



Incorporate Fallback Mechanisms For Your Amazon Bedrock RAG Pipelines

Bilge Yücel
AWS Community Day Türkiye 2024



About me



Bilge Yücel

- 🥑 Developer Relations Engineer at deepset
🇩🇪
 - 🚧 Open source LLM Framework: Haystack
 - 🎓 Sabancı University B.Sc.
 - 🧠 KU Leuven M.Sc. 🇧🇪
 - 📍 Istanbul, Turkey
-  @bilgeycl
-  Bilge Yücel

Agenda

- ★ Retrieval Augmented Generation
- ★ Haystack & Amazon Bedrock
- ★ Fallback Mechanisms
- ★ Q&A



Large Language Models (LLMs)

The image shows a dark-themed interface of a large language model, likely ChatGPT. On the left, there is a green circular profile picture with the letters 'BI' inside. To its right, the word 'You' is displayed in white. Below this, a question is asked: 'What is Tortured Poets Department album about?'. On the left side of the main message area, there is a circular icon containing a stylized blue and white knot or swirl pattern. To its right, the text 'ChatGPT' is written in white. The AI's response follows: 'The "Tortured Poets Department" album you're referring to doesn't seem to be widely recognized or part of mainstream culture as of my last update in January 2022. If it's a recent release or an independent project, I might not have information about it. Could you provide more context or details about the album?'. At the bottom left of the message area, there are two small white icons: a square with a vertical line and a square with a diagonal line.

Retrieval Augmented Generation

Retrieval Augmented Generation

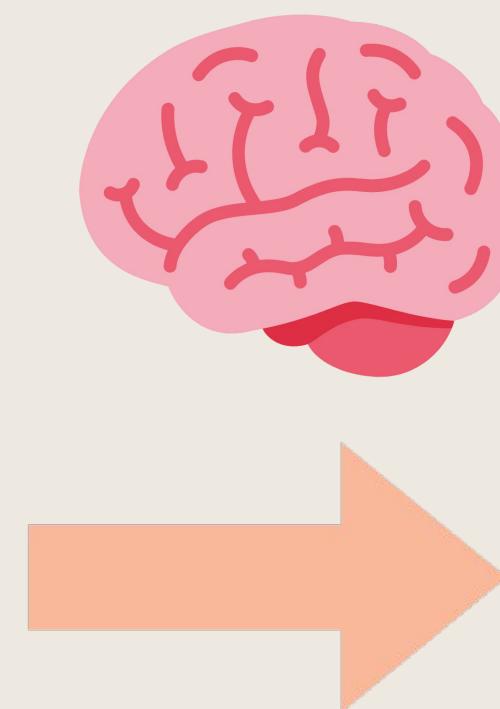


Retrieval Augmented Generation

Answer the following query
given the documents

(Some Relevant
Documents)

What is Tortured Poets
Department album about?



Fans have theorized
that the upcoming
album will detail Swift's
split from actor Joe
Alwyn, whom she dated
for six years before
splitting in April 2023.
Not only do the song
titles...





Haystack
by deepset



- Haystack is an open source Python framework for building production-ready LLM applications



Haystack
by deepset



- Haystack is an open source Python framework for building production-ready LLM applications
- Prototyping, evaluation, deployment, monitoring...

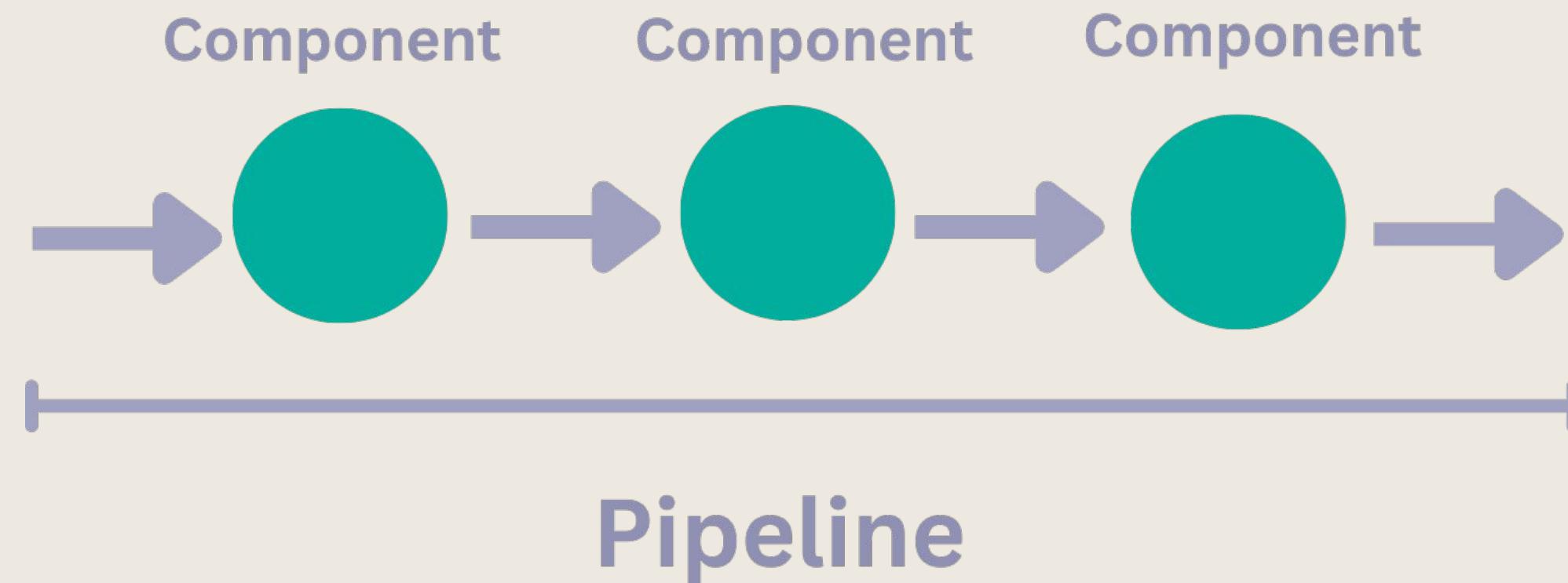


Haystack

by deepset

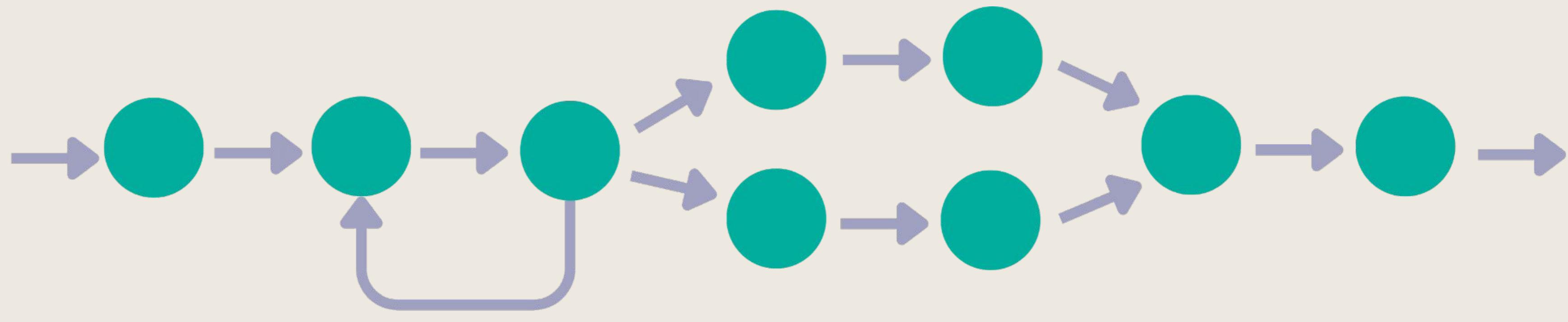


- Haystack is an open source Python framework for building production-ready LLM applications
- Prototyping, evaluation, deployment, monitoring...
- Building blocks: Components & Pipelines



