

# Princeton A.I. Sandbox

# Spyglass MTG Overview

## Data, AI, and Security on the Microsoft Clouds

*Spyglass is a Specialized Expert Consulting firm focused on helping clients select, architect, implement, migrate to and manage their Microsoft Technology.*

- **A Women Owned/Women Led Company**
- **Headquartered in Lincoln, RI**
- **10+ years providing Microsoft Solutions to US clients**
- **25+ years average experience on our team**
- **5 Solutions Badges**
- **10 Advanced Specializations**
- **90% of our business is repeat business**
- **80% of our new business comes from Microsoft**



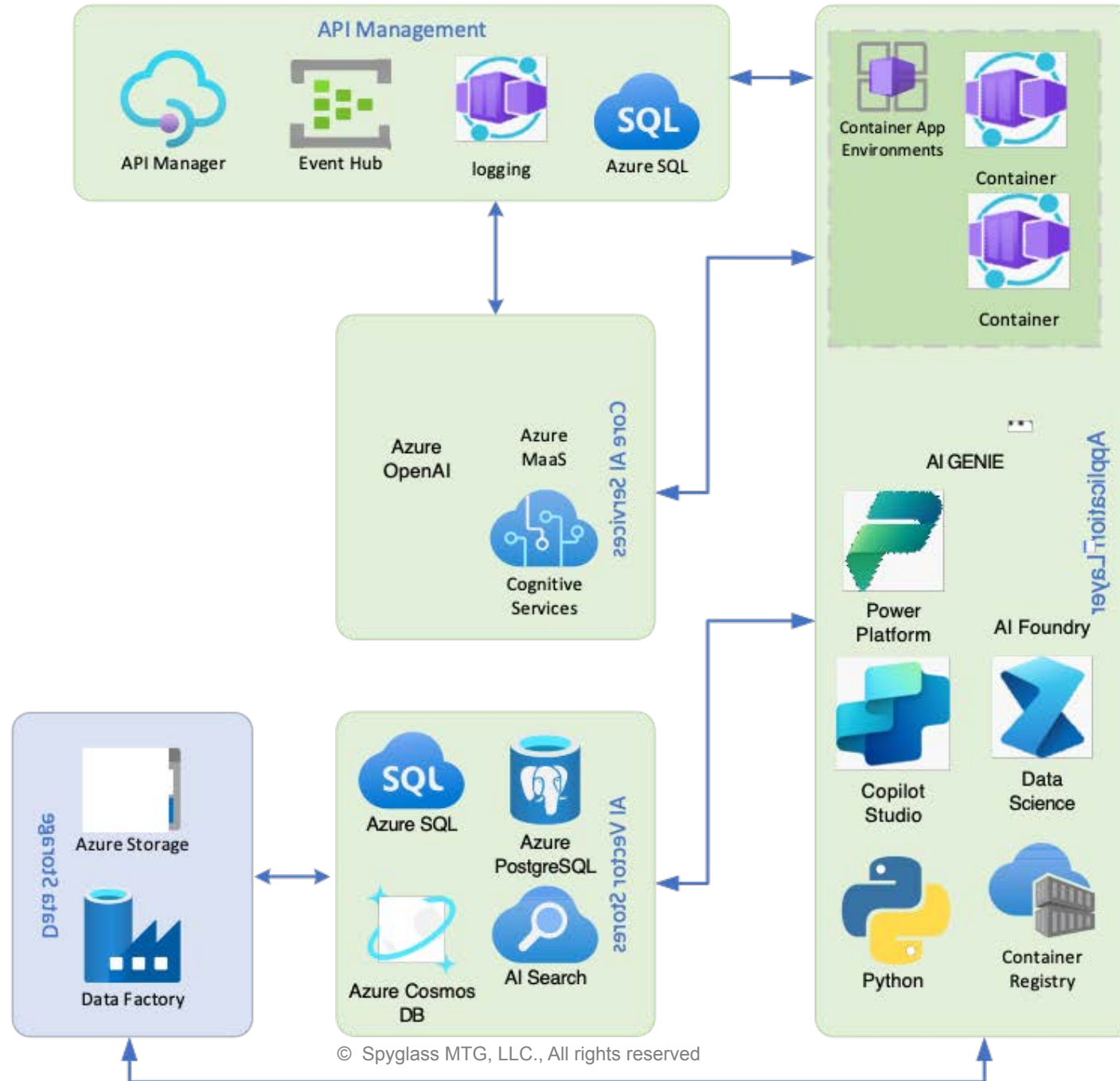
# Agenda

- AI Sandbox overview
- Models
- RAG Primer
- RAG Techniques
  - Advanced RAG
  - Prompt Stuffing
  - GraphRAG
- Document Intelligence
- LLMOps
- Useful Links



# AI Sandbox

# AI Sandbox



# Hello, Rudy Sandoval

How can I help you today?

Type a new question...



Chat ✕

GPT 4o ✎

Assistants ⋮

# Hello, Rudy Sandoval

How can I help you today?

Type a new question...

### Model Settings

Select model \*

- GPT 4o
- GPT 4o mini
- GPT 4o**
- Llama 3.3 70B Instruct
- Mistral Large
- DeepSeek R1
- GPT o1 Mini

Instructions Default Answer Summarize Editor

You are an AI assistant that helps people find information.

Save Cancel

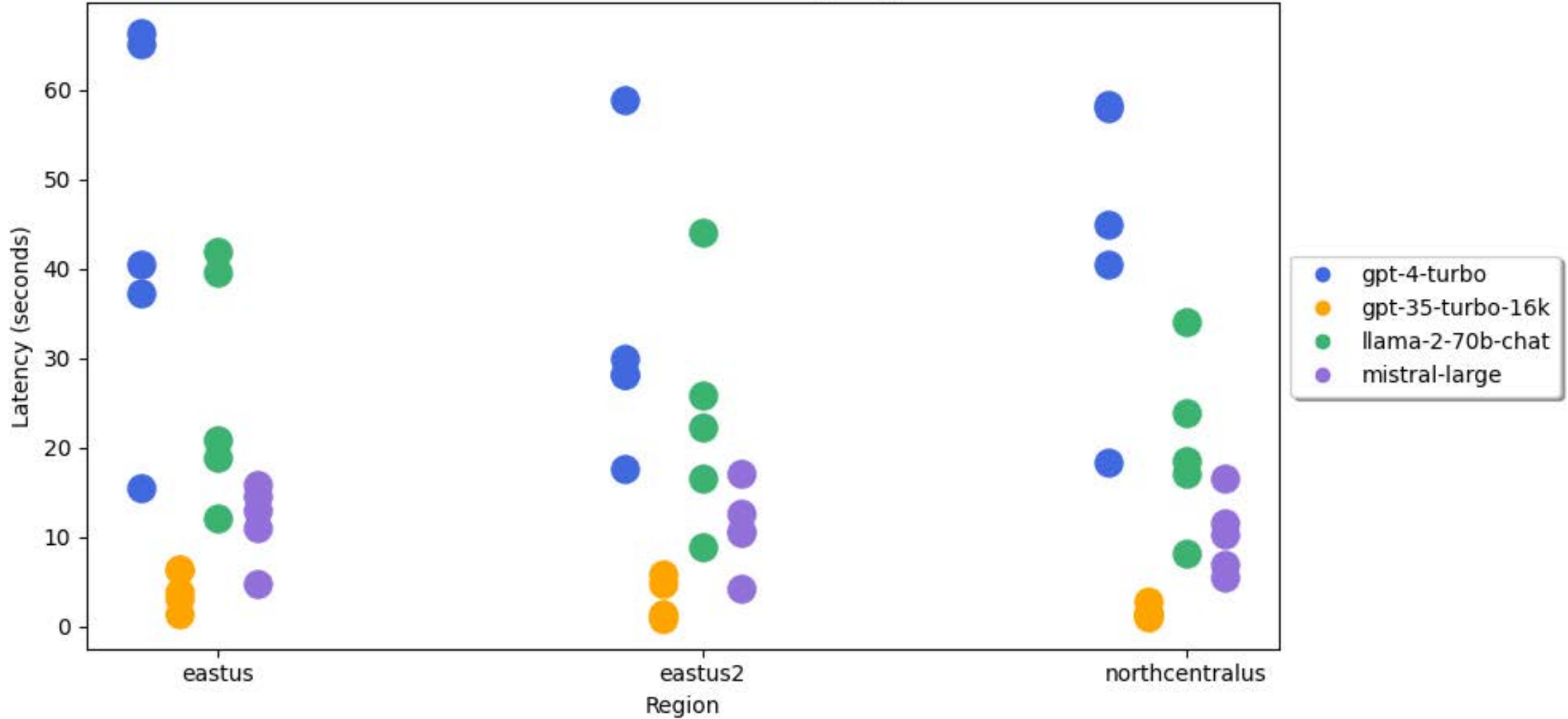
### Chat

- GPT 4o

### Assistants

# Latency Graphs

LLM Model Inference Times by Region





# Models Available

🏠 Collections ▾ 🏠 Industry ▾ 🏠 Capabilities ▾ ▶ Deployment options ▾ 🏠 Inference tasks ▾ 🏠 Fine-tuning tasks ▾ 🏠 Licenses ▾ 🔍 Compare models

