Building an Multimodal Knowledge Assistant

Jerry Liu

September 23, 2024

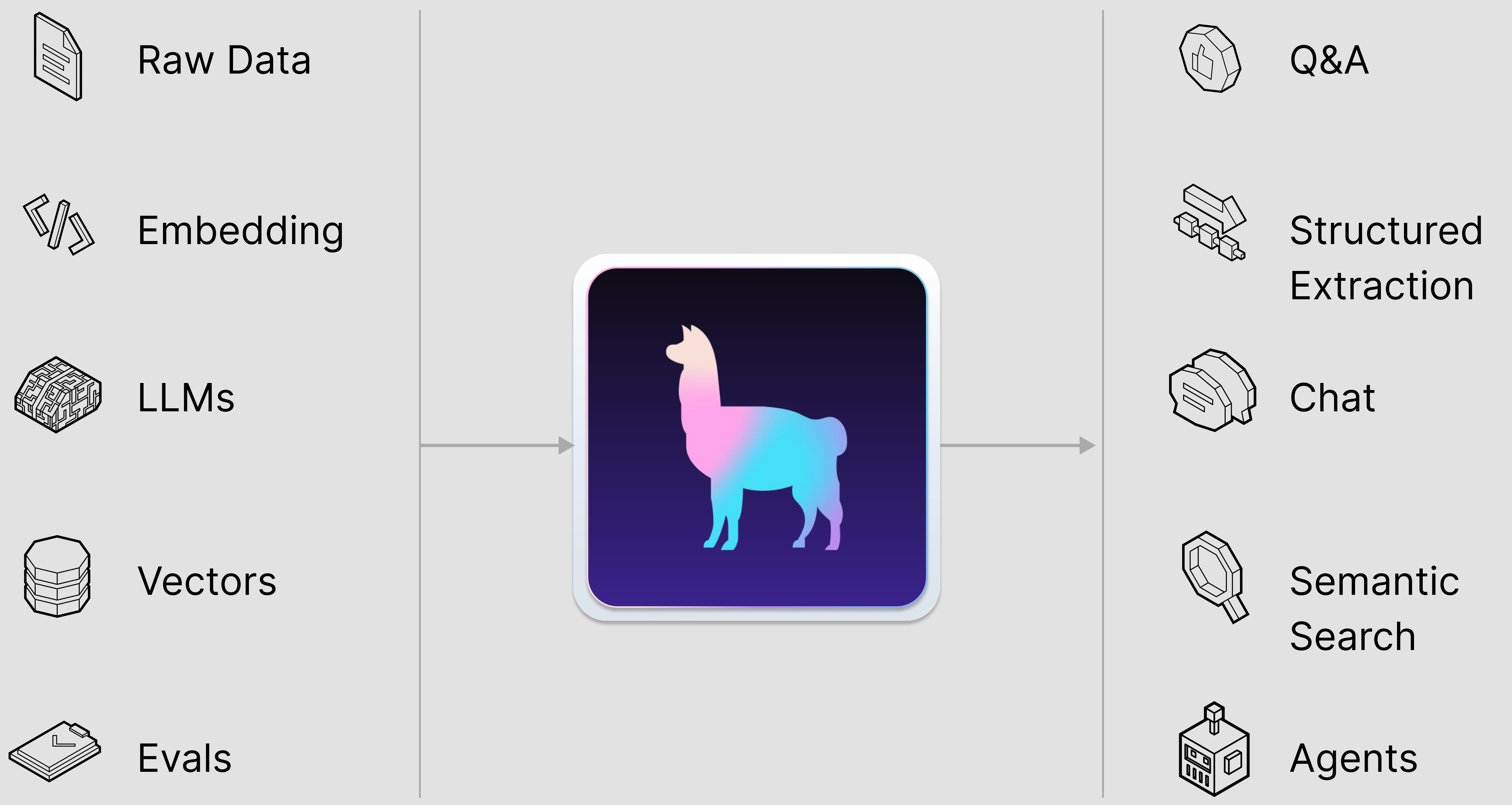
LlamaIndex:
Build Production LLM Apps over Enterprise Data

LlamaIndex helps any developer build **context-augmented LLM apps from prototype to production.**

Open-Source: Leading developer toolkit for building production LLM apps over data.
Docs: https://docs.llamaindex.ai/
Repo: https://github.com/run-llama/llama_index

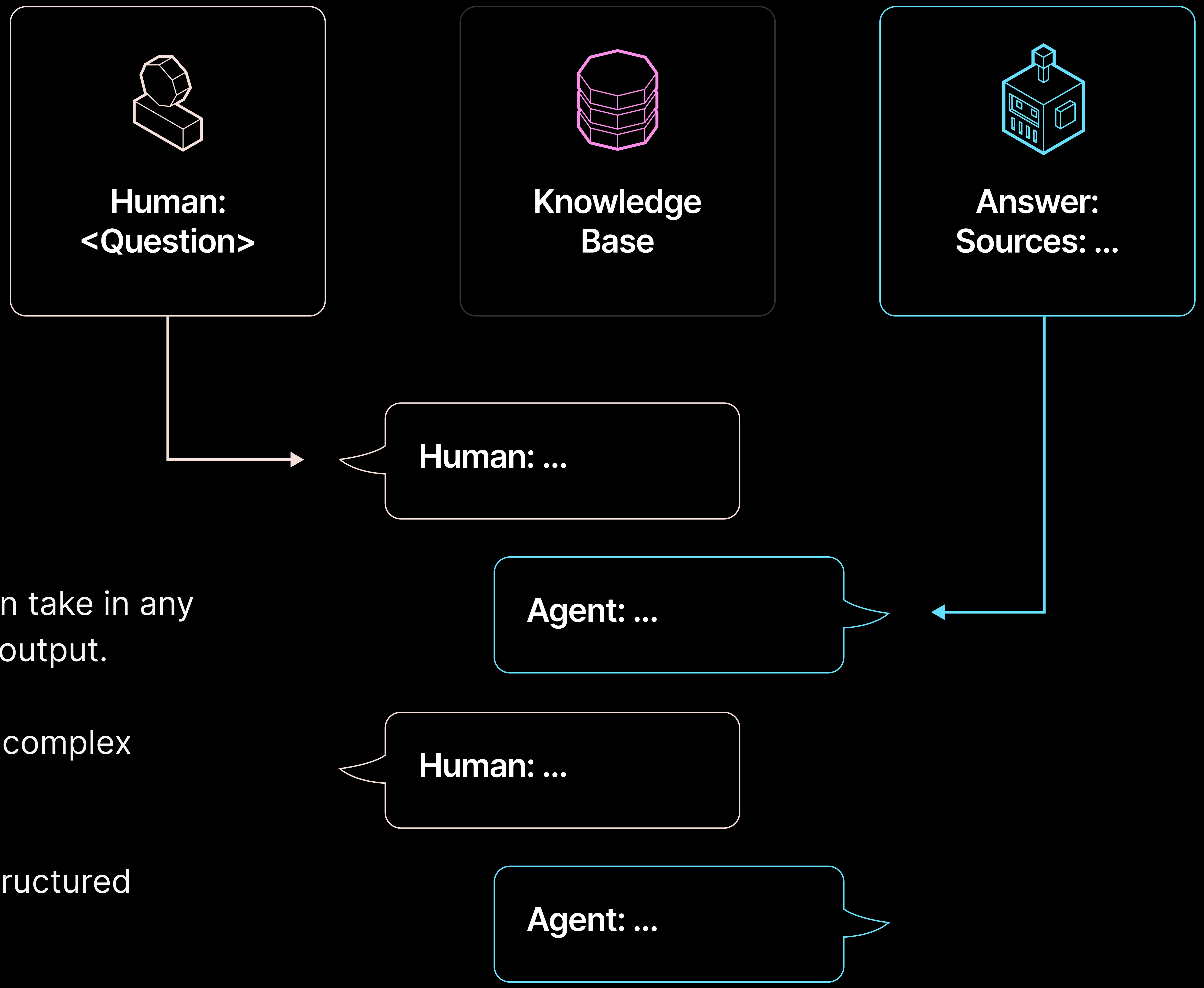
LlamaCloud: A centralized knowledge interface for your production LLM application.