elle

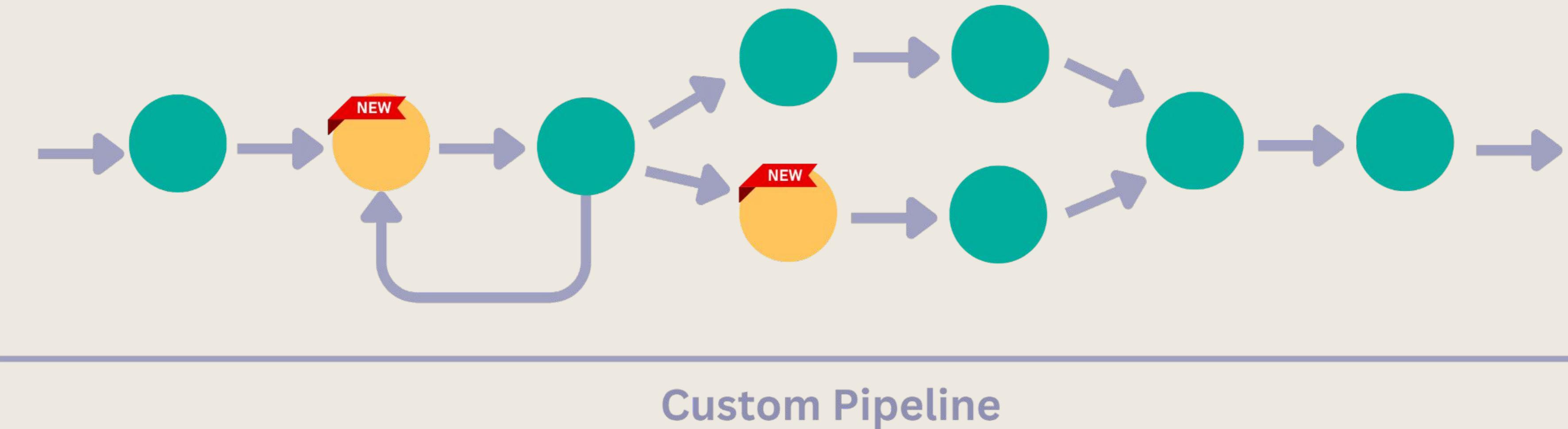
Haystack



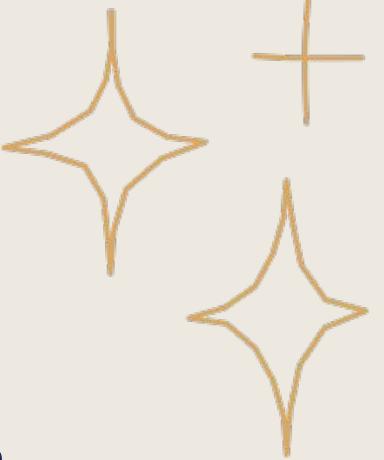
Custom Pipeline

elle

Haystack



Amazon Bedrock



- Foundation models from **Anthropic**, **Cohere**, **Meta**, **Mistral AI**, **Amazon** and more through single API



ANTHROPIC



Amazon Bedrock



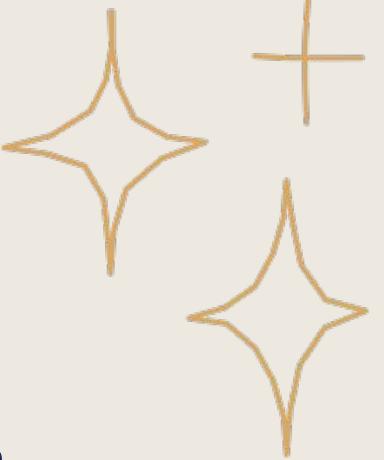
- Foundation models from **Anthropic**, **Cohere**, **Meta**, **Mistral AI**, **Amazon** and more through single API
- Serverless, no infrastructure management



ANTHROPIC



Amazon Bedrock



- Foundation models from **Anthropic**, **Cohere**, **Meta**, **Mistral AI**, **Amazon** and more through single API
- Serverless, no infrastructure management
- Data is not shared with model providers



ANTHROPIC



Amazon Bedrock



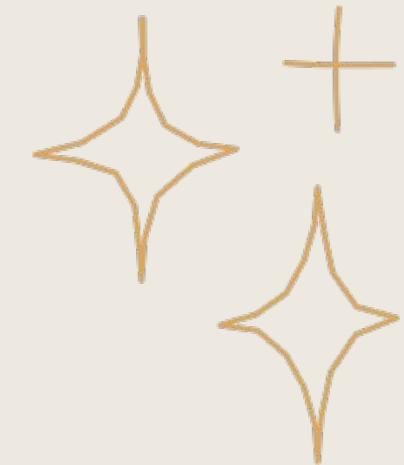
- Foundation models from **Anthropic**, **Cohere**, **Meta**, **Mistral AI**, **Amazon** and more through single API
- Serverless, no infrastructure management
- Data is not shared with model providers
- Haystack supports it!