🔍 Search Models 1861

<b>o3-mini</b> Chat completion	<b>DeepSeek-R1</b> Chat completion	<b>Phi-4</b> Chat completion	<b>gpt-4o-mini-audio-preview</b> Audio generation	<b>gpt-4o-mini-realtime-preview</b> Audio generation	<b>o1</b> Chat completion	<b>o1-mini</b> Chat completion
<b>gpt-4o</b> Chat completion	<b>gpt-4o-mini</b> Chat completion	<b>gpt-4o-audio-preview</b> Audio generation	<b>gpt-4o-realtime-preview</b> Audio generation	<b>Muse</b> Image to image	<b>Cohere-rerank-v3.5</b> Text classification	<b>Stable-Diffusion-3.5-Large</b> Text to image, Image to image
<b>Stable-Image-Ultra</b> Text to image	<b>Stable-Image-Core</b> Text to image	<b>Gretel-Navigator-Tabular</b> Chat completion, Data generation	<b>o1-preview</b> Chat completion	<b>Llama-3.3-70B-Instruct</b> Chat completion	<b>tsuzumi-7b</b> Chat completion	<b>Bria-2.3-Fast</b> Text to image
<b>Ministral-3B</b> Chat completion	<b>Prism</b> Zero-shot image classification	<b>Virchow</b> image-feature-extraction	<b>Virchow2</b> image-feature-extraction	<b>Cohere-embed-v3-english</b> Embeddings	<b>Cohere-embed-v3-multi...</b> Embeddings	<b>Llama-3.2-11B-Vision-Ins...</b> Chat completion
<b>gpt-4</b> Chat completion	<b>gpt-4-32k</b> Chat completion	<b>Phi-3-small-8k-instruct</b> Chat completion	<b>AI21-Jamba-1.5-Mini</b> Chat completion	<b>AI21-Jamba-1.5-Large</b> Chat completion	<b>Cohere-command-r-08...</b> Chat completion	<b>Cohere-command-r-plus...</b> Chat completion
<b>Cohere-rerank-v3-english</b> Text classification	<b>Cohere-rerank-v3-multiling...</b> Text classification	<b>text-embedding-3-large</b> Embeddings	<b>text-embedding-3-small</b> Embeddings	<b>tts</b> Text to speech	<b>tts-hd</b> Text to speech	<b>whisper</b> Speech recognition
<b>Deci-DeciLM-7B-instruct</b> Text generation	<b>Deci-DeciLM-7B</b> Text generation	<b>Deci-DeciCoder-1b</b> Text generation	<b>snowflake-arctic-base</b> Text generation	<b>dall-e-3</b> Text to image	<b>dall-e-2</b> Text to image	<b>text-embedding-ada-002</b> Embeddings
<b>davinci-002</b> Completions						

# Serverless Models

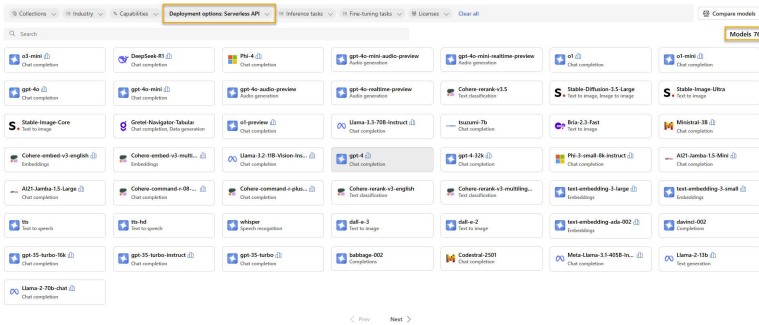
☰ Collections ▾ ☰ Industry ▾ ✨ Capabilities ▾ **Deployment options: Serverless API** ▾ ☰ Inference tasks ▾ ☰ Fine-tuning tasks ▾ 🏷 Licenses ▾ Clear all 🔍 Compare models

🔍 Search **Models 76**

<b>o3-mini</b> Chat completion	<b>DeepSeek-R1</b> Chat completion	<b>Phi-4</b> Chat completion	<b>gpt-4o-mini-audio-preview</b> Audio generation	<b>gpt-4o-mini-realtime-preview</b> Audio generation	<b>o1</b> Chat completion	<b>o1-mini</b> Chat completion
<b>gpt-4o</b> Chat completion	<b>gpt-4o-mini</b> Chat completion	<b>gpt-4o-audio-preview</b> Audio generation	<b>gpt-4o-realtime-preview</b> Audio generation	<b>Cohere-rerank-v3.5</b> Text classification	<b>Stable-Diffusion-3.5-Large</b> Text to image, image to image	<b>Stable-Image-Ultra</b> Text to image
<b>Stable-Image-Core</b> Text to image	<b>Gretel-Navigator-Tabular</b> Chat completion, Data generation	<b>o1-preview</b> Chat completion	<b>Llama-3.3-70B-Instruct</b> Chat completion	<b>tsuzumi-7b</b> Chat completion	<b>Bria-2.3-Fast</b> Text to image	<b>Ministral-3B</b> Chat completion
<b>Cohere-embed-v3-english</b> Embeddings	<b>Cohere-embed-v3-multi-lingual</b> Embeddings	<b>Llama-3.2-11B-Vision-Instruct</b> Chat completion	<b>gpt-4</b> Chat completion	<b>gpt-4-32k</b> Chat completion	<b>Phi-3-small-8k-instruct</b> Chat completion	<b>AI21-Jamba-1.5-Mini</b> Chat completion
<b>AI21-Jamba-1.5-Large</b> Chat completion	<b>Cohere-command-r-08</b> Chat completion	<b>Cohere-command-r-plus</b> Chat completion	<b>Cohere-rerank-v3-english</b> Text classification	<b>Cohere-rerank-v3-multilingual</b> Text classification	<b>text-embedding-3-large</b> Embeddings	<b>text-embedding-3-small</b> Embeddings
<b>tts</b> Text to speech	<b>tts-hd</b> Text to speech	<b>whisper</b> Speech recognition	<b>dall-e-3</b> Text to image	<b>dall-e-2</b> Text to image	<b>text-embedding-ada-002</b> Embeddings	<b>davinci-002</b> Completions
<b>gpt-35-turbo-16k</b> Chat completion	<b>gpt-35-turbo-instruct</b> Chat completion	<b>gpt-35-turbo</b> Chat completion	<b>babbage-002</b> Completions	<b>Codestral-2501</b> Chat completion	<b>Meta-Llama-3.1-405B-Instruct</b> Chat completion	<b>Llama-2-13b</b> Text generation
<b>Llama-2-70b-chat</b> Chat completion						

< Prev Next >

# Serverless Models Cont.




















































<b>Llama-2-7b-chat</b> Chat completion	<b>Llama-2-7b</b> Text generation	<b>Mistral-large</b> Chat completion	<b>Phi-3-medium-4k-instruct</b> Chat completion	<b>jais-30b-chat</b> Chat completion	<b>Llama-2-70b</b> Text generation	<b>TimeGEN-1</b> Forecasting
<b>Mistral-small</b> Chat completion	<b>Phi-3.5-MoE-instruct</b> Chat completion	<b>Phi-3.5-vision-instruct</b> Chat completion	<b>Meta-Llama-3-8B-Instruct</b> Chat completion	<b>Meta-Llama-3.1-70B-Ins...</b> Chat completion	<b>Meta-Llama-3.1-8B-Instr...</b> Chat completion	<b>Phi-3-medium-128k-inst...</b> Chat completion
<b>Meta-Llama-3-70B-Instr...</b> Chat completion	<b>Phi-3-mini-128k-instruct</b> Chat completion	<b>Cohere-command-r-plus</b> Chat completion	<b>Phi-3-small-128k-instruct</b> Chat completion	<b>Phi-3-mini-4k-instruct</b> Chat completion	<b>Mistral-large-2407</b> Chat completion	<b>Mistral-Nemo</b> Chat completion
<b>Llama-3.2-90B-Vision-In...</b> Chat completion	<b>Llama-2-13b-chat</b> Chat completion	<b>Phi-3.5-mini-instruct</b> Chat completion	<b>Cohere-command-r</b> Chat completion	<b>Mistral-Large-2411</b> Chat completion		

# Managed Compute

Collections: Hugging Face ▾ Industry ▾ Capabilities ▾ Deployment options: Managed compute ▾ Inference tasks ▾ Fine-tuning tasks ▾ Licenses ▾ Clear all Compare models