Link: https://cloud.llamaindex.ai/



Red Hat
OPENBB
replit
Berkeley UNIVERSITY OF CALIFORNIA
DATASTAX
T Systems
Weights & Biases
adyen
Jasper
GUNDERSON DETTMER
springworks
Softtek
Clearbit

Building a Knowledge Assistant



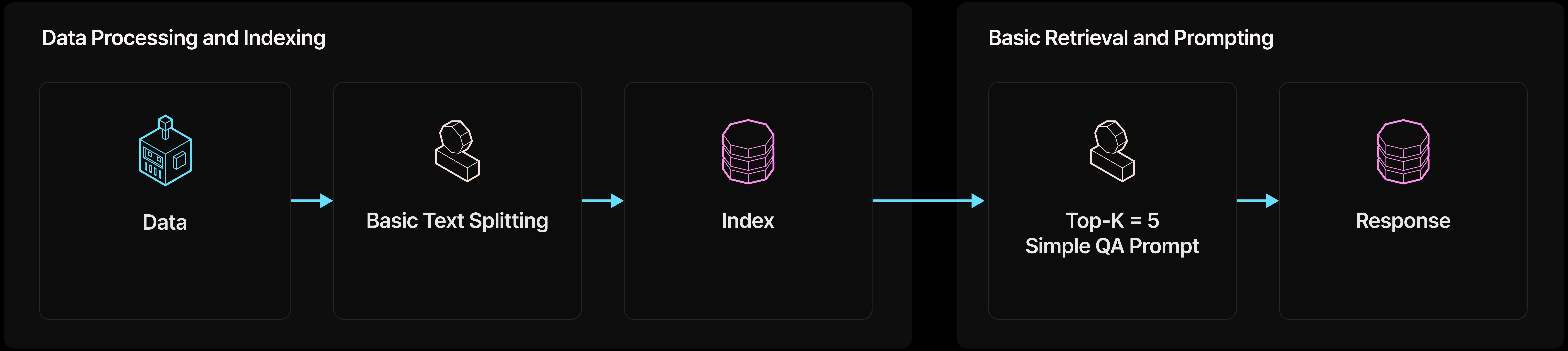
Goal: Build an interface that can take in any task as input and give back an output.

Input forms: simple questions, complex questions, research tasks

Output forms: short answer, structured output, research report

Knowledge Assistant with Basic RAG

⚠ Naive data processing, primitive retrieval interface
⚠ Poor query understanding/planning
⚠ No function calling or tool use
⚠ Stateless, no memory



Can we do more?

There's many questions/tasks that naive RAG can't give an answer to

- 🚫 Hallucinations
- 🚫 Limited time savings
- 🚫 Limited decision-making enhancement

How do we aim to build a production-ready knowledge assistant?

A Better Knowledge Assistant

1. **High-quality Multimodal RAG**
2. Complex output generation
3. Agentic reasoning over complex inputs
4. Towards a scalable, full-stack application

A Better Knowledge Assistant

1. High-quality Multimodal RAG
2. **Complex output generation**
3. Agentic reasoning over complex inputs
4. Towards a scalable, full-stack application