ANTHROPIC

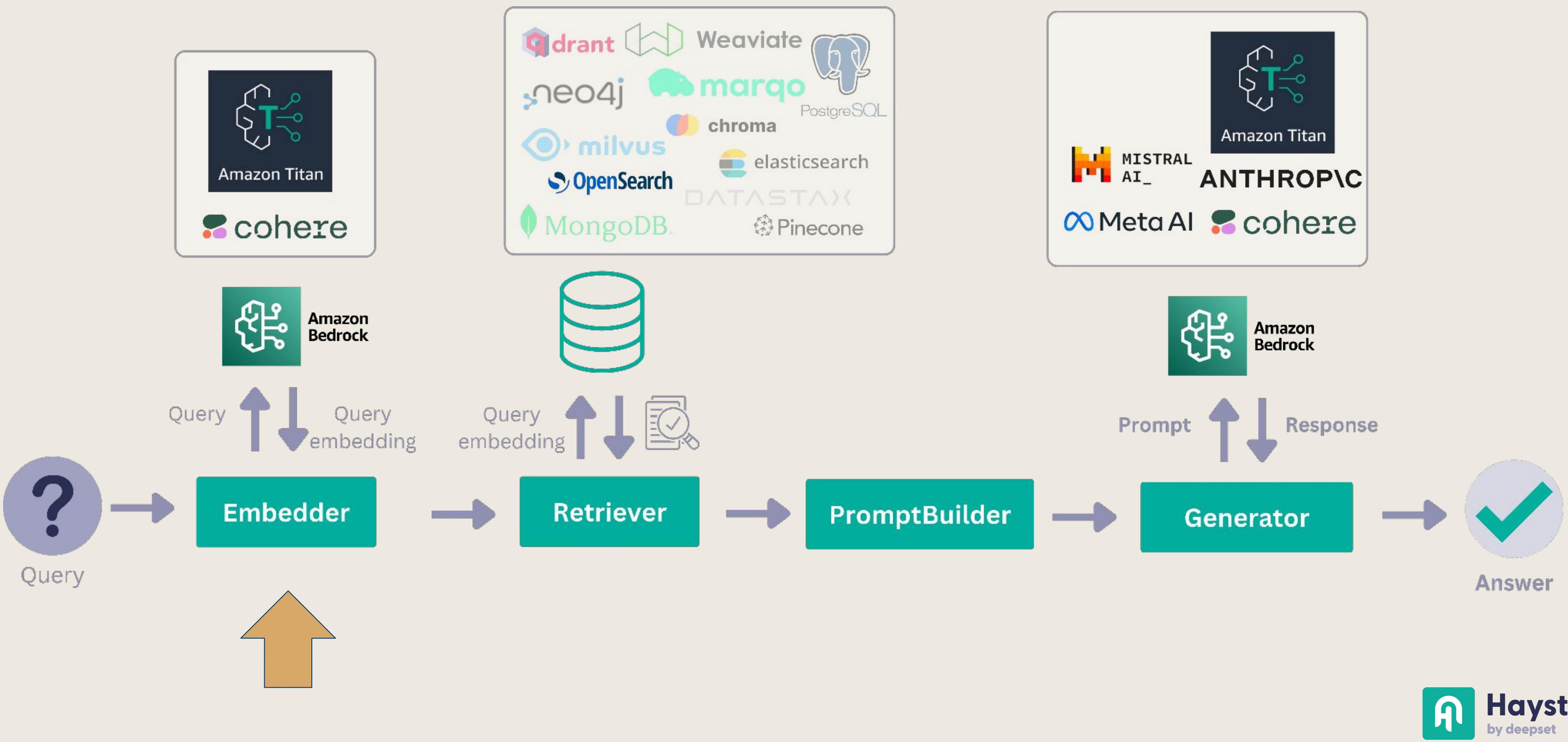


Models

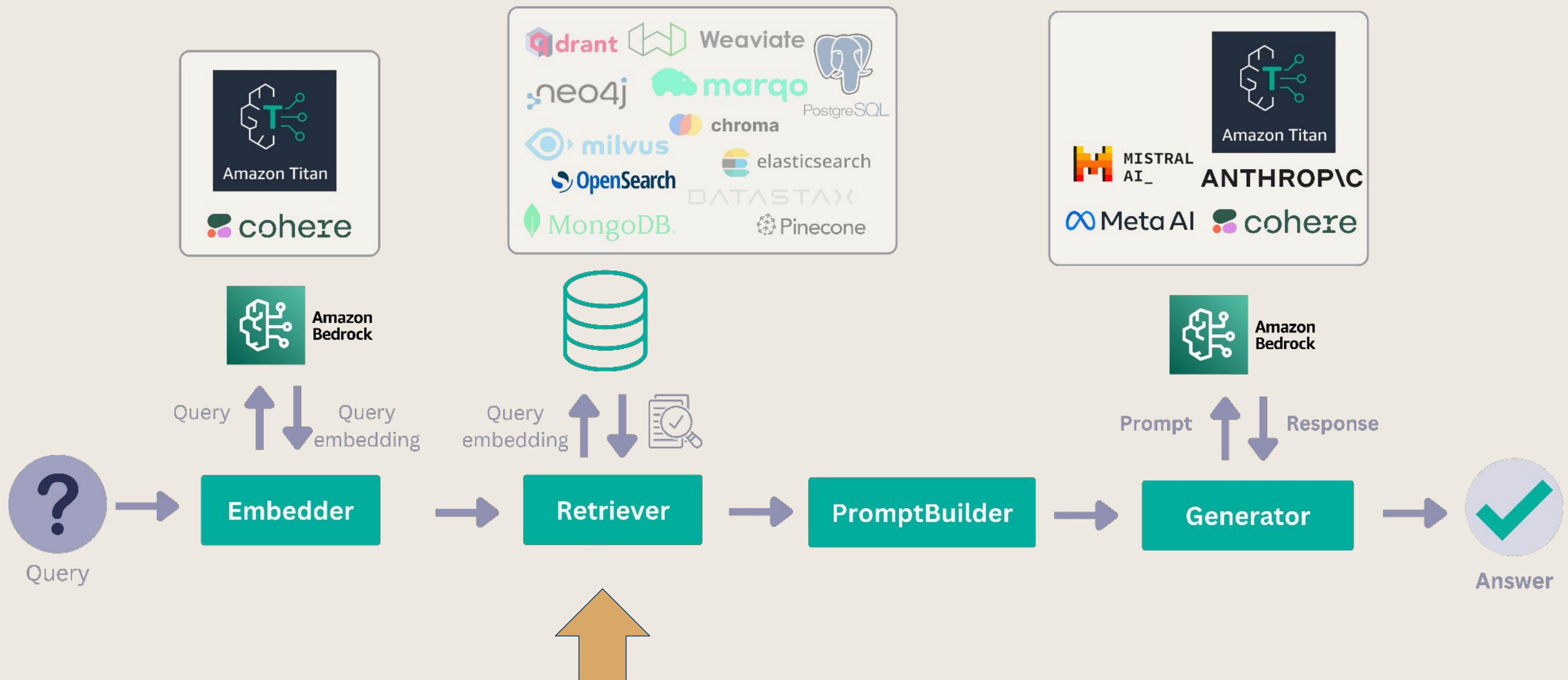


Provider	Model name	Version	Model ID
Amazon	Titan Text G1 - Express	1.x	amazon.titan-text-express-v1
Amazon	Titan Text G1 - Lite	1.x	amazon.titan-text-lite-v1
Amazon	Titan Text Premier	1.x	amazon.titan-text-premier-v1:0
Amazon	Titan Embeddings G1 - Text	1.x	amazon.titan-embed-text-v1
Amazon	Titan Embedding Text v2	1.x	amazon.titan-embed-text-v2:0
Amazon	Titan Multimodal Embeddings G1	1.x	amazon.titan-embed-image-v1
Amazon	Titan Image Generator G1	1.x	amazon.titan-image-generator-v1
Anthropic	Claude	2.0	anthropic.claude-v2
Anthropic	Claude	2.1	anthropic.claude-v2:1
Anthropic	Claude 3 Sonnet	1.0	anthropic.claude-3-sonnet-20240229-v1:0
Anthropic	Claude 3 Haiku	1.0	anthropic.claude-3-haiku-20240307-v1:0
Anthropic	Claude 3 Opus	1.0	anthropic.claude-3-opus-20240229-v1:0
Anthropic	Claude Instant	1.x	anthropic.claude-instant-v1
AI21 Labs	Jurassic-2 Mid	1.x	ai21.j2-mid-v1
AI21 Labs	Jurassic-2 Ultra	1.x	ai21.j2-ultra-v1
Cohere	Command	14.x	cohere.command-text-v14