Search Models 1620

 <b>huggingface4-zephyr-7b-...</b> Text generation	 <b>vagosolutions-llama-3.1-sau...</b> Text generation	 <b>freedomintelligence-aceppt...</b> Text generation	 <b>bin12345-autocoder</b> Text generation	 <b>alibaba-nlp-gte-large-en-v1.5</b> sentence-similarity	 <b>sonnenblume-bert-base-un...</b> Fill mask	 <b>helsinki-nlp-opus-mt-fr-ru</b> Translation
 <b>naver-efficient-splade-v-lar...</b> Fill mask	 <b>oliverguhr-fullstop-punctua...</b> Token classification	 <b>moussakam-arabart</b> Fill mask	 <b>deepset-gbert-base</b> Fill mask	 <b>cointegrated-rut5-base</b> Text to text generation	 <b>m3hrdadfi-typo-detector-di...</b> Token classification	 <b>hyunwoongko-asian-bart-e...</b> Text to text generation
 <b>staka-fugumt-ja-en</b> Translation	 <b>activebus-bert-review</b> Fill mask	 <b>toddgoldfarb-cadet-tiny</b> Conversational	 <b>facebook-blenderbot-1b-dis...</b> Conversational	 <b>monologg-kobigbird-bert-...</b> Fill mask	 <b>persiannlp-mt5-large-parsi...</b> Text to text generation	 <b>hf-internal-testing-tiny-ran...</b> Question answering
 <b>yoshitomo-matsubara-bert-...</b> Text classification	 <b>allenai-tk-instruct-base-def...</b> Text to text generation	 <b>vinai-phobert-base-v2</b> Fill mask	 <b>asi-gpt-fr-cased-base</b> Text generation	 <b>castorini-doc2query-t5-bas...</b> Text to text generation	 <b>aidenh20-dnabert-500down</b> Text classification	 <b>hf-internal-testing-tiny-ran...</b> Fill mask
 <b>microsoft-tapex-base</b> Table question answering	 <b>google-pegasus-newsroom</b> Summarization	 <b>andreaskoepf-pythia-1.4b-g...</b> Text generation	 <b>moritzlaurer-deberta-v3-ba...</b> Text classification	 <b>ckiplab-albert-tiny-chinese</b> Fill mask	 <b>aubmindlab-bert-base-arab...</b> Fill mask	 <b>togethercomputer-gpt-jt-m...</b> Text generation
 <b>microsoft-dialogrpt-width</b> Text classification	 <b>embeddia-crosloengual-bert</b> Fill mask	 <b>hf-internal-testing-tiny-ran...</b> Fill mask	 <b>nchunlp-chinese-question-a...</b> Question answering	 <b>alirezamsh-small100</b> Translation	 <b>flax-community-t5-recipe-g...</b> Text to text generation	 <b>monohime-rubert-base-cas...</b> Text classification
 <b>yiyanghust-finbert-fls</b> Text classification	 <b>luodian-llama-7b-hf</b> Text generation	 <b>mariagrاندury-roberta-bas...</b> Text classification	 <b>eleutherai-pythia-160m-ded...</b> Text generation	 <b>smanjil-german-medbert</b> Fill mask	 <b>allenai-unifiedqa-v2-t5-larg...</b> Text to text generation	 <b>ingen51-dialogpt-medium-...</b> Conversational