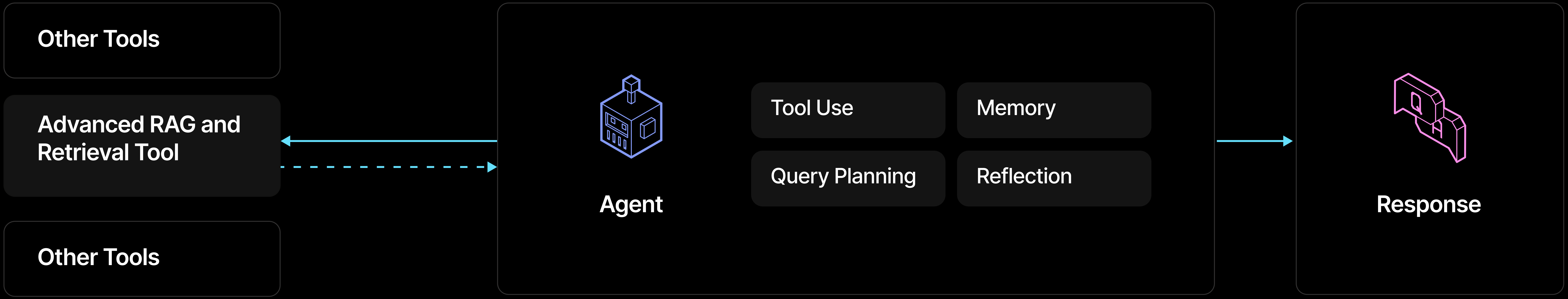
Report Generation

Data Analysis

Action-Taking

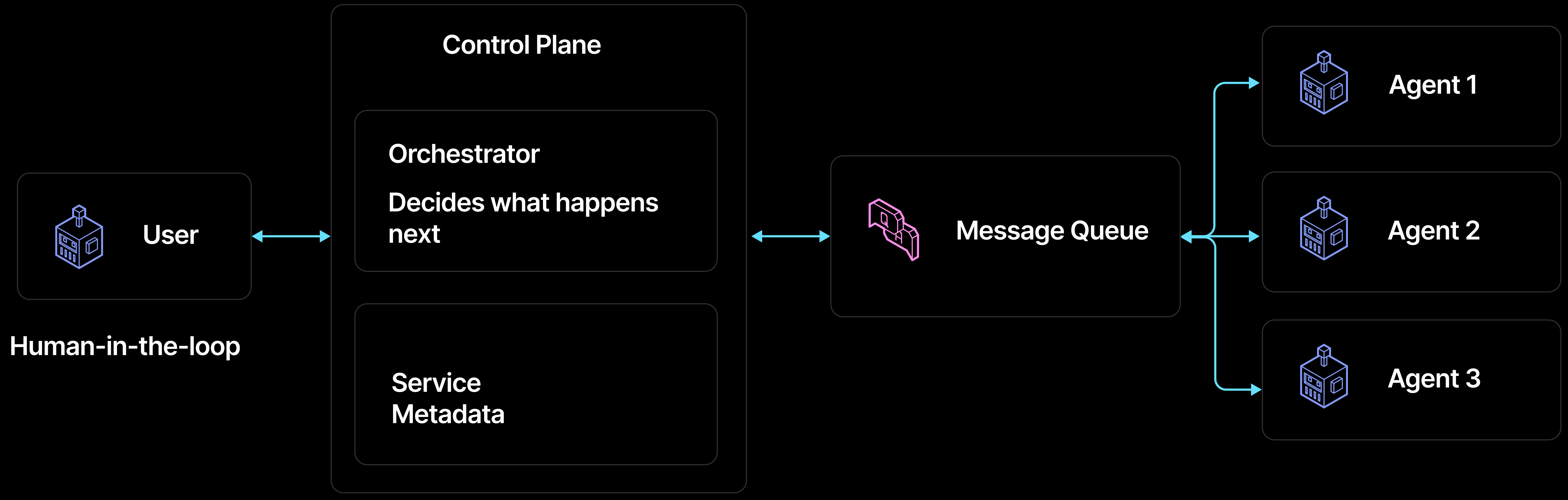
A Better Knowledge Assistant

1. High-quality Multimodal RAG
2. Complex output generation
3. **Agentic reasoning over complex inputs**
4. Towards a scalable, full-stack application



A Better Knowledge Assistant

1. High-quality Multimodal RAG
2. Complex output generation
3. Agentic reasoning over complex inputs
4. **Towards a scalable, full-stack application**

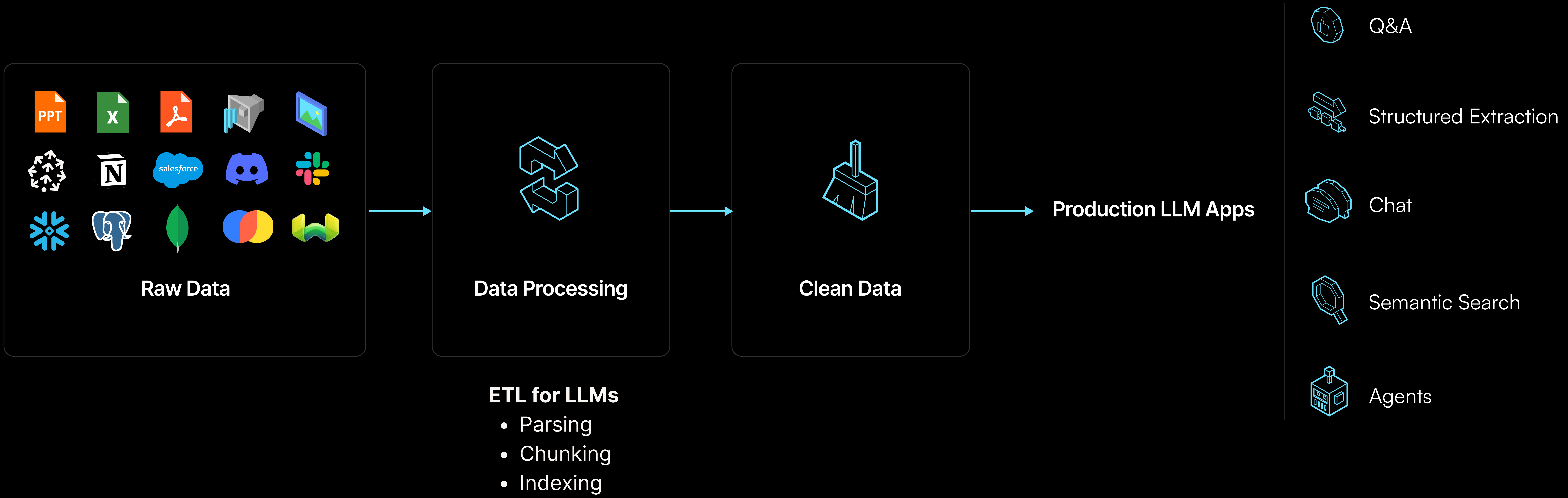


Setting up Multimodal RAG

Any LLM App is only as Good as your Data

Garbage in = garbage out

Good data quality is a **necessary** component of any production LLM app.