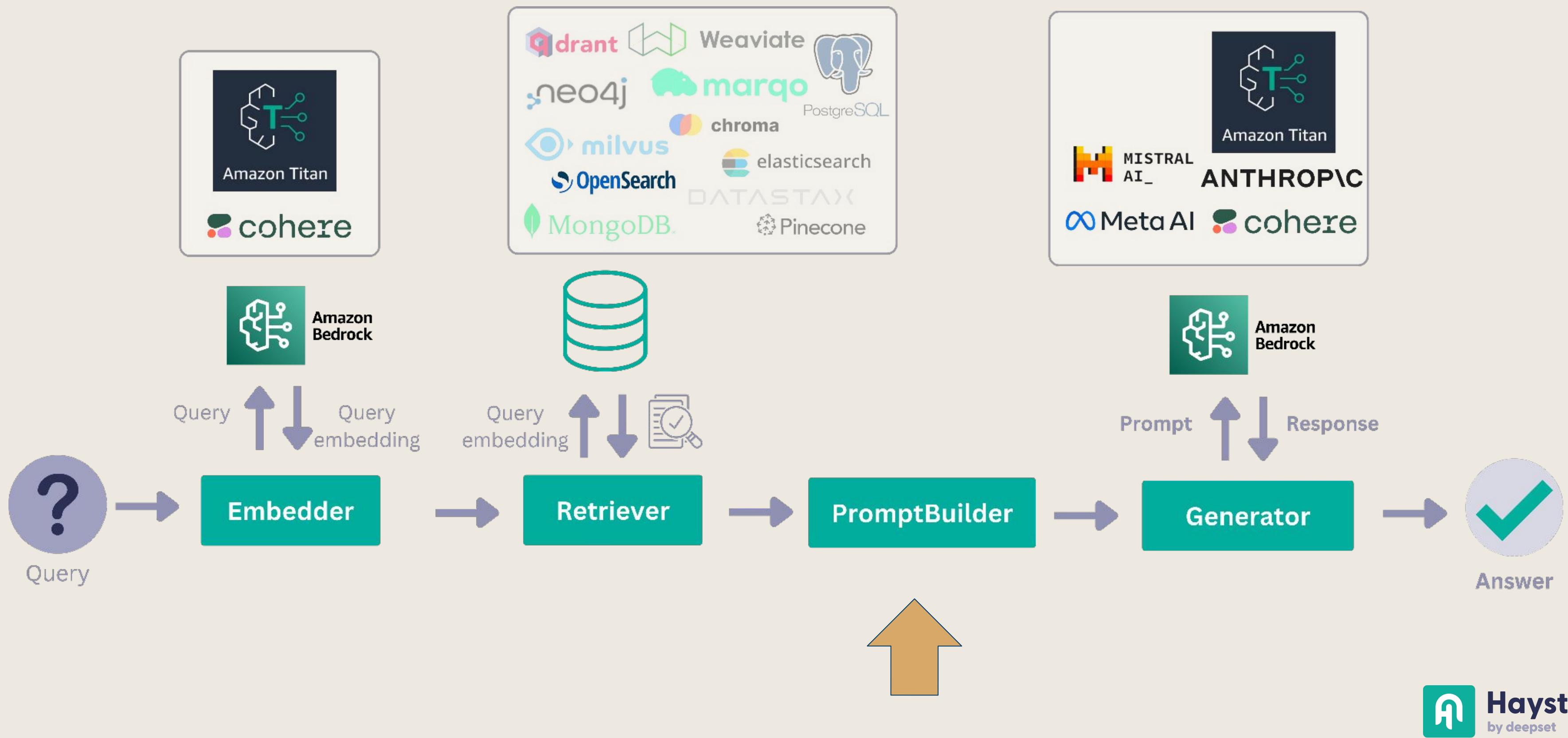
RAG PIPELINE



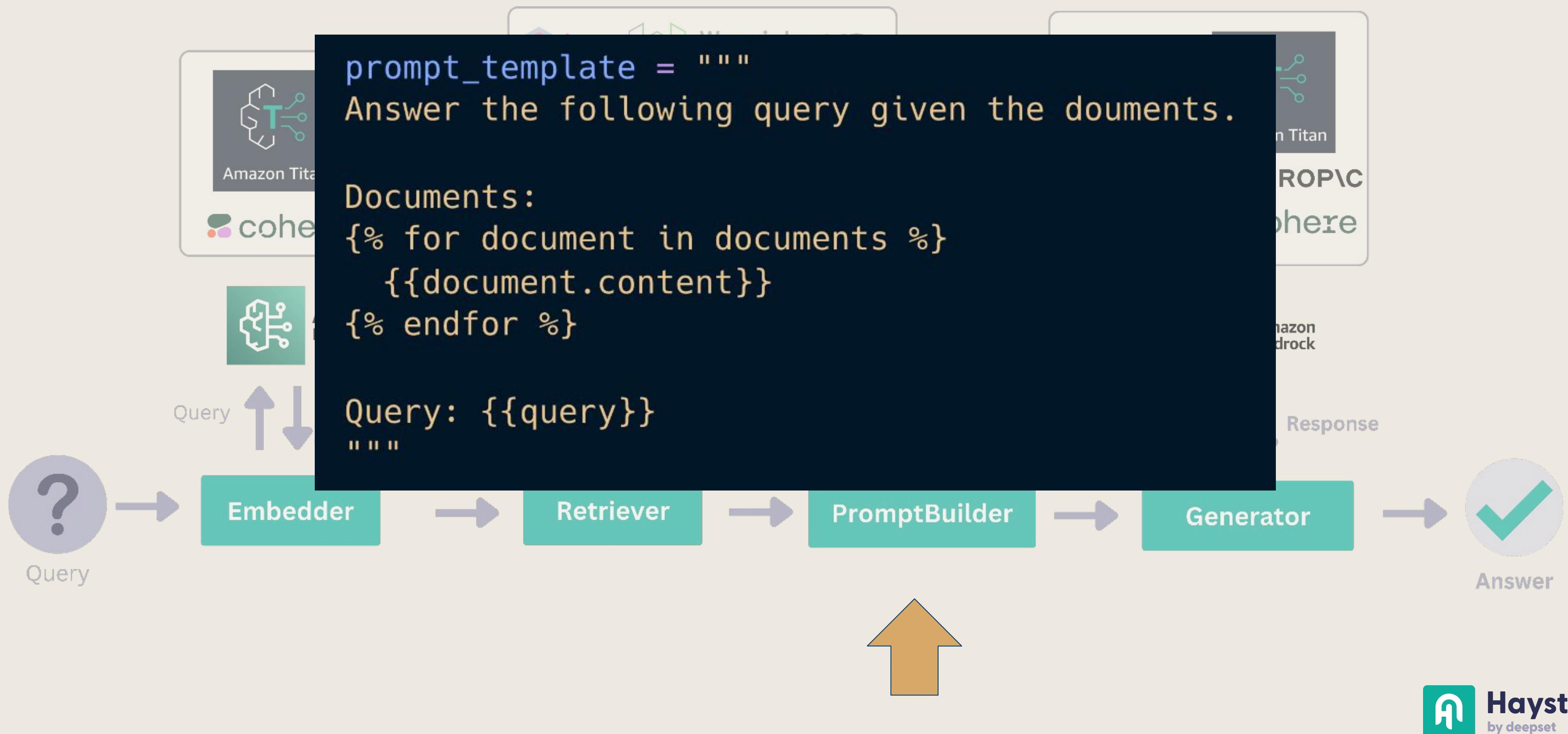
RAG PIPELINE



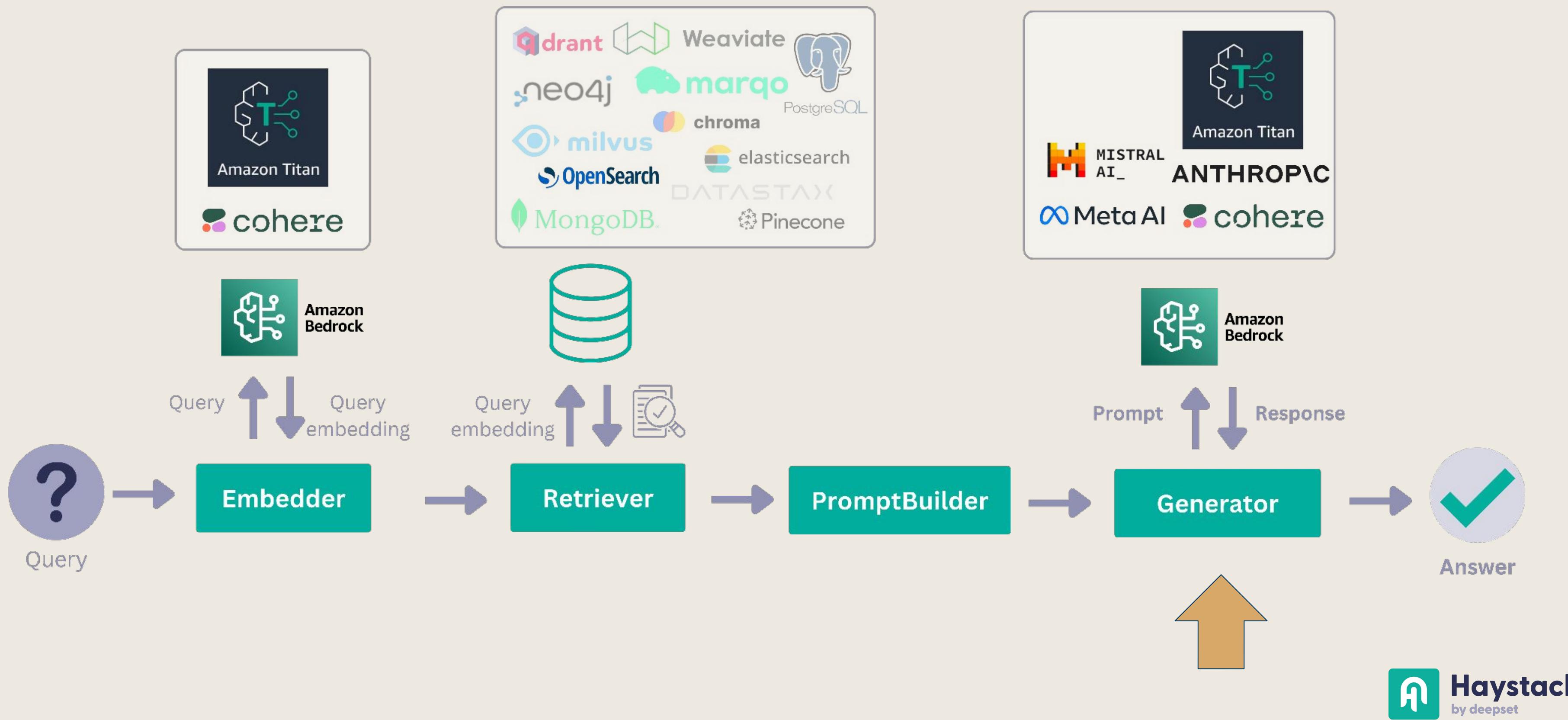
RAG PIPELINE



RAG PIPELINE



RAG PIPELINE



RAG PIPELINE



```
rag_pipeline = Pipeline()
rag_pipeline.add_component("text_embedder", AmazonBedrockTextEmbedder(model="amazon.titan-embed-text-v1"))
rag_pipeline.add_component("retriever", InMemoryEmbeddingRetriever(document_store=document_store))
rag_pipeline.add_component("prompt_builder", PromptBuilder(template=prompt_template))
rag_pipeline.add_component("generator", AmazonBedrockGenerator(model='amazon.titan-text-express-v1'))

rag_pipeline.connect("text_embedder.embedding", "retriever.query_embedding")
rag_pipeline.connect("retriever", "prompt_builder.documents")
rag_pipeline.connect("prompt_builder", "generator")

query="What's Tortured Poets Department album about?"

rag_pipeline.run({"text_embedder":{"text": query}, "prompt_builder":{"query": query}})

