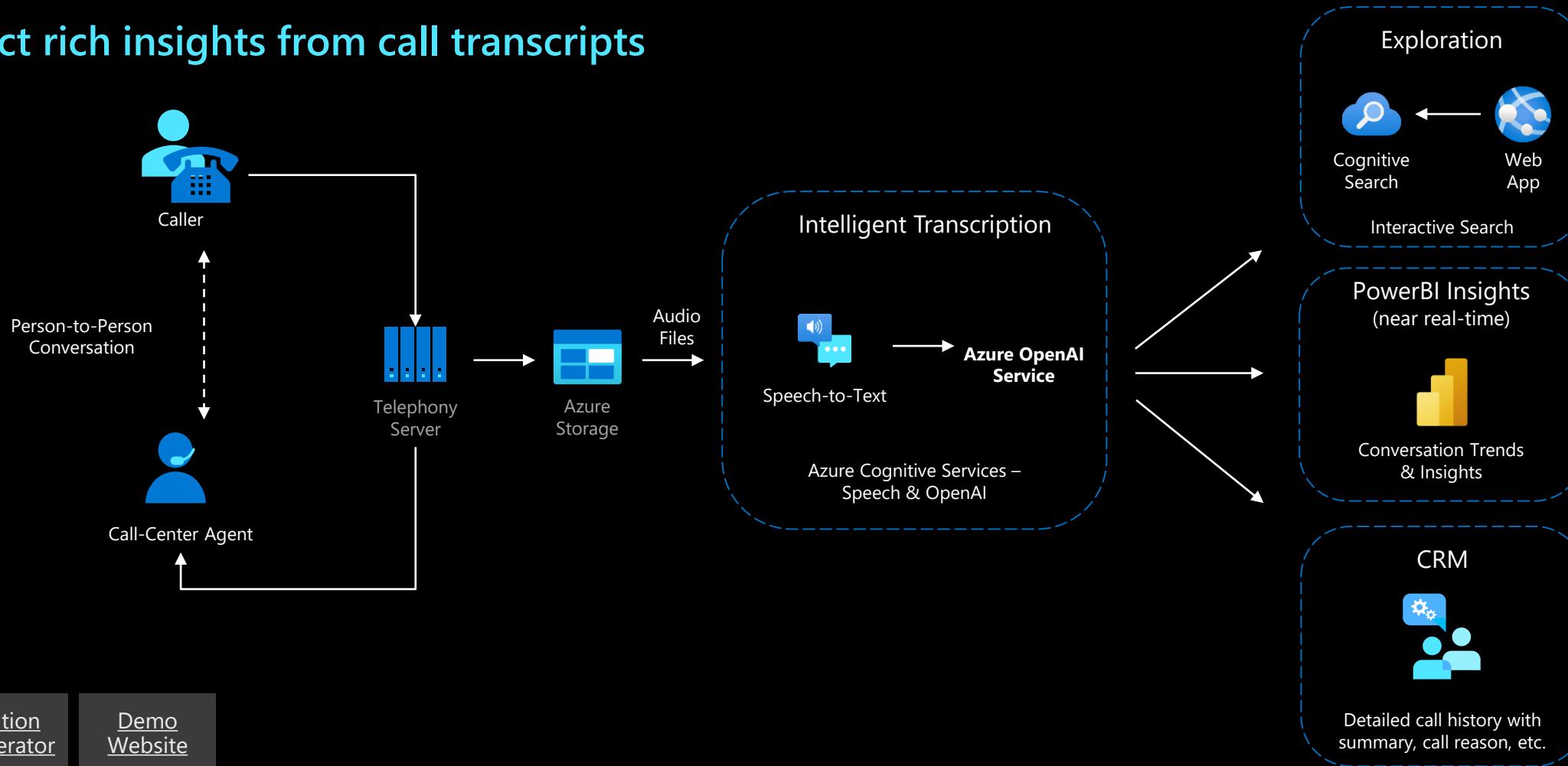
Azure OpenAI | GPT-3 sample high-level architecture