Case Study: Complex Documents

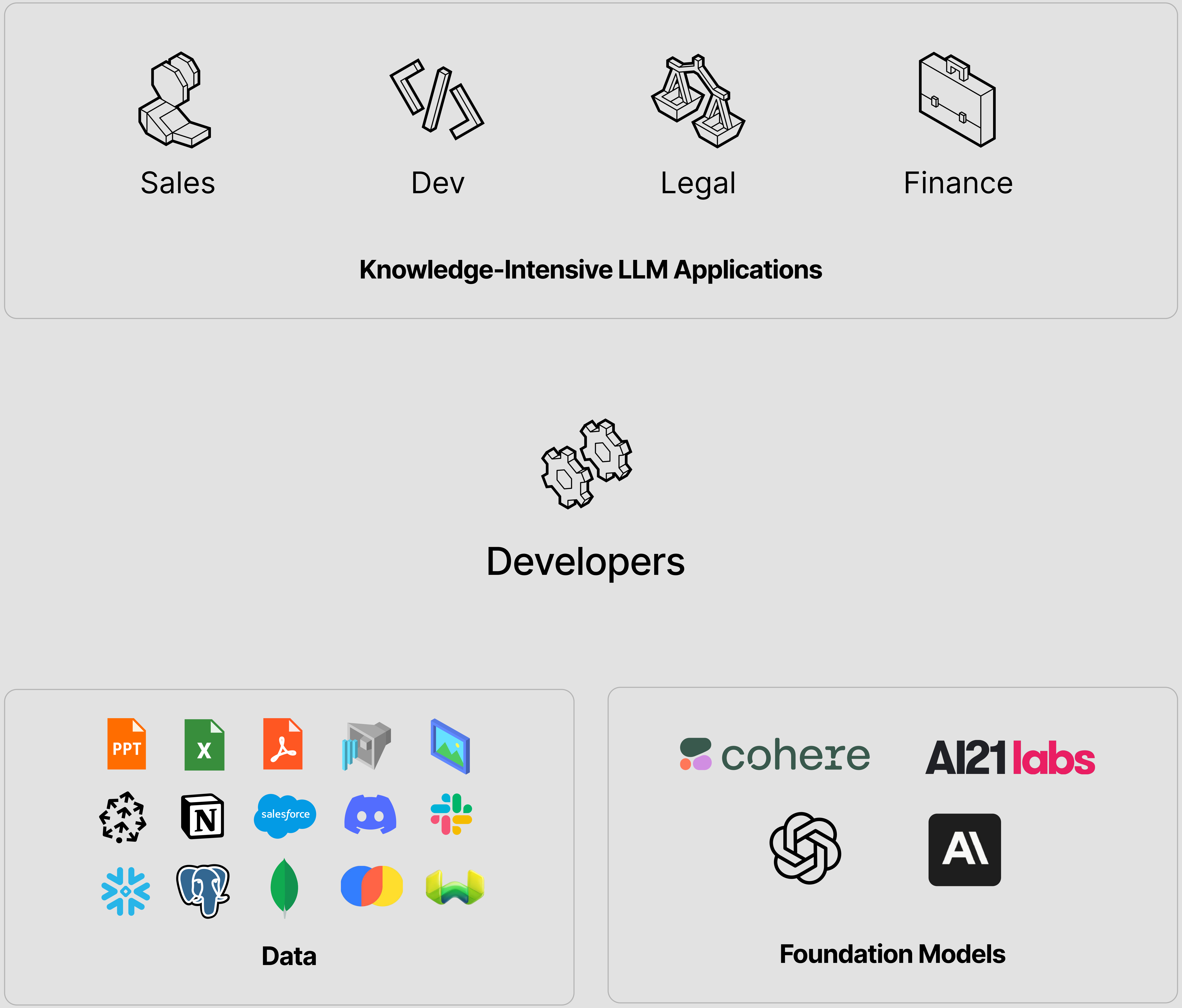
A lot of documents can be classified as **complex**:

- Embedded Tables, Charts, Images
- Irregular Layouts
- Headers/Footers

Users want to ask research questions over this data:

- Simple pointed questions
- Multi-document comparisons
- Research tasks

Building a production-ready knowledge assistants over this complex data is **challenging**.



An LLM-Native Document Parser

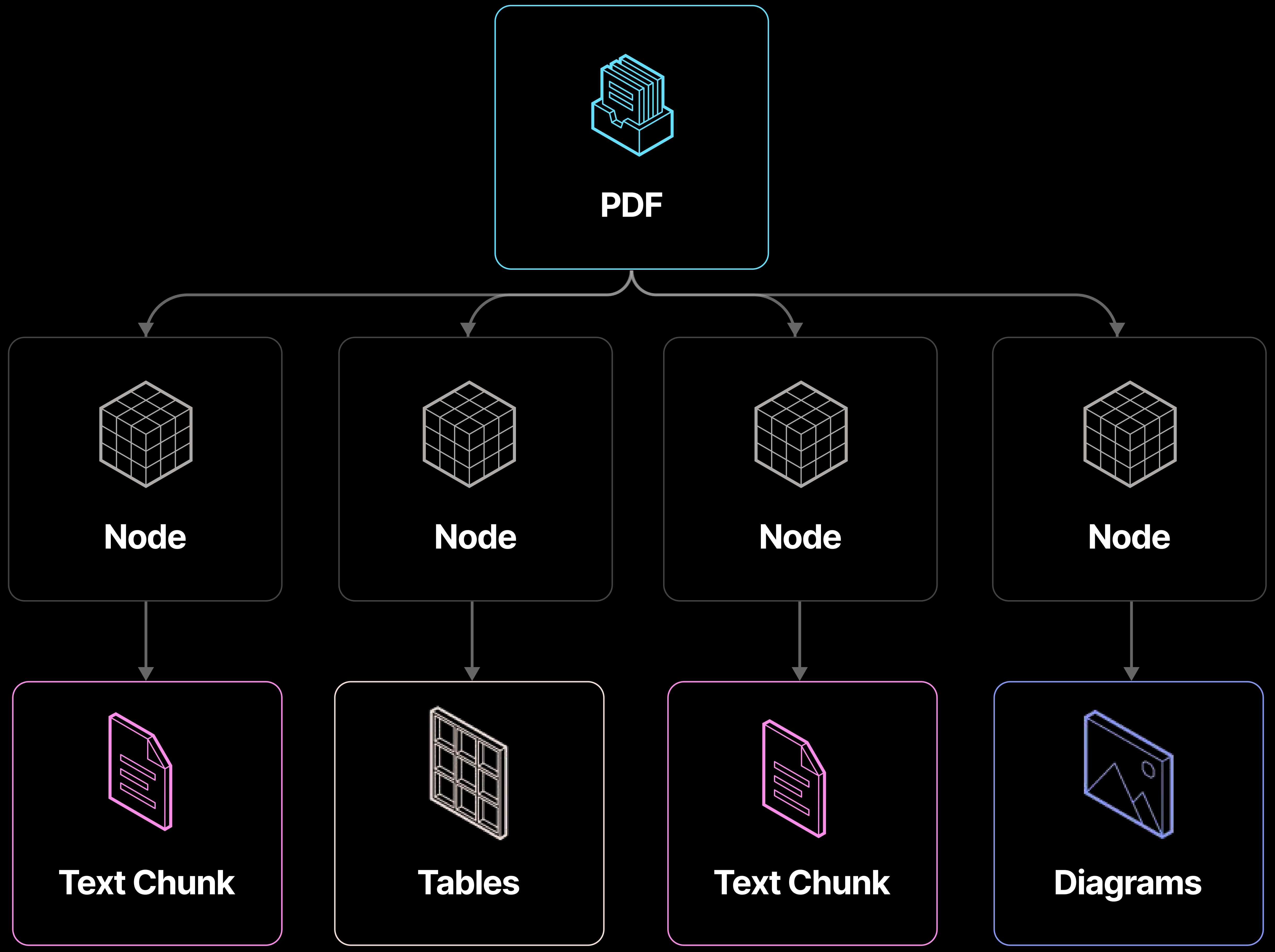
An ideal **GenAI-native parser** can structure complex document data for any downstream use case.