>>> Fans have theorized that the upcoming album will detail Swift's split from actor Joe Alwyn, whom she dated for six years before splitting in April 2023. Not only do the song titles include references to the English actor and their former relationship, but Swift recently teased that the album is about heartbreak as she shared five exclusive playlists with Apple Music.\nAccording to a press release, with each playlist, Swift has chosen songs from her own catalog that fit each stage of a breakup: denial, anger, bargaining, depression and acceptance.
```

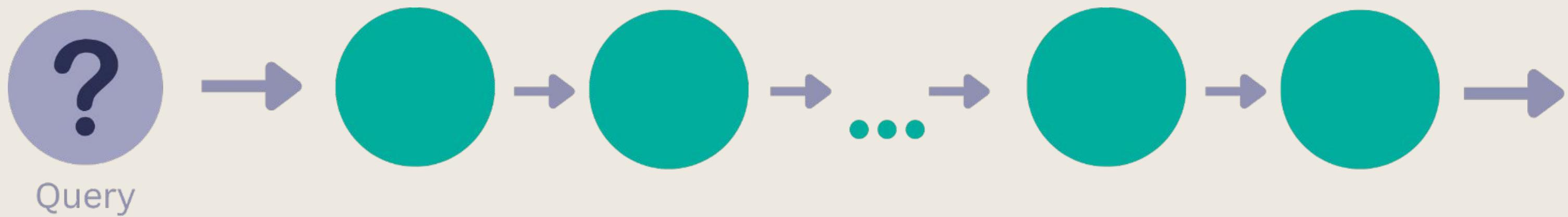


Fallback Mechanisms

- ★ Routing/Branching
- ★ Looping

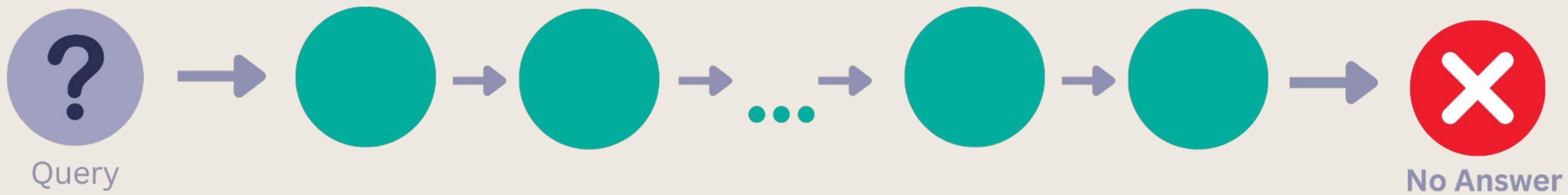


RAG PIPELINE



*When is Taylor Swift's
next concert?*

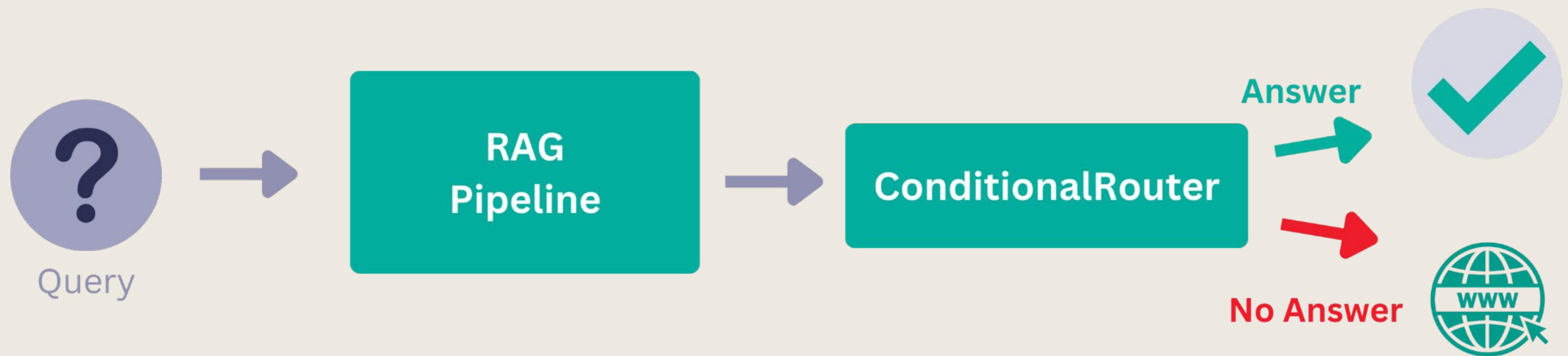
RAG PIPELINE



*When is Taylor Swift's
next concert?*

*The documents do not
provide any
information about
Taylor Swift's
upcoming concert.*

Routing



Routing - Prompt Template



```
prompt_template = """
```

Answer the following query given the documents.