# Accelerate innovation with Azure AI

Hosted Models and services to accelerate teams

## AI services



Bot Service



AI Search



Document Intelligence



Video Indexer



Metrics Advisor



Immersive Reader



Vision



Speech



Language

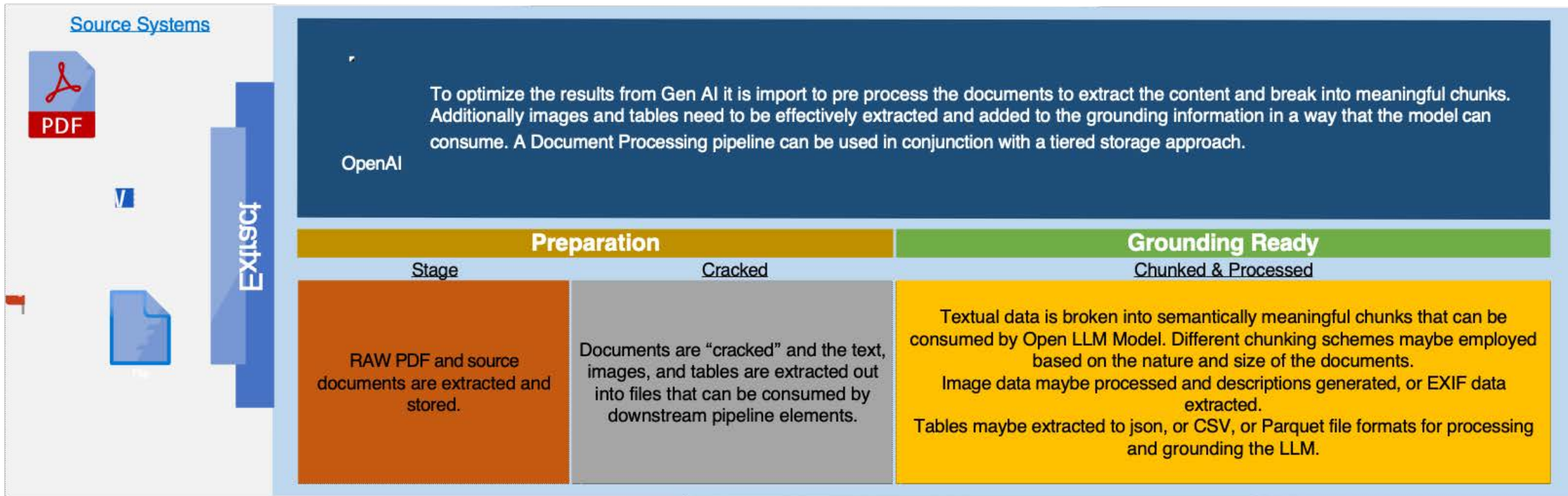


Decision

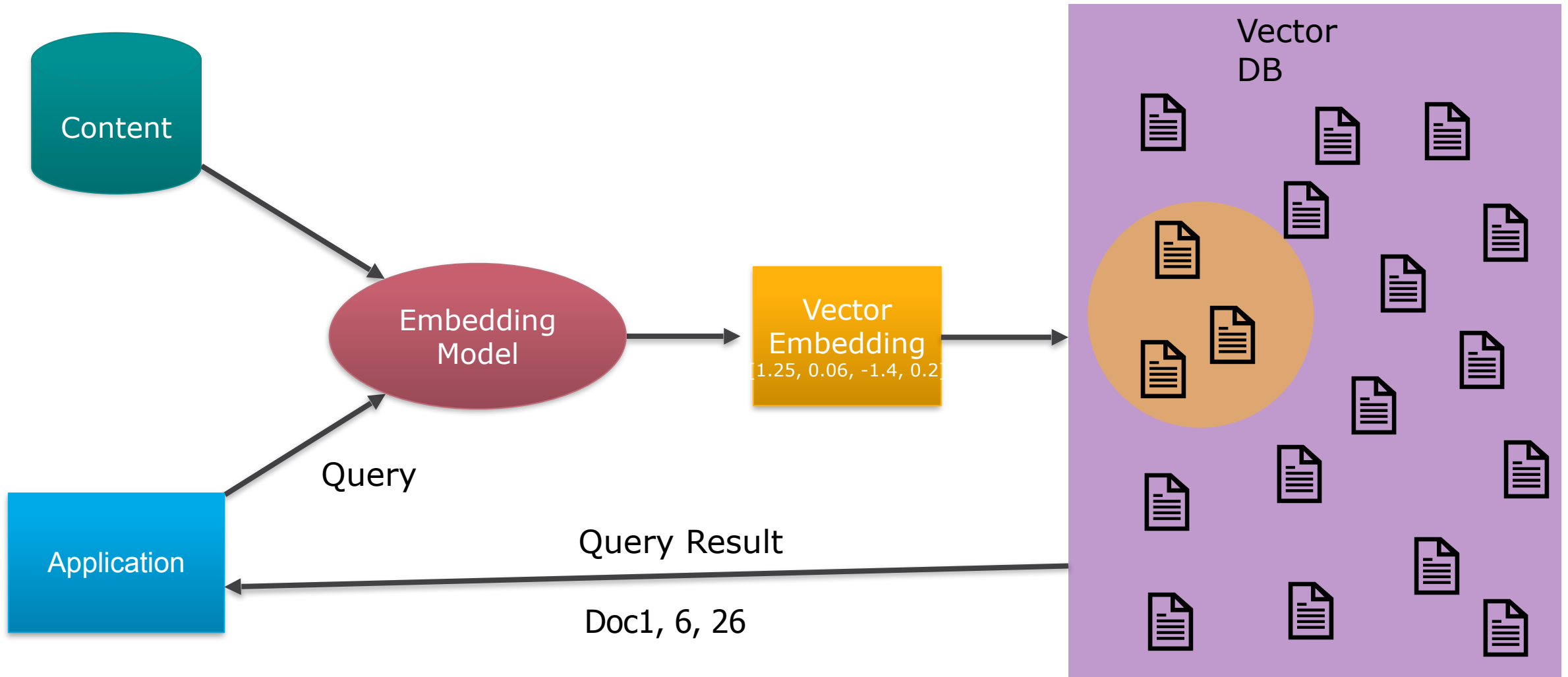


OpenAI Service

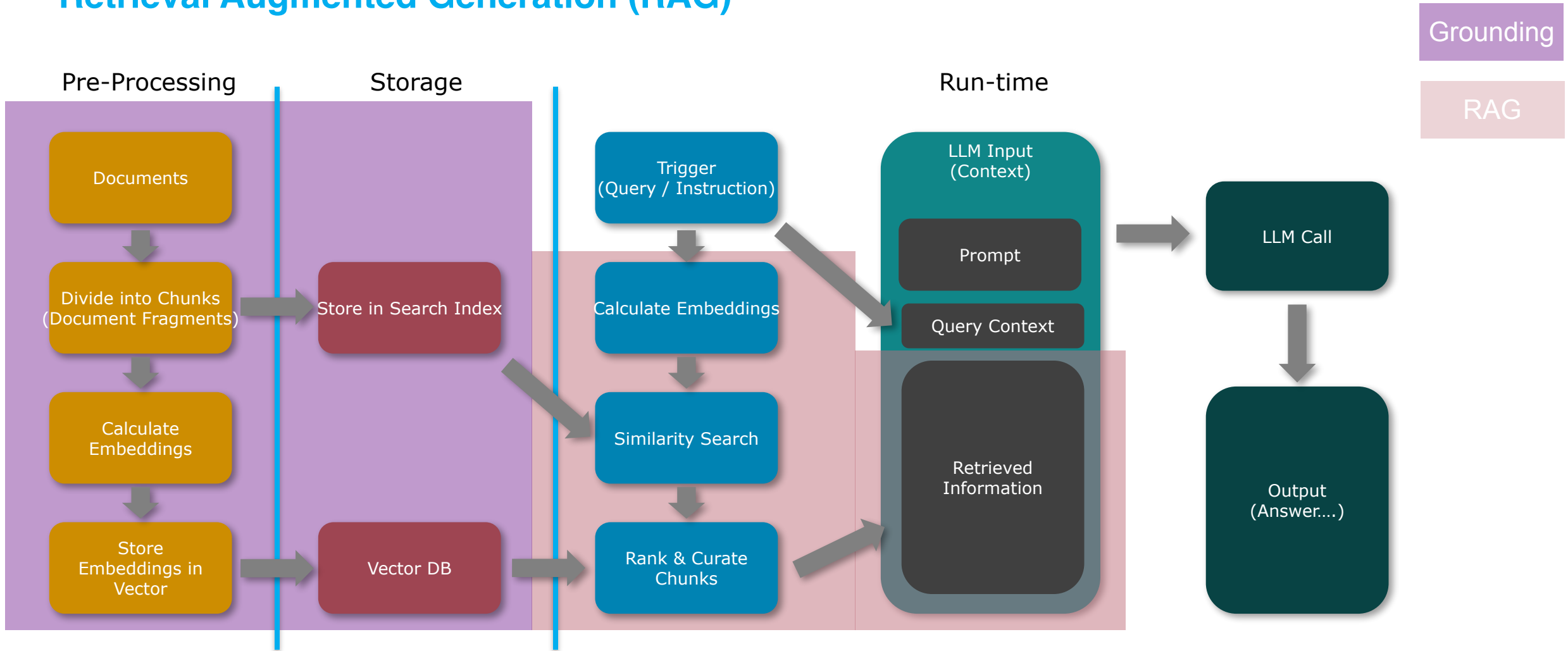
# Grounding Process Detail



# RAG Primer: Vector DB

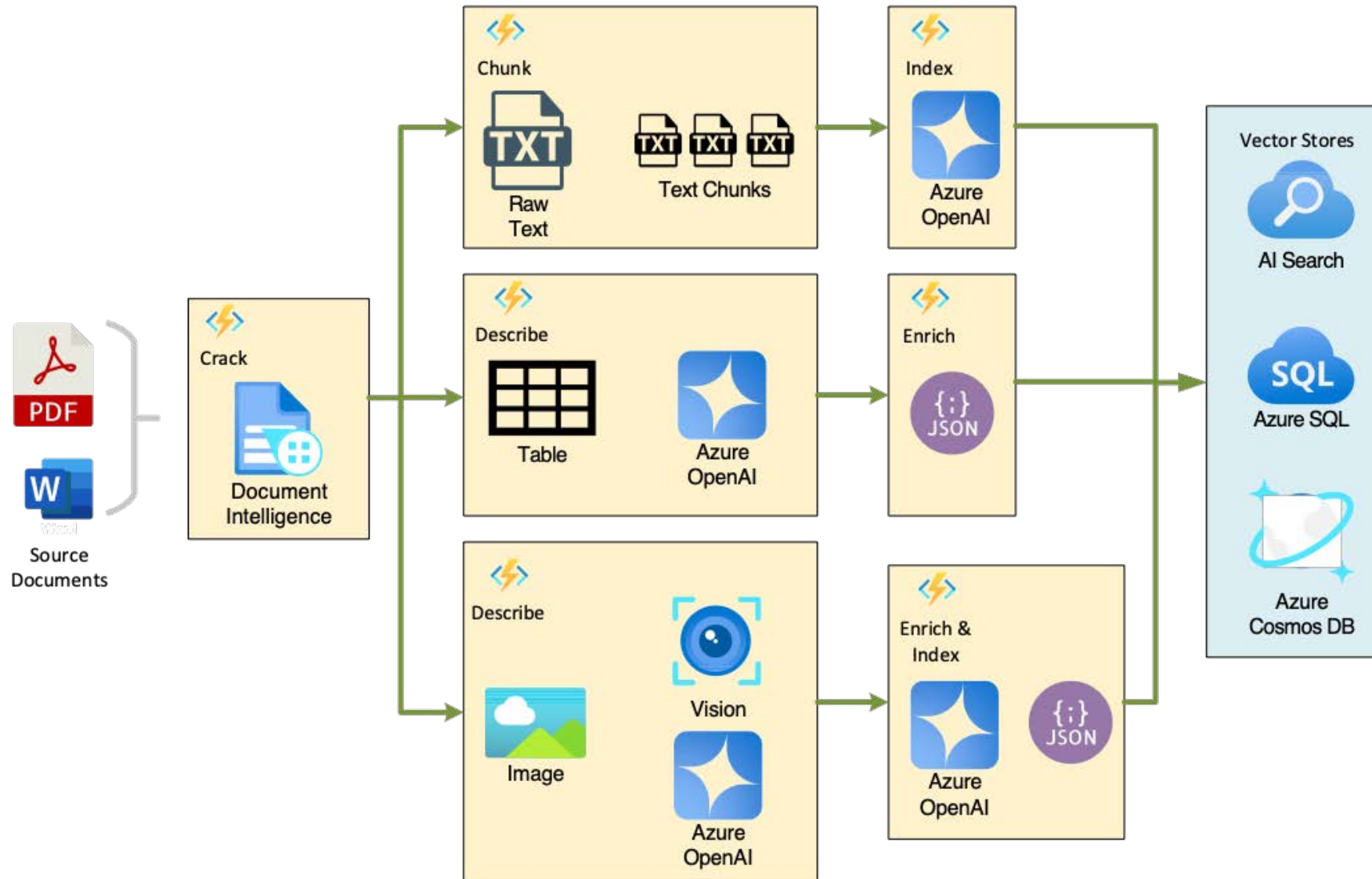


# Using Your Own Data: “Grounding” & Retrieval Augmented Generation (RAG)





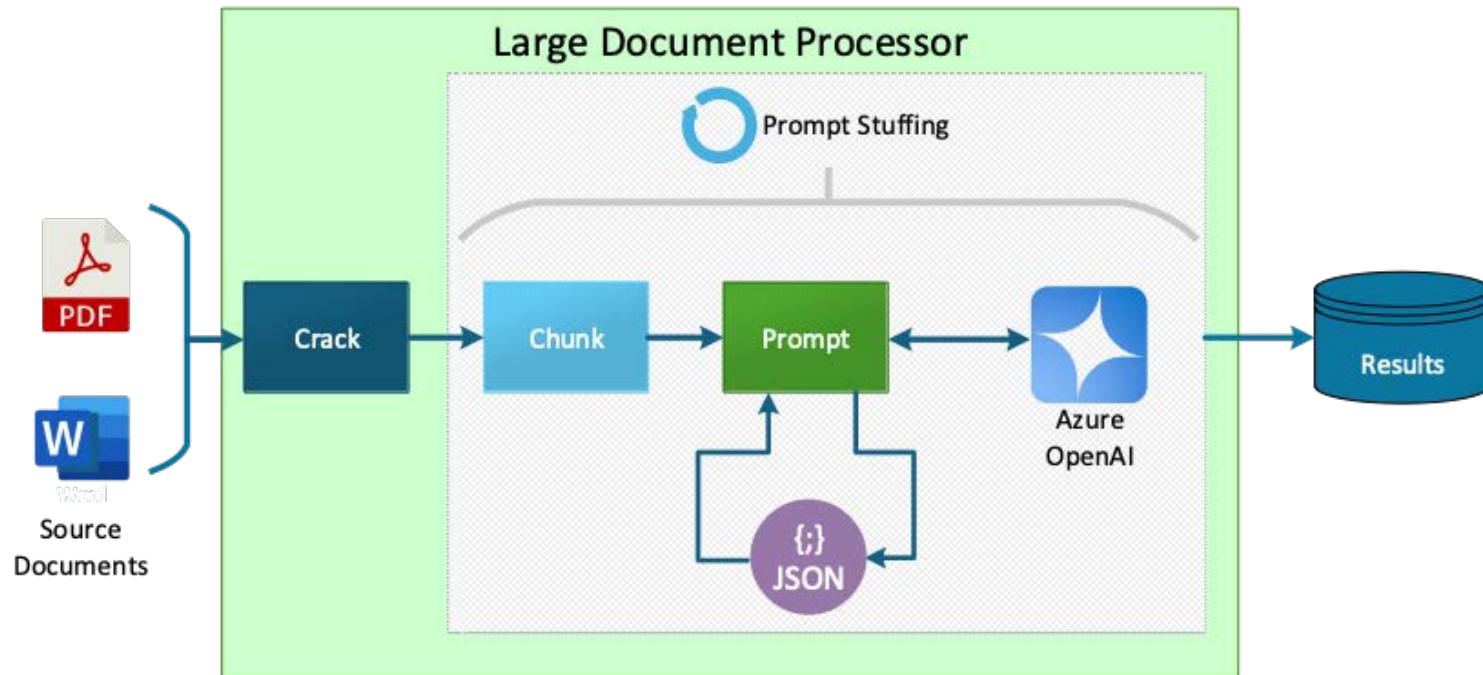
# Advanced RAG



# Limitations of RAG for Large, Dense Documents

- **Chunking constraints:** Documents must be split into chunks, and the system is typically configured to select how many chunks to retrieve. Critical information might be left out if the system is forced to pick only a few chunks.
- **Loss of document integrity:** Splitting a document into disjointed pieces can disrupt logical flow, especially if concepts span multiple sections.
- **Context window exhaustion:** If most of the document is relevant—for instance, 95% of a 300,000-token text—retrieving all of it may exceed the LLM’s context limit .
- **Performance degradation with increased context length:** While GPT-4o has a context window of 128K tokens, studies have shown that LLMs often experience a decline in reasoning performance when processing inputs that approach or exceed approximately 50% of their maximum context length. This suggests that for GPT-4o, performance issues might arise with inputs around 64K tokens.