Requirements

- Parse tables accurately into text and semi-structured representations
- Parse text into semantically coherent chunks
- Extract visual elements (images/diagrams/charts) into structured formats *and* return image chunks.
- Automated metadata extraction

Non-Requirements

- Extract detailed JSONs for every element
- Extract bounding boxes



"As an AI Applied Data Scientist who was granted one of the first ML patents in the U.S., and who is building cutting-edge AI capabilities at one of the world's largest Private Equity Funds, I can confidently say that LlamaParse from LlamaIndex is currently the best technology I have seen for parsing complex document structures for Enterprise RAG pipelines. Its ability to preserve nested tables, extract challenging spatial layouts, and images is key to maintaining data integrity in advanced RAG and agentic model building."

Dean Barr, Applied AI Lead at Carlyle

LlamaParse

Advanced document parser specifically for reducing LLM hallucinations

20k+

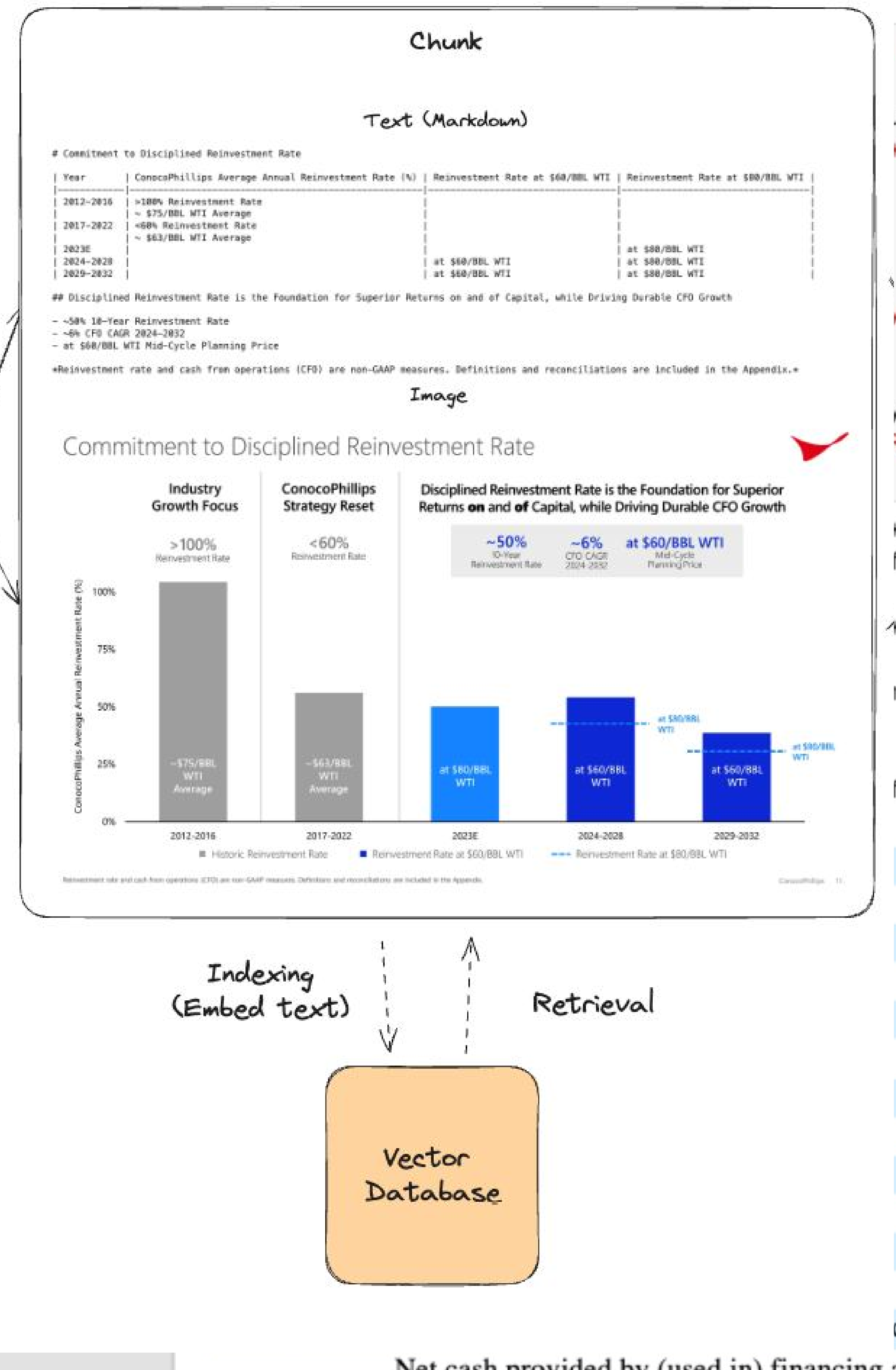
unique users

25M+

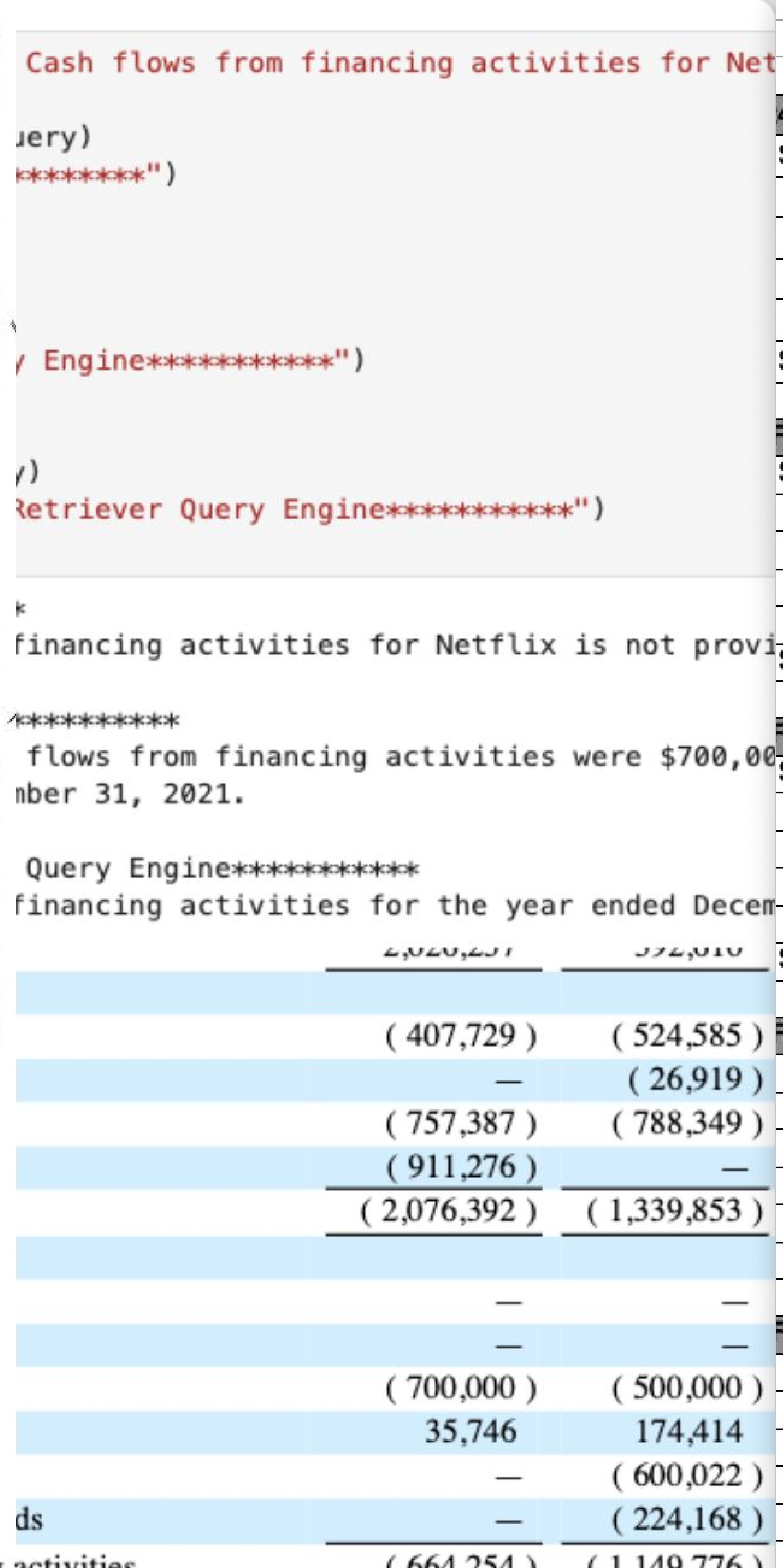
pages processed

Use Cases

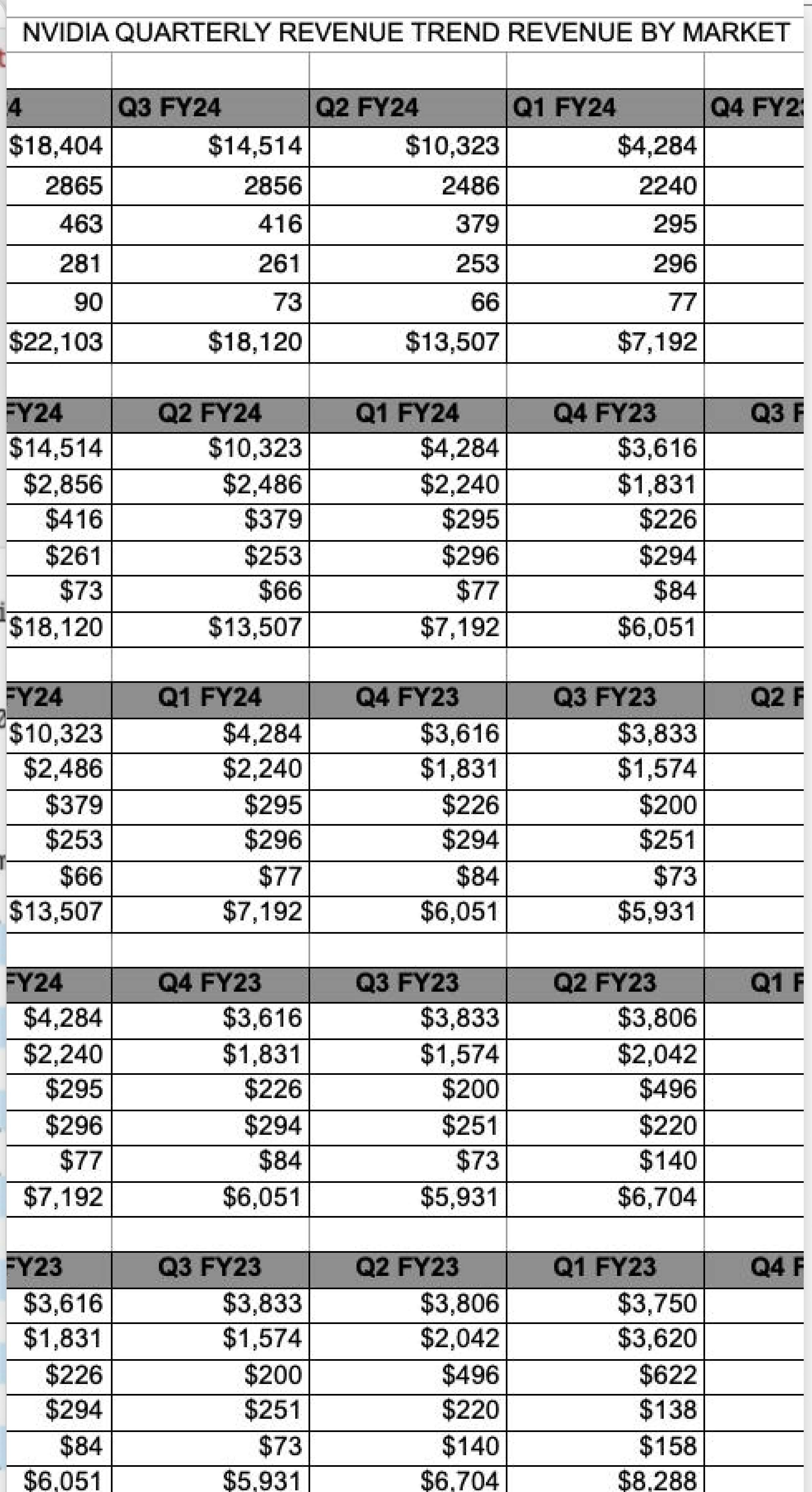
Multimodal RAG



Annual Reports (Tables)