If the answer is not contained within the documents reply with 'no_answer'

Documents:

```
{% for document in documents %}  
    {{document.content}}  
{% endfor %}
```

Query: {{query}}

```
"""
```

Routing - Conditional Router

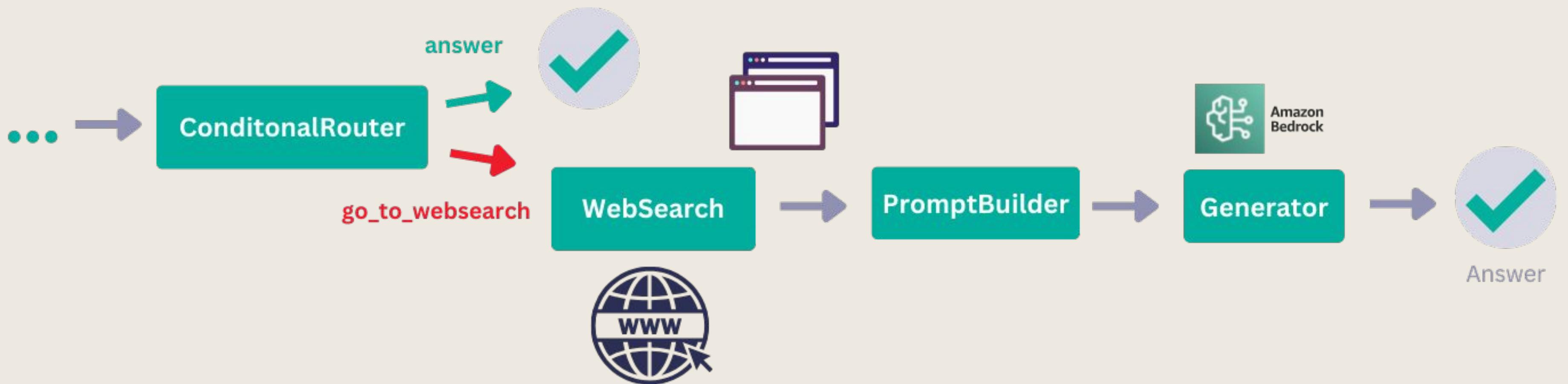
```
● ● ●

from haystack.components.routers import ConditionalRouter

routes = [
    {
        "condition": "{{'no_answer' in replies[0]}}",
        "output": "{{query}}",
        "output_name": "go_to_websearch",
        "output_type": str,
    },
    {
        "condition": "{{'no_answer' not in replies[0]}}",
        "output": "{{replies[0]}}",
        "output_name": "answer",
        "output_type": str,
    },
]

router = ConditionalRouter(routes)
```

Fallback to Web



Fallback to Web



Fallback to Web

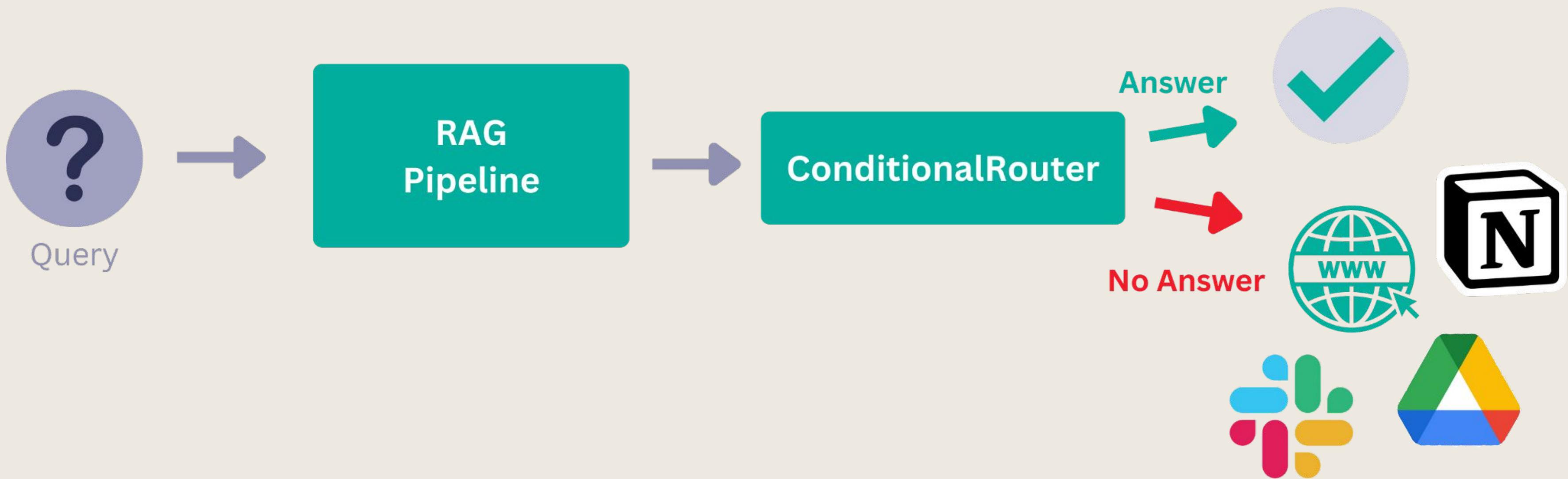


*When is Taylor Swift's
next concert?*

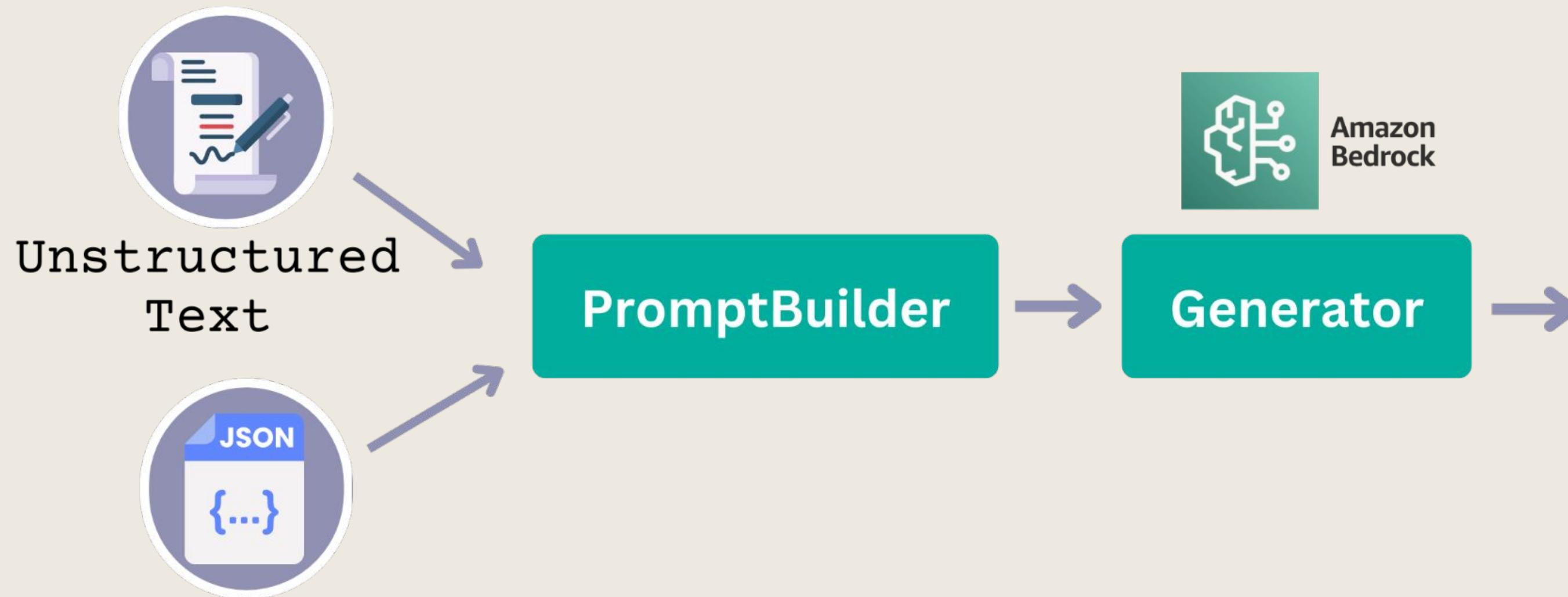


*Taylor Swift's next
concert is on May 17,
2024, at Friends
Arena in Solna,
Sweden.*