# Prompt Stuffing



## How It Works

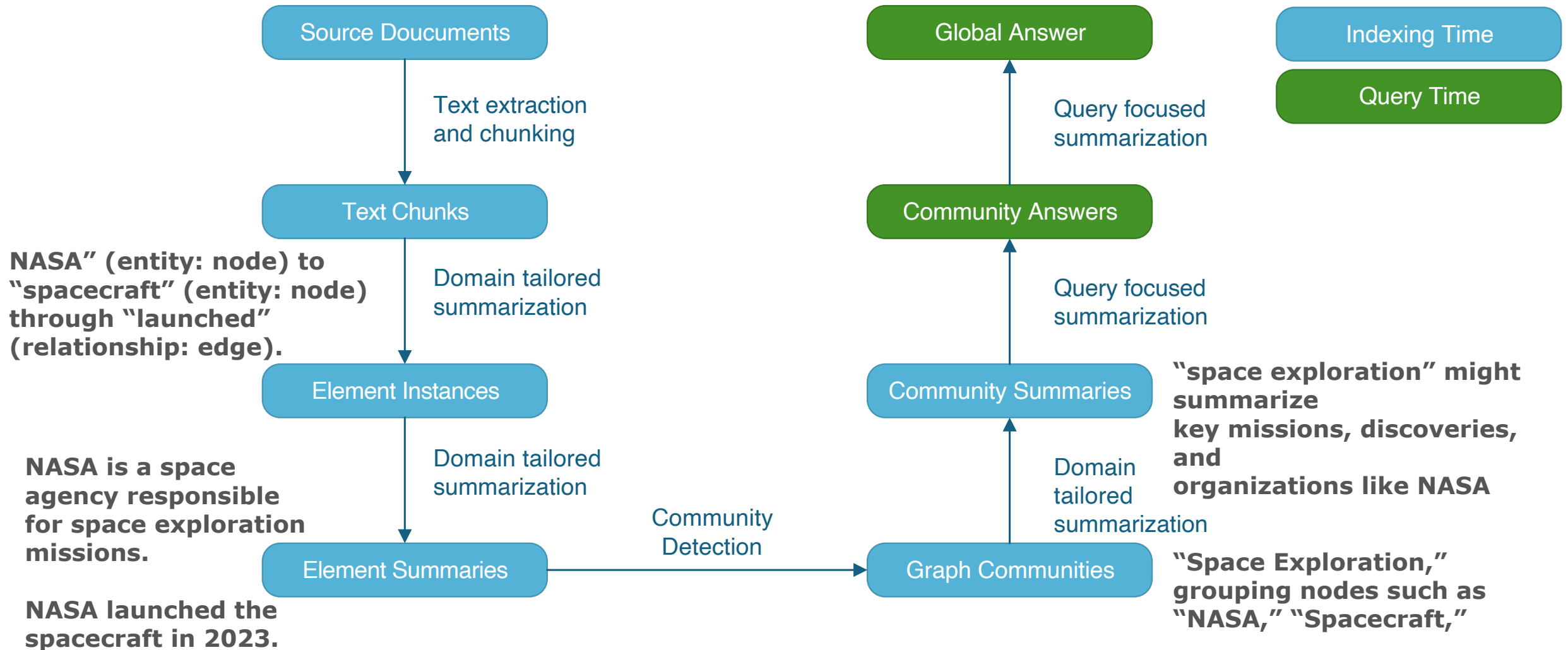
- 1. Chunk the document dynamically:** Instead of retrieving small pieces, you select sizable fragments (e.g., 40,000–60,000 tokens based on the exact LLM being used) and feed them to the LLM in sequence along with the prompt.
- 2. Produce structured JSON summaries:** For each segment, the model outputs a JSON object containing key points, references (like page numbers or section titles), and any critical facts.
- 3. Iteratively build the context:** The JSON summary from each step is reintroduced alongside the next text segment, creating a chain of prompts where new content is processed in the light of previously extracted data.
- 4. Generate compressed, information-rich output:** By continuously condensing key insights, the final JSON structure captures the entire document without exceeding the LLM's context limit.

# Advantages Over Traditional RAG

- Full-document coverage: Every section is processed in turn, ensuring no portion is inadvertently excluded.
- No dependency on external retrieval: You do not need to maintain a vector database or indexing system.
- Higher accuracy: Since all content is examined, crucial details aren't missed due to the configured chunk selection limit.
- Scalability: Large documents can be handled in segments, limited only by the size of the final JSON output rather than the LLM's prompt window.

# GraphRAG

# GraphRAG



## Graph RAG

- GraphRAG is good when **useful insight spans many documents**
- GraphRAG is capable of providing high degrees of insight on **complex information topics**
- GraphRAG is intended to be deployed and used with a domain specific corpus of text data.
- GraphRAG performs best on materials that is **entity rich** – entities being people, places, things, or objects that can be uniquely identified.

# Graph RAG vs Basic RAG

what risks are you seeing in the commercial real estate industry, which companies are most impacted

## GraphRAG (GPT-4o mini)

### Overview of Risks in the Commercial Real Estate Industry

The commercial real estate industry is currently facing a multitude of significant risks that are reshaping its landscape. These risks are primarily driven by economic factors such as rising interest rates, inflation, and changing consumer behaviors, which may lead to decreased demand for commercial properties and increased vacancy rates.

### Key Risks Identified

- 1. Rising Interest Rates:** The most pressing concern is the impact of rising interest rates, which may lead to increased borrowing costs for companies involved in commercial real estate. This financial strain can reduce investment in properties and negatively affect property valuations, creating challenges for maintaining profitability [Data: Reports (554, 337, 763, 548, 430, +more)].
- 2. Economic Uncertainty:** High inflation and moderating labor market conditions are influencing consumer purchasing behavior, which in turn affects the overall economic performance. Companies like Walmart are adjusting their pricing strategies to navigate these challenges, indicating a broader impact on retail and commercial real estate dynamics [Data: Reports (774)].
- 3. Changing Market Dynamics:** The shift towards remote work and e-commerce has altered demand for traditional office spaces and retail locations. Companies heavily invested in these sectors may experience financial strain as occupancy rates decline and rental income diminishes [Data: Reports (490, 786, 693, 295, 733)].
- 4. Liquidity Challenges:** The industry is also grappling with liquidity challenges, prompting companies to reconsider their cash balances and investment strategies. This situation is particularly complex for those with high debt loads and limited access to capital markets.

## Baseline RAG (GPT-4o mini)

The documents provided do not contain specific information about the risks in the commercial real estate industry or which companies are most impacted. Therefore, I cannot provide an answer based on the documents.

### Companies Most Impacted

Several companies are particularly vulnerable to these risks:

- **Wells Fargo:** As a major player in the commercial real estate sector, Wells Fargo is closely monitoring increased charge-offs and a decline in revenue and loan balances, which raises concerns about its stability and growth potential [Data: Reports (537)].
- **JPMorgan Chase and Truist Securities:** These companies may experience impacts on their asset management and investment banking divisions due to market fluctuations and changing consumer behaviors, which are closely tied to the health of the commercial real estate sector [Data: Reports (548, 763)].
- **Real Estate Investment Trusts (REITs):** Companies heavily invested in commercial real estate, particularly those reliant on traditional office spaces, may face adverse effects from the shift towards remote work and changing tenant preferences, leading to higher vacancy rates and decreased rental income [Data: Reports (440, 502, 783, 549, 442)].

### Conclusion

In summary, the commercial real estate industry is navigating a complex landscape of risks primarily driven by economic fluctuations, rising interest rates, and changing consumer behaviors. Companies such as Wells Fargo, JPMorgan Chase, and various REITs are among those most impacted, facing challenges that could significantly affect their financial performance and market stability. Stakeholders must closely monitor these trends to adapt their strategies accordingly and mitigate potential risks.



# Graph RAG vs Basic RAG

what industries are impacted by issues with supply chain

## GraphRAG (GPT-4o mini)

### Overview of Industries Impacted by Supply Chain Issues

Supply chain disruptions have become a significant concern across various industries, affecting their operational efficiency, profitability, and market competitiveness. Below is a summary of the key industries impacted by these challenges:

#### Technology and Semiconductor Industry

The technology sector, particularly the semiconductor industry, is heavily affected by supply chain issues. Disruptions in the production and distribution of essential components, such as semiconductors, can lead to delays in product launches and increased costs for companies like TSMC and AMD. The reliance on advanced packaging technologies, such as CoWoS, further emphasizes the need for efficient supply chains to meet customer demand and enhance capacity [Data: Reports (352, 691, 592, 593, 598, +more)].