Excel Sheets



NVIDIA QUARTERLY REVENUE TREND REVENUE BY MARKET

4	Q3 FY24	Q2 FY24	Q1 FY24	Q4 FY2
$18,404	$14,514	$10,323	$4,284	
2865	2856	2486	2240	
463	416	379	295	
281	261	253	296	
90	73	66	77	
$22,103	$18,120	$13,507	$7,192	

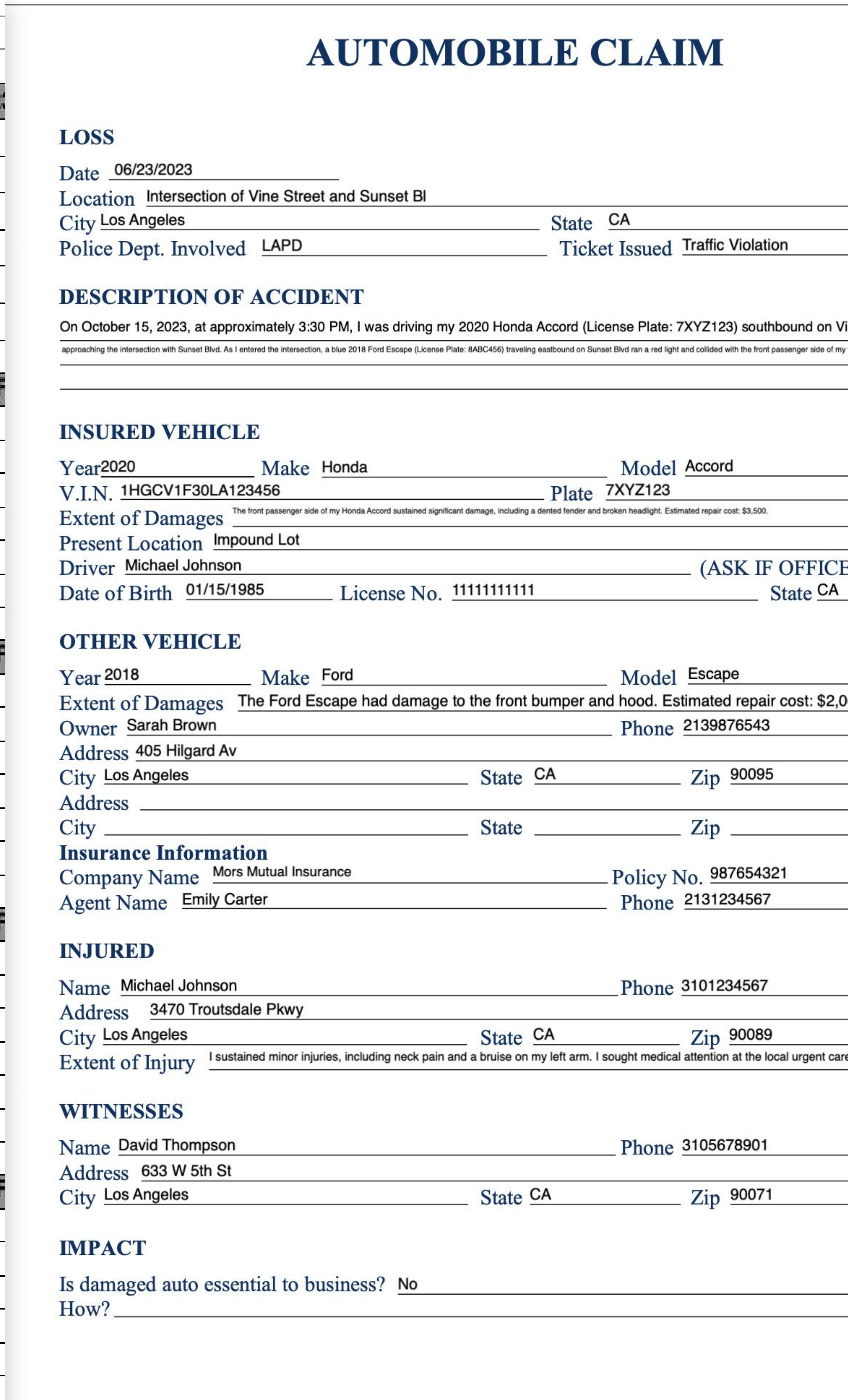
FY24	Q2 FY24	Q1 FY24	Q4 FY23	Q3 F
$14,514	$10,323	$4,284	$3,616	
$2,856	$2,486	$2,240	$1,831	
$416	$379	$295	$226	
$261	$253	$296	$294	
$73	$66	$77	$84	
$18,120	$13,507	$7,192	$6,051	

FY24	Q1 FY24	Q4 FY23	Q3 FY23	Q2 F
$10,323	$4,284	$3,616	$3,833	
$2,486	$2,240	$1,831	$1,574	
$379	$295	$226	$200	
$253	$296	$294	$251	
$66	$77	$84	$73	
$13,507	$7,192	$6,051	$5,931	

FY24	Q4 FY23	Q3 FY23	Q2 FY23	Q1 F
$4,284	$3,616	$3,833	$3,806	
$2,240	$1,831	$1,574	$2,042	
$295	$226	$200	$496	
$296	$294	$251	$220	
$77	$84	$73	$140	
$7,192	$6,051	$5,931	$6,704	

FY23	Q3 FY23	Q2 FY23	Q1 FY23	Q4 F
$3,616	$3,833	$3,806	$3,750	
$1,831	$1,574	$2,042	$3,620	
$226	$200	$496	$622	
$294	$251	$220	$138	
$84	$73	$140	$158	
$6,051	$5,931	$6,704	$8,288	