Document processing and summarization



Contact Center Analytics using Speech API & Azure OpenAI

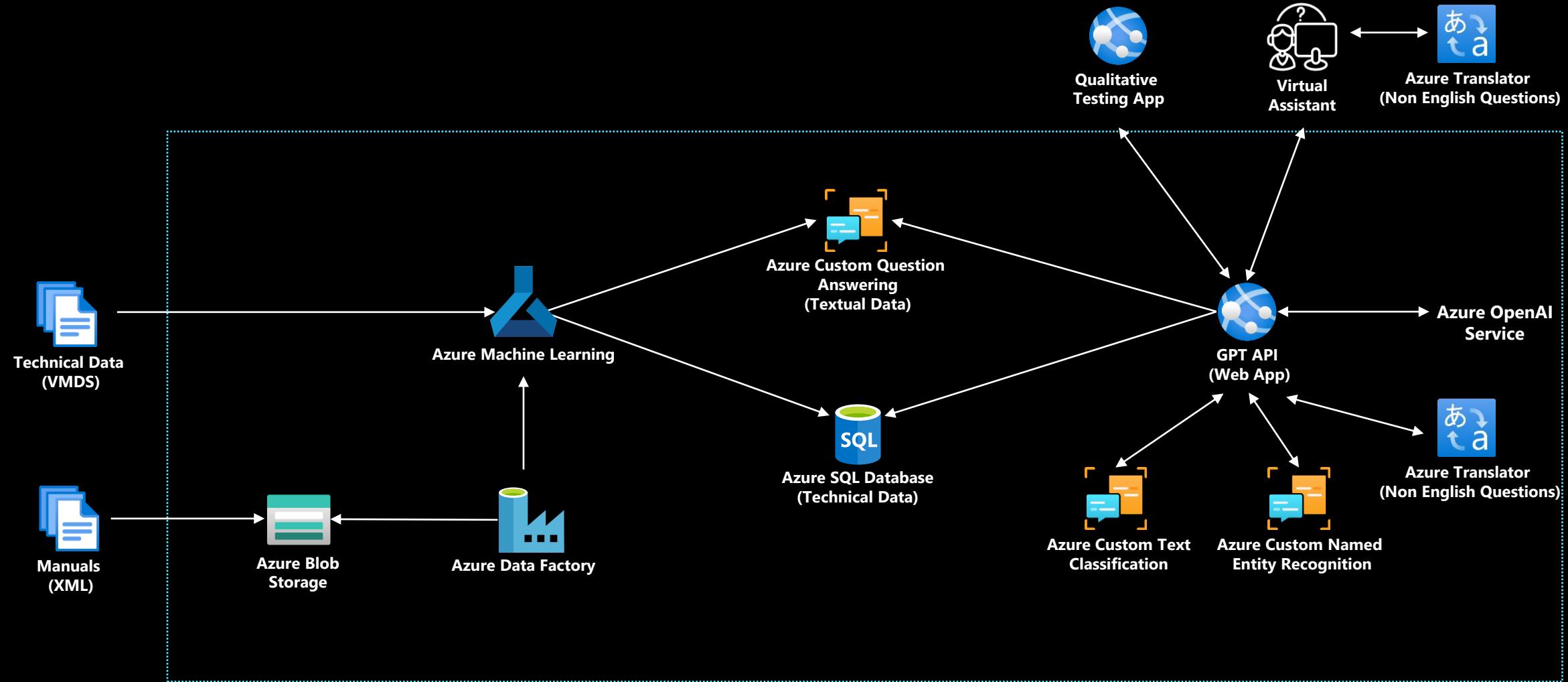
Extract rich insights from call transcripts



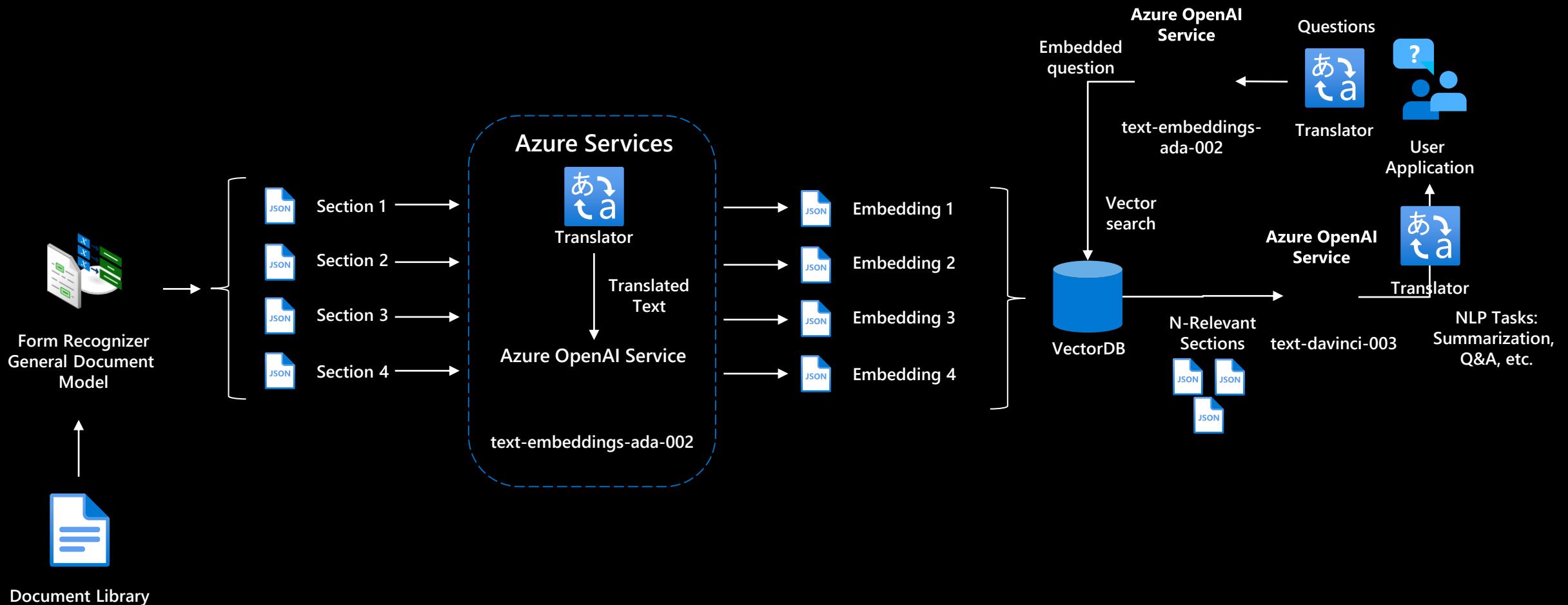
[Solution Accelerator](#)

[Demo Website](#)

Azure OpenAI Virtual Assistant architecture example



Document Embedding with Translation



Safety Execution Workflow

□ People & Policy