#### Automotive Industry

The automotive industry faces substantial challenges due to supply chain disruptions, particularly concerning the availability of critical parts and materials. This can result in production halts and delays in vehicle deliveries, affecting overall sales and profitability. The shift towards electric vehicles adds another layer of complexity, as companies like Polestar and those operating in the Ghent Facility must navigate operational challenges and production difficulties [Data: Reports (391, 59, 402, 691, 555, +more)].

#### Manufacturing Sector

The manufacturing industry is significantly impacted by supply chain issues, as disruptions can lead to delays in production and increased operational costs. The reliance on timely delivery of raw materials and components is crucial for maintaining efficiency and profitability [Data: Reports (456, 69, 334, 272, 800, +more)].

#### Retail Industry

The retail sector, including major players like Walmart and Costco, is also affected by supply chain challenges. Disruptions can lead to inventory

## Baseline RAG (GPT-4o mini)

The documents indicate that the automotive industry, specifically Volvo, is impacted by supply chain issues. Volvo has been dealing with increased tariffs on batteries and has been focusing on creating a more resilient supply chain to manage these challenges. Additionally, the beverage industry, as represented by Coca-Cola, is experiencing supply chain shortages, which have affected product availability and market share, particularly in Mexico and Brazil. The retail industry, specifically Walmart, has also been working on improving its supply chain, which has resulted in better quality and availability of fresh food products.

context: economic supply chain disruptions (774, 768, 775, 776, +more)].

#### Healthcare and Pharmaceutical Industry

The healthcare and pharmaceutical sectors are significantly impacted by supply chain issues, particularly regarding the availability of medications and medical supplies. Disruptions can lead to shortages of critical treatments, affecting patient care and market access. Companies like Novartis and those involved in medical technology must navigate these challenges to ensure compliance and operational success [Data: Reports (293, 784, 292, 377, +more)].

#### Conclusion

In summary, supply chain issues are pervasive across multiple industries, including technology, automotive, manufacturing, retail, and healthcare. Each sector faces unique challenges that require strategic planning and effective management to mitigate the impacts of these disruptions. Addressing supply chain dynamics will be crucial for maintaining competitiveness and ensuring operational efficiency in the face of ongoing global challenges.

# Document Intelligence

## Study guide for Exam AZ-900: Microsoft Azure Fundamentals

### Purpose of this document

This study guide should help you understand what to expect on the exam and includes a summary of the topics the exam might cover and links to additional resources. The information and materials in this document should help you focus your studies as you prepare for the exam.

#### Useful links

#### Description

[Review the skills measured as of October 28, 2022](#)

This list represents the skills measured AFTER the date provided. Study this list if you plan to take the exam AFTER that date.

[Review the skills measured prior](#)

Study this list of skills if you take your exam PRIOR to the date provided.

# Document Intelligence

## Study guide for Exam AZ-900: Microsoft Azure Fundamentals

### Purpose of this document

This study guide should help you understand what to expect on the exam and includes a summary of the topics the exam might cover and links to additional resources. The information and materials in this document should help you focus your studies as you prepare for the exam.

Useful links	Description
<a href="#">Review the skills measured as of October 26, 2022</a>	This list represents the skills measured AFTER the date provided. Study this list if you plan to take the exam AFTER that date.
<a href="#">Review the skills measured prior</a>	Study this list of skills if you take your exam PRIOR to the date provided.

```
File Edit Selection View Go Run Terminal Help Search
{} AZ-900.json X
C: > Users > IanDicker > AppData > Local > Microsoft > Windows > INetCache > Content.Outlook > 9CMPYWI > {} AZ-900.json > ...
1  {
2    "apiVersion": "2024-07-31-preview",
3    "modelId": "prebuilt-layout",
4    "stringIndexType": "textElements",
5    "content": "Exam AZ-900: Microsoft Azure Fundamentals\nStudy guide for Exam AZ-900: Microsoft Azure Fundamentals\nPurpose of this document\nThis study guide should help you u
6    "pages": [
7      {
8        "pageNumber": 1,
9        "angle": 0.027639290317893028,
10       "width": 8.5,
11       "height": 11,
12       "unit": "inch",
13       "words": [
14         {
15           "content": "Exam",
16           "polygon": [
17             4.7665,
18             0.5132,
19             5.1342,
20             0.5129,
21             5.1341,
22             0.6852,
23             4.7662,
24             0.6834
25           ],
26           "confidence": 0.992,
27           "span": {
28             "offset": 0,
29             "length": 4
30           }
31         },
32         {
33           "content": "AZ-900:",
34           "polygon": [
35             5.1716,
36             0.5129,
37             5.7058,
38             0.5112,
39             5.7057,
40             0.6864,
41             5.1715,
42             0.6851
```

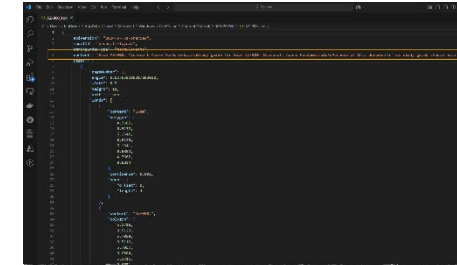
# Document Intelligence

## Study guide for Exam AZ-900: Microsoft Azure Fundamentals

### Purpose of this document

This study guide should help you understand what to expect on the exam and includes a summary of the topics the exam might cover and links to additional resources. The information and materials in this document should help you focus your studies as you prepare for the exam.

Useful links	Description
<a href="#">Review the skills measured as of October 26, 2022</a>	This list represents the skills measured AFTER the date provided. Study this list if you plan to take the exam AFTER that date.
<a href="#">Review the skills measured prior</a>	Study this list of skills if you take your exam PRIOR to the date provided.



```
4424 ],
4425 "lines": [
4426 {
4427   "content": "Exam AZ-900: Microsoft Azure Fundamentals",
4428   "polygon": [
4429     4.7664,
4430     0.5099,
4431     7.7596,
4432     0.5131,
4433     7.7594,
4434     0.6886,
4435     4.7662,
4436     0.6853
4437   ],
4438   "spans": [
4439     {
4440       "offset": 0,
4441       "length": 41
4442     }
4443   ]
4444 },
4445 {
4446   "content": "Study guide for Exam AZ-900: Microsoft",
4447   "polygon": [
4448     0.7415,
4449     1.2517,
4450     7.344,
4451     1.2478,
4452     7.3443,
4453     1.6426,
4454     0.7417,
4455     1.6465
4456   ],
4457   "spans": [
4458     {
```

# Document Intelligence

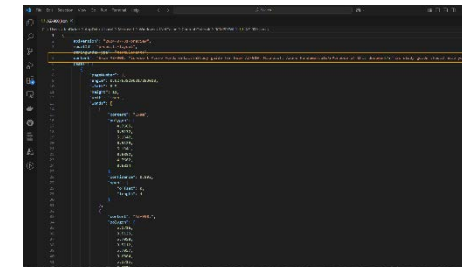
```
1 {
2   "apiVersion": "2024-07-31-preview",
3   "modelId": "prebuilt-layout",
4   "stringIndexType": "textElements",
5   "content": "Exam AZ-900: Microsoft Azure Fundamentals\nStudy guide for Exam AZ-900: Microsoft Azure Fundamentals\nPurpose of this document\nThis study guide should help
6 > "pages": [...],
47384 },
47385 "tables": [
47386   {
47387     "rowCount": 11,
47388     "columnCount": 2,
47389     "cells": [
47390       {
47391         "kind": "columnHeader",
47392         "rowIndex": 0,
47393         "columnIndex": 0,
47394         "content": "Useful links",
47395         "boundingRegions": [
47396           {
47397             "pageNumber": 1,
47398             "polygon": [
47399               0.7275,
47400               3.3883,
47401               2.1116,
47402               3.3883,
47403               2.1116,
47404               3.912,
47405               0.7275,
47406               3.8976
47407             ]
47408           }
47409         ],
47410         "spans": [
47411           {
47412             "offset": 400,
47413             "length": 12
47414           }
47415         ],
47416         "elements": [
47417           "/paragraphs/4"
47418         ]
47419       },
47420       {
47421         "kind": "columnHeader",
47422         "rowIndex": 0,
47423         "columnIndex": 1,
47424         "content": "Description"
47425       }
47426     ]
47427   }
47428 ]
47429 }
```

## Study guide for Exam AZ-900: Microsoft Azure Fundamentals

### Purpose of this document

This study guide should help you understand what to expect on the exam and includes a summary of the topics the exam might cover and links to additional resources. The information and materials in this document should help you focus your studies as you prepare for the exam.

Useful links	Description
<a href="#">Review the skills measured as of October 26, 2022</a>	This list represents the skills measured AFTER the date provided. Study this list if you plan to take the exam AFTER that date.
<a href="#">Review the skills measured prior</a>	Study this list of skills if you take your exam PRIOR to the date provided.