Forms



AUTOMOBILE CLAIM

LOSS
Date 06/23/2023
Location Intersection of Vine Street and Sunset Bl
City Los Angeles State CA
Police Dept. Involved LAPD Ticket Issued Traffic Violation

DESCRIPTION OF ACCIDENT
On October 15, 2023, at approximately 3:30 PM, I was driving my 2020 Honda Accord (License Plate: 7XYZ123) southbound on Vi

INSURED VEHICLE
Year 2020 Make Honda Model Accord
V.I.N. 1HGCV1F30LA123456 Plate 7XYZ123
Extent of Damages The front passenger side of my Honda Accord sustained significant damage, including a dented fender and broken headlight. Estimated repair cost: $3,500.
Present Location Impound Lot
Driver Michael Johnson (ASK IF OFFICE
Date of Birth 01/15/1985 License No. 11111111111 State CA

OTHER VEHICLE
Year 2018 Make Ford Model Escape
Extent of Damages The Ford Escape had damage to the front bumper and hood. Estimated repair cost: $2,0
Owner Sarah Brown Phone 2139876543
Address 405 Hilgard Av
City Los Angeles State CA Zip 90095
Address
City State Zip
Insurance Information
Company Name Mors Mutual Insurance Policy No. 987654321
Agent Name Emily Carter Phone 2131234567

INJURED
Name Michael Johnson Phone 3101234567
Address 3470 Troutsdale Pkwy
City Los Angeles State CA Zip 90089
Extent of Injury I sustained minor injuries, including neck pain and a bruise on my left arm. I sought medical attention at the local urgent care

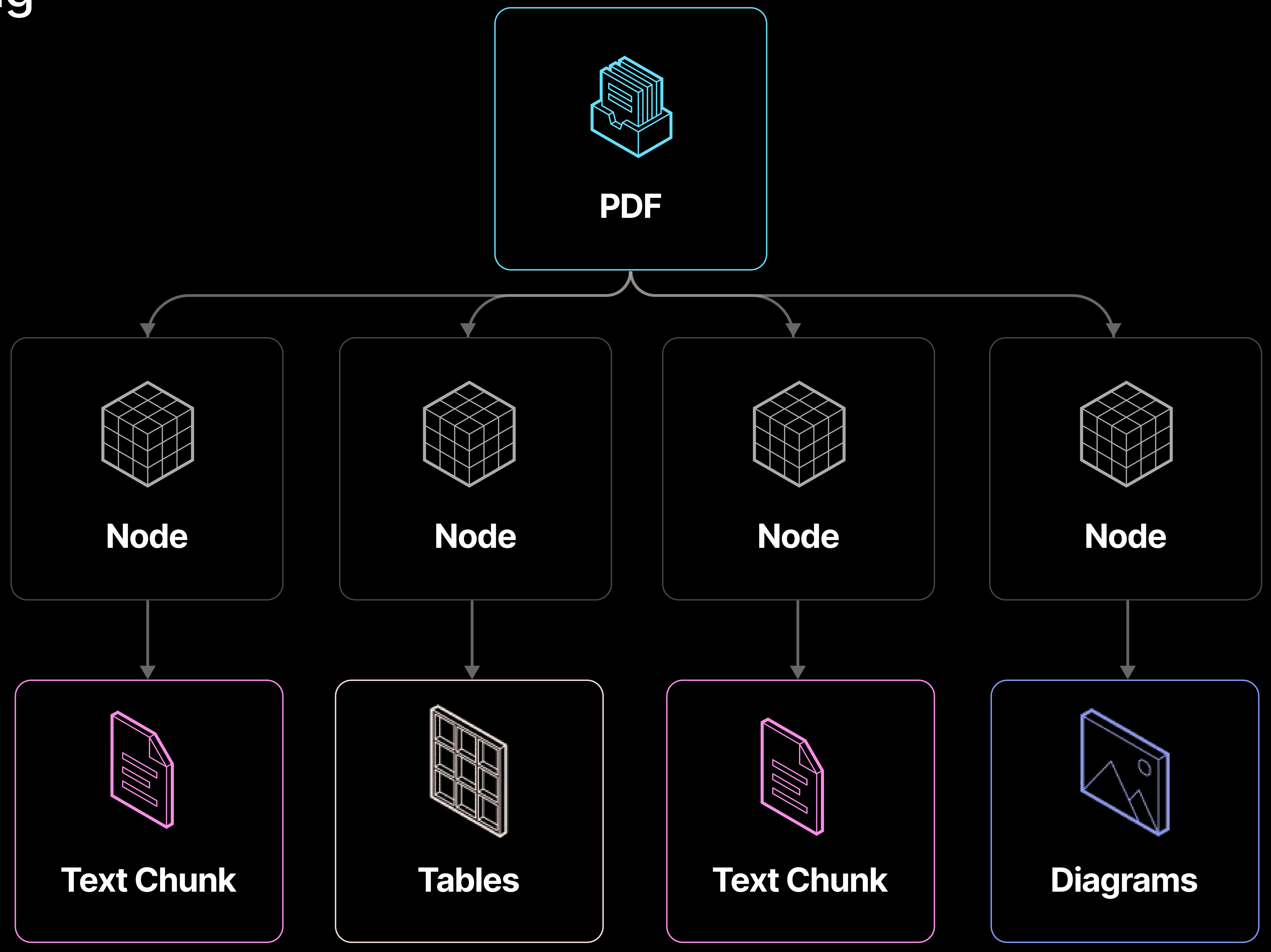
WITNESSES
Name David Thompson Phone 3105678901
Address 633 W 5th St
City Los Angeles State CA Zip 90071

IMPACT
Is damaged auto essential to business? No
How?

Advanced Parsing + Advanced Indexing

You can combine parsing with **hierarchical indexing and retrieval** to model heterogeneous unstructured/tabular/multimodal data within a document.

1. Parse documents into elements: text chunks, tables, images, and more.
2. For each element, extract **one or more** text representations that can be indexed.
3. Do **recursive retrieval**



Multimodal RAG Pipeline

A true multimodal RAG pipeline stores both text and image chunks for use within a multi-modal LLM

Indexing

1. Parse document into text and image chunks with LlamaParse
2. Link each text chunk to image chunk through metadata
3. Embed and index text chunks

Retrieval