Fallback to Other Sources

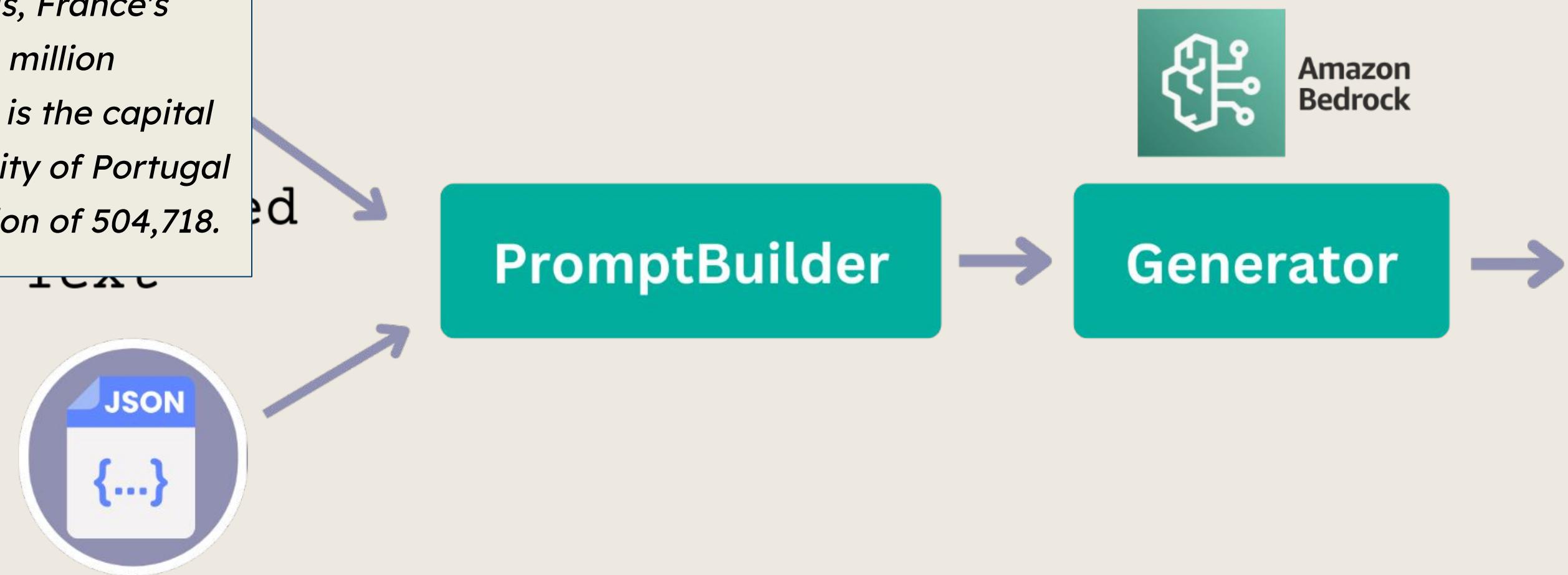


Extracting Structured Data



Extracting Structured Data

Berlin is the capital of Germany. It has a population of 3,850,809. Paris, France's capital, has 2.161 million residents. Lisbon is the capital and the largest city of Portugal with the population of 504,718.



Extracting Structured Data

Berlin is the capital of Germany. It has a population of 3,850,809. Paris, France's capital, has 2.161 million residents. Lisbon is the capital and the largest city of Portugal with the population of 504,718.

→
Text

 JSON

```
class City(BaseModel):
    name: str
    country: str
    population: int
```



```
class CitiesData(BaseModel):
    cities: List[City]
```

PromptBuilder →



Generator →

Extracting Structured Data

Berlin is the capital of Germany. It has a population of 3,850,809. Paris, France's capital, has 2.161 million residents. Lisbon is the capital and the largest city of Portugal with the population of 504,718.

→
Text

→
JSON

```
class City(BaseModel):  
    name: str  
    country: str  
    population: int
```



```
class CitiesData(BaseModel):  
    cities: List[City]
```

PromptBuilder

Generator



Amazon
Bedrock



Invalid JSON

Extracting Structured Data

Berlin is the capital of Germany. It has a population of 3,850,809. Paris, France's capital, has 2.161 million residents. Lisbon is the capital and the largest city of Portugal with the population of 504,718.

Text



```
class City(BaseModel):  
    name: str  
    country: str  
    population: int
```



```
class CitiesData(BaseModel):  
    cities: List[City]
```

PromptBuilder



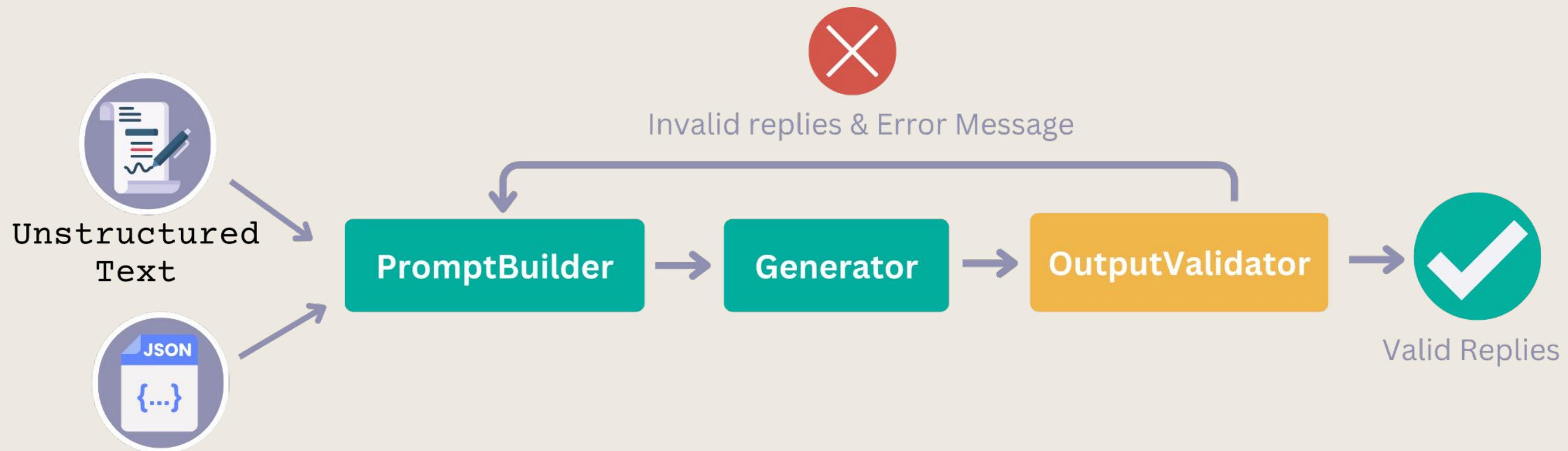
Generator



Amazon
Bedrock

```
{  
    "cities": [  
        {  
            "name": "Berlin",  
            "population": 3850809,  
            "country": "Germany"  
        },  
        {  
            "name": "Paris",  
            "population": 2161000,  
            "country": "France"  
        },  
        {  
            "name": "Lisbon",  
            "population": 504718  
        }  
    ]  
}
```