# Document Intelligence

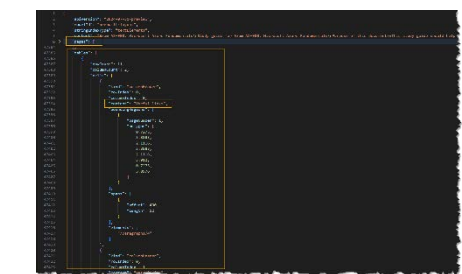
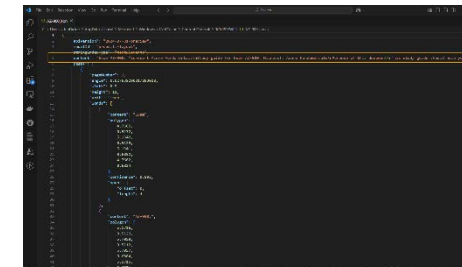
```
1 {
2   "apiVersion": "2024-07-31-preview",
3   "modelId": "prebuilt-layout",
4   "stringIndexType": "textElements",
5   "content": "Exam AZ-900: Microsoft Azure Fundamentals\nStudy guide for Exam AZ-900: Microsoft Azure Fundamentals\nPurpose of this document\nThis study guide should help you u
6 > "pages": [ ...
47384 ],
47385 > "tables": [ ...
50293 ],
50294 "paragraphs": [
50295   {
50296     "spans": [
50297       {
50298         "offset": 0,
50299         "length": 41
50300       }
50301     ],
50302     "boundingRegions": [
50303       {
50304         "pageNumber": 1,
50305         "polygon": [
50306           4.7664,
50307           0.5099,
50308           7.7596,
50309           0.5131,
50310           7.7594,
50311           0.6886,
50312           4.7662,
50313           0.6853
50314         ]
50315       }
50316     ],
50317     "role": "pageHeader",
50318     "content": "Exam AZ-900: Microsoft Azure Fundamentals"
50319   },
50320   {
50321     "spans": [
50322       {
50323         "offset": 42,
50324         "length": 57
50325       }
50326     ],
50327     "boundingRegions": [
50328       {
50329         "pageNumber": 1,
50330         "polygon": [
```

## Study guide for Exam AZ-900: Microsoft Azure Fundamentals

### Purpose of this document

This study guide should help you understand what to expect on the exam and includes a summary of the topics the exam might cover and links to additional resources. The information and materials in this document should help you focus your studies as you prepare for the exam.

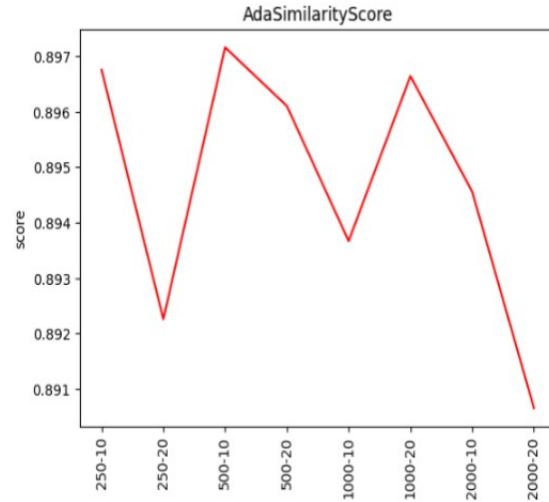
Useful links	Description
<a href="#">Review the skills measured as of October 26, 2022</a>	This list represents the skills measured AFTER the date provided. Study this list if you plan to take the exam AFTER that date.
<a href="#">Review the skills measured prior</a>	Study this list of skills if you take your exam PRIOR to the date provided.



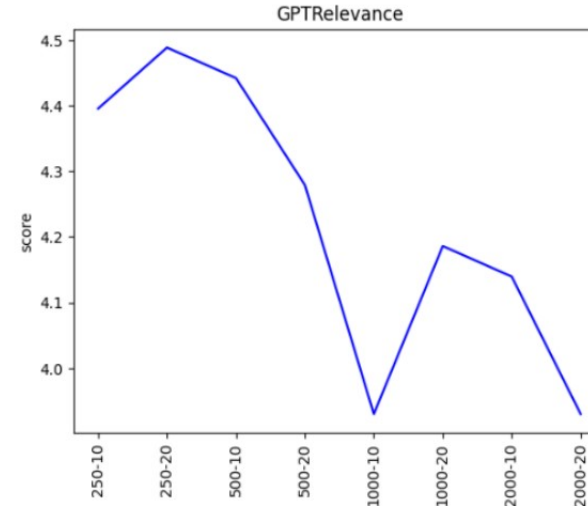
# LLM Ops



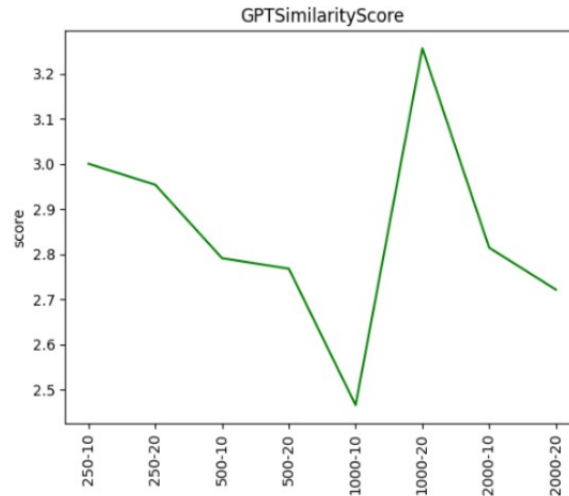
# Why Basic Chunking is Not Enough



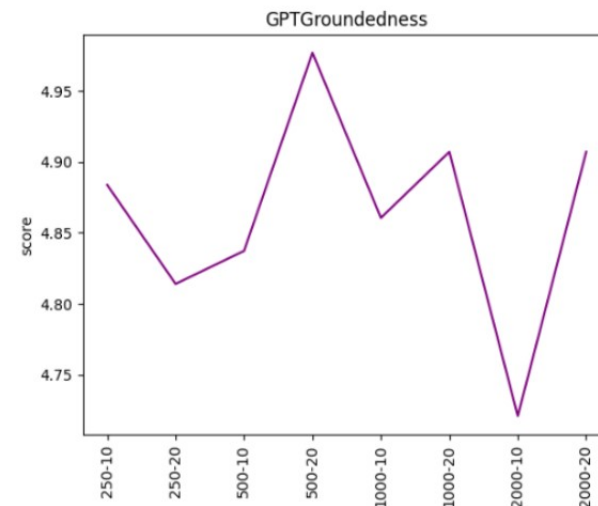
Ada Similarity Score = mathematically calculates the similarity between the embedded ground truth answer and the embedded chat's response



GPT Relevance = has GPT give a score (5 to 1, with 5 being the best) as to whether the chat's response is relevant to the question being asked

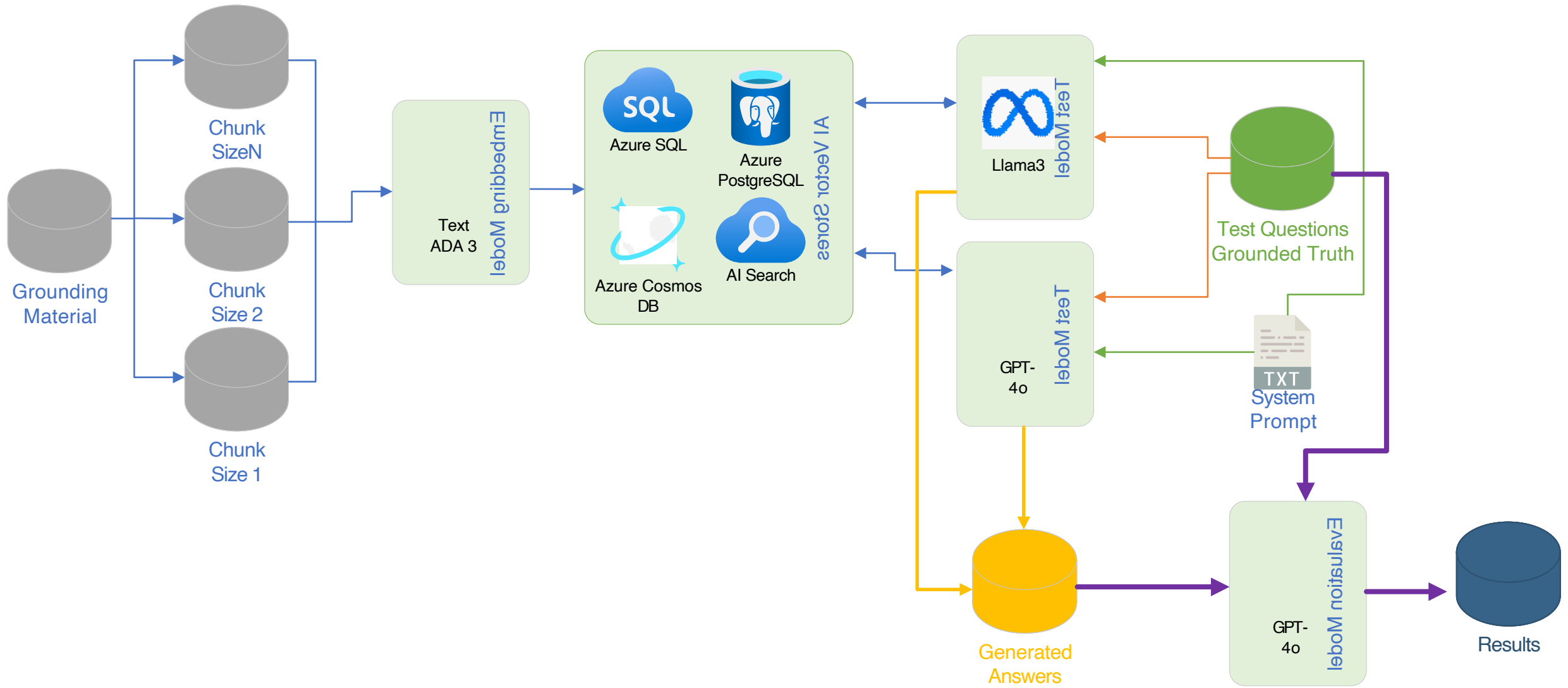


GPT Similarity = has GPT give a score (5 to 1, with 5 being the best) as to whether the ground truth answer is similar to the chat's response



GPT Groundedness = has GPT give a score (5 to 1, with 5 being the best) as to whether the chat's response is based on the context it received from the search index chunks

# LLMOps



# GPT Groundedness

## Definition

- Groundedness measures **how well the model's predicted answers align with the context provided**. Even if the responses are factually accurate, they must be verifiable within the given context to be considered well-grounded. This metric does not have any implications for the factual accuracy of the response.