Looping



OutputValidator



```
import pydantic
from haystack import component

@Component
class OutputValidator:
    def __init__(self, pydantic_model: pydantic.BaseModel):
        self.pydantic_model = pydantic_model

    @component.output_types(valid_replies=List[str], invalid_replies=Optional[List[str]],
                           error_message=Optional[str])
    def run(self, replies: List[str]):

        try:
            output_dict = json.loads(replies[0])
            self.pydantic_model.parse_obj(output_dict)

        return {"valid_replies": replies}

    except (ValueError, ValidationError) as e:
        return {"invalid_replies": replies, "error_message": str(e)}
```

Prompt Template



```
prompt_template = """
Create a JSON object from the information present in this passage: {{passage}}.
Only use information that is present in the passage. Follow this JSON schema, but only return
the actual instances without any additional schema definition:
{{schema}}
Make sure your response is a dict and not a list.
{% if invalid_replies and error_message %}
    You already created the following output in a previous attempt: {{invalid_replies}}
    However, this doesn't comply with the format requirements from above and triggered this Python
exception: {{error_message}}
    Correct the output and try again. Just return the corrected output without any extra
explanations.
{% endif %}
"""
```

Looping

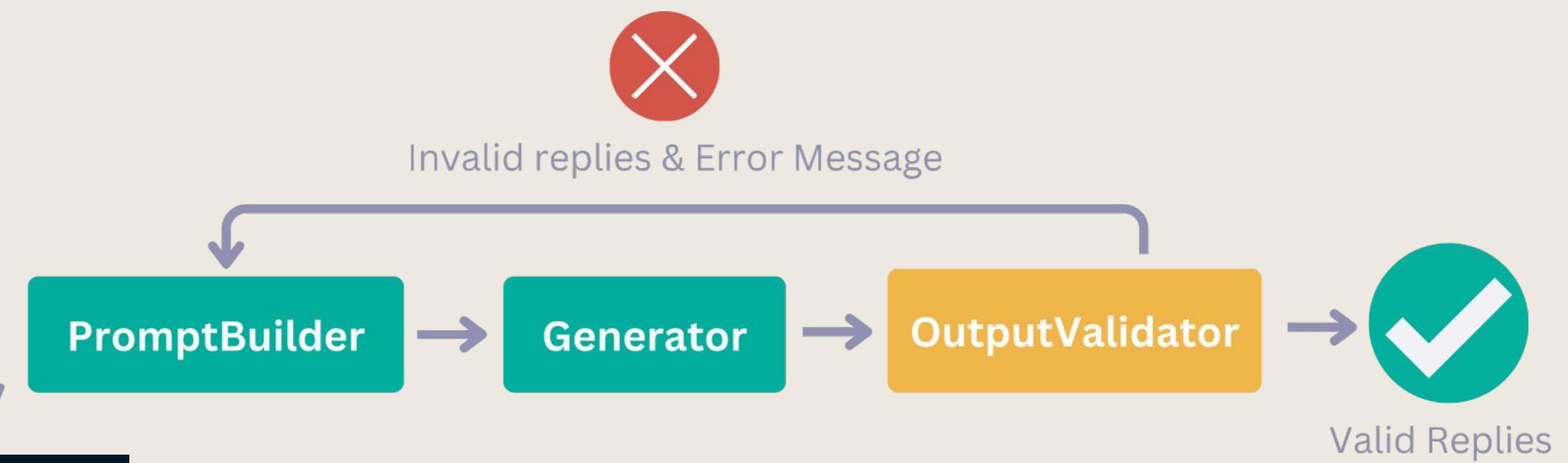
Berlin is the capital of Germany. It has a population of 3,850,809. Paris, France's capital, has 2.161 million residents. Lisbon is the capital and the largest city of Portugal with the population of 504,718.

Text



```
class City(BaseModel):  
    name: str  
    country: str  
    population: int
```

```
class CitiesData(BaseModel):  
    cities: List[City]
```



Looping

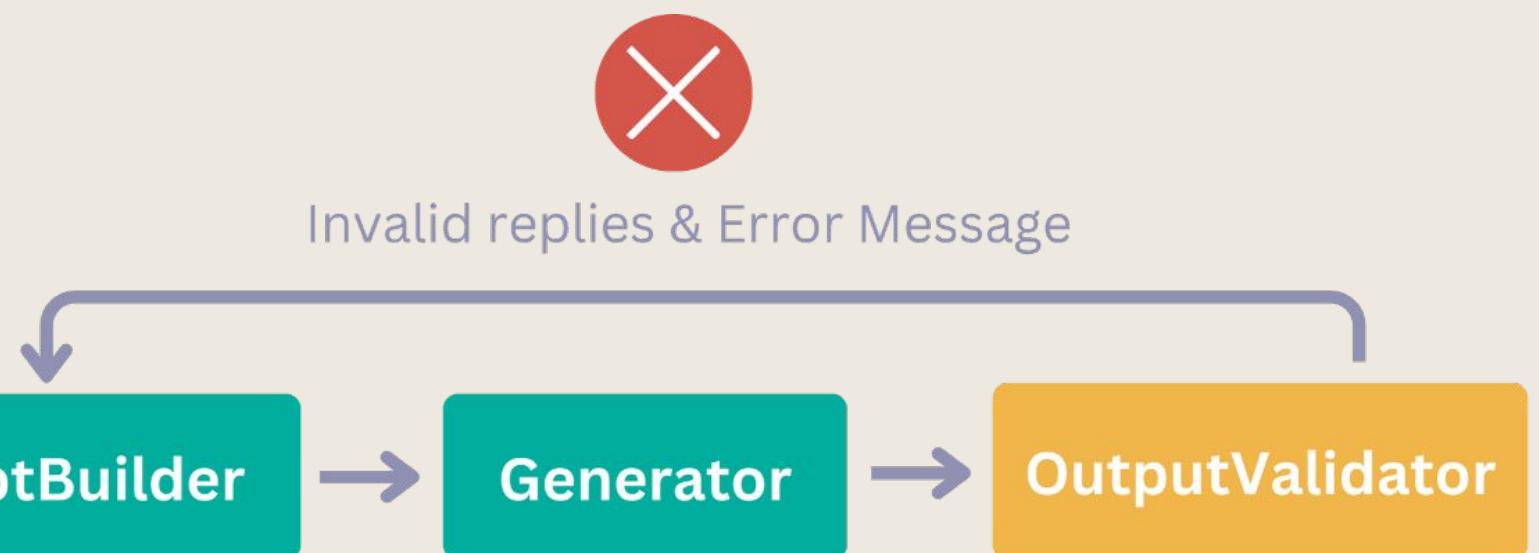
Berlin is the capital of Germany. It has a population of 3,850,809. Paris, France's capital, has 2.161 million residents. Lisbon is the capital and the largest city of Portugal with the population of 504,718.

Text



```
class City(BaseModel):  
    name: str  
    country: str  
    population: int
```

```
class CitiesData(BaseModel):  
    cities: List[City]
```



```
{  
    "cities": [  
        {  
            "name": "Berlin",  
            "country": "Germany",  
            "population": 3850809  
        },  
        {  
            "name": "Paris",  
            "country": "France",  
            "population": 2161000  
        },  
        {  
            "name": "Lisbon",  
            "country": "Portugal",  
            "population": 504718  
        }  
    ]  
}
```

Sources



Fallback to Web



🔄 Generating
Structured Output



Teşekkürler



@bilgeycl



Bilge Yücel



bilgeyucel

Haystack is open source! 



haystack.deepset.ai