## Importance

- Groundedness ensures that the model's generated content is not detached from reality or context. It helps prevent the model from producing answers that sound plausible but lack proper grounding.

## Evaluation

- During evaluation, the model's responses are assessed based on their alignment with the context. A well designed prompt sent to the Azure Open AI service is used to rate the groundedness on a scale of 1 – 5.
- Azure AI Studio provides tools to log, view, and analyze detailed evaluation metrics, including groundedness.

# GPT Retrieval Score

## Definition

- Retrieval Score measures **the quality and relevance of the retrieved document** or external information **to the question that has been asked, and its intent**
- It assesses how well the retrieved content aligns with the user's query and conversation history

## Importance

- Retrieval Score ensures that the model's generated responses are based on information that is likely to be relevant to the answer.
- High retrieval quality leads to more accurate and context-aware answers.
- It enhances the overall conversational experience by providing well-researched content.

## Evaluation

- During evaluation, the retrieved documents are assessed both individually and as a group based on their alignment with the question that has been asked. This assessment is done by a well-engineered prompt to the generative AI model. The AI model returns a score on a scale from 1-5, as well as the reasoning as to why it chose this score.

# GPT Relevance

## Definition:

- GPT Relevance refers to how well the model's **generated responses align with the context provided by the user and the conversation.**
- It measures the extent to which the answers produced by the model are contextually appropriate and logically connected to the input.
- It does this by generating a reference response based purely on the question and retrieved documents. It then compares this to the full answer. If the reference response is deemed to be acceptably similar to the actual answer, it scores highly.

## Importance:

- Relevance ensures that the AI-generated content makes sense within the given context.
- High relevance leads to more meaningful responses.
- It enhances the overall quality of interactions, making them more useful and engaging. Higher relevance scores are also expected to correlate with a higher probability of a correct answer

## Evaluation:

- During evaluation, the model's responses are assessed based on their alignment with the context. This evaluation is performed by the Open AI model, and returns scores on a scale from 1-5.
- Metrics related to relevance, coherence, and contextual understanding are used to quantify this alignment.
- A well-referenced and contextually relevant answer receives a higher GPT Relevance score.
- Remember, GPT Relevance bridges the gap between generated content and context, making AI conversations more accurate and contextually aware.

# GPT Similarity

## Definition:

- GPT Similarity refers to the **measurement of semantic similarity between different pieces of text** using models like GPT-4.
- It assesses how closely related the generated answer is to a user defined ground truth that is assumed to be correct. This makes this metric an imperfect proxy for accuracy.

## Importance:

- GPT Similarity helps determine how well a language model captures semantic relationships.
- It enables applications such as question answering, text summarization, and information retrieval.
- High similarity scores indicate that the model can answer questions with some degree of accuracy, and that the answers may align semantically with what an expert end-user may wish to see

## Evaluation:

- To evaluate GPT Similarity, **we compare the model's predicted answers with reference ground truth answers.** Both answers are sent to the Open AI model, alongside some example evaluations, and the result is a score from 1-5 and the reasoning behind this score.
- A higher similarity score indicates better alignment with the expected meaning. Although imperfect, GPT-4 often picks up on factual discrepancies between the two answers, even if they are semantically similar. This gives us a reasonable proxy for correctness in the absence of an expert evaluator.

# ADA Cosine Similarity

## Definition:

- Cosine Similarity measures **how similar two non-zero vectors are** in a multi-dimensional space.
- It computes the cosine of the angle between the vectors, which is the dot product of the vectors divided by the product of their magnitudes (lengths). 1 indicates equality between the vectors. 0 indicates perfect dissimilarity (“orthogonality”) between the vectors.
- The vectors are generated using semantic embeddings from the ADA model in Open AI.

## Importance:

- Cosine Similarity helps us understand the relationship between two elements based on their direction rather than individual values.
- It’s widely used in various applications, including recommendation systems, text analysis, and image processing. This technique is purely statistical in nature, making it **less prone to variance than the GPT-focused metrics**, at the cost of being slightly more complex to interpret.
- High cosine similarity indicates strong alignment between vectors, making it valuable for tasks like document similarity and clustering.

## Evaluation:

- **A ground truth response and the model response are both sent to the ADA model**, and the resulting vector embeddings are compared using the cosine similarity metric.
- The resulting similarity ranges from -1 (opposite) to 1 (same), with 0 indicating orthogonality (no correlation).
- Remember, cosine similarity provides a powerful way to measure similarity and dissimilarity between vectors, enhancing our ability to analyze data.

# LLM Ops

	A	B	C	D	E	F	G	H	I
1	Retrieval Content	Ground Truth	Question	Generated Answer	Relevance	Groundedness	Retrieval Score	Cosine Similarity	GPT Similarity
2			How do I protect the storage of sterno in metal cans in cardboard boxes on 12 ft shelves?		4	5	5	1	1
3			Can I store ignitable liquids in portable containers in an inside cutoff room?		4	5	5	1	2
4			Follow-up: Can you provide more information on specific guidelines or restrictions? (Cutoff room question)		4	5	5	1	1
	Content: service to the minimum necessary for efficient	For more specific guidelines on storing ignitable	2nd Follow-up: Do you need more details on	Yes, I can provide more information about the					



# Useful Links

- <https://tinyurl.com/oit-sandbox>
- [Azure-Samples/azureai-basic-python: Basic sample for deploying chat web apps with Azure AI Studio and SDKs](#)
- [Azure-Samples/contoso-creative-writer: A creative writing multi-agent solution to help users write articles.](#)
- [Azure-Samples/contoso-chat: This sample has the full End2End process of creating RAG application with Prompty and Azure AI Foundry.](#)
- [GraphRAG: Unlocking LLM discovery on narrative private data - Microsoft Research](#)

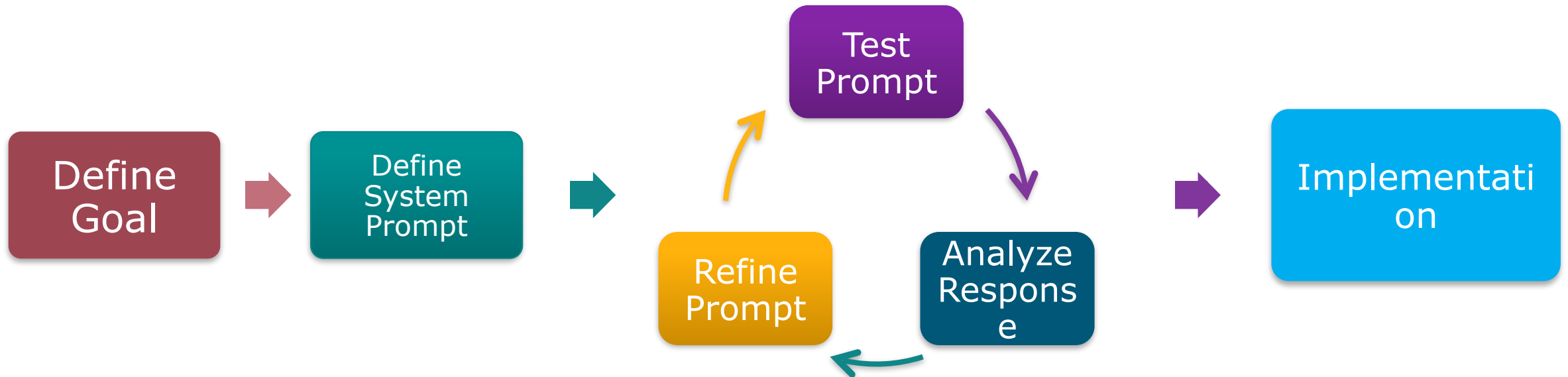


# Thank you

Ian Dicker  
[idicker@spyglassmtg.com](mailto:idicker@spyglassmtg.com)

# Developing a Prompt

- Prompt engineering is the art of communicating with a Generative AI model.
- Prompt engineering is the process of refining interactions with AI systems, such as ChatGPT, to produce optimal responses. A prompt engineer crafts the right question or command that will guide the AI to deliver the most accurate and useful answer.



# Responsible AI

Spyglass follows the Microsoft guidelines for responsible AI

## Fairness

- How might an AI system allocate opportunities, resources, or information in ways that are fair to the humans who use it?

## Reliability & Safety

- How might the system function well for people across different use conditions and contexts, including ones it was not originally intended for?

## Privacy & Security

- How might the system be designed to support privacy and security?

## Inclusiveness

- How might the system be designed to be inclusive of people of all abilities?

## Transparency

- How might people misunderstand, misuse, or incorrectly estimate the capabilities of the system?

## Accountability

- How can we create oversight so that humans can be accountable and in control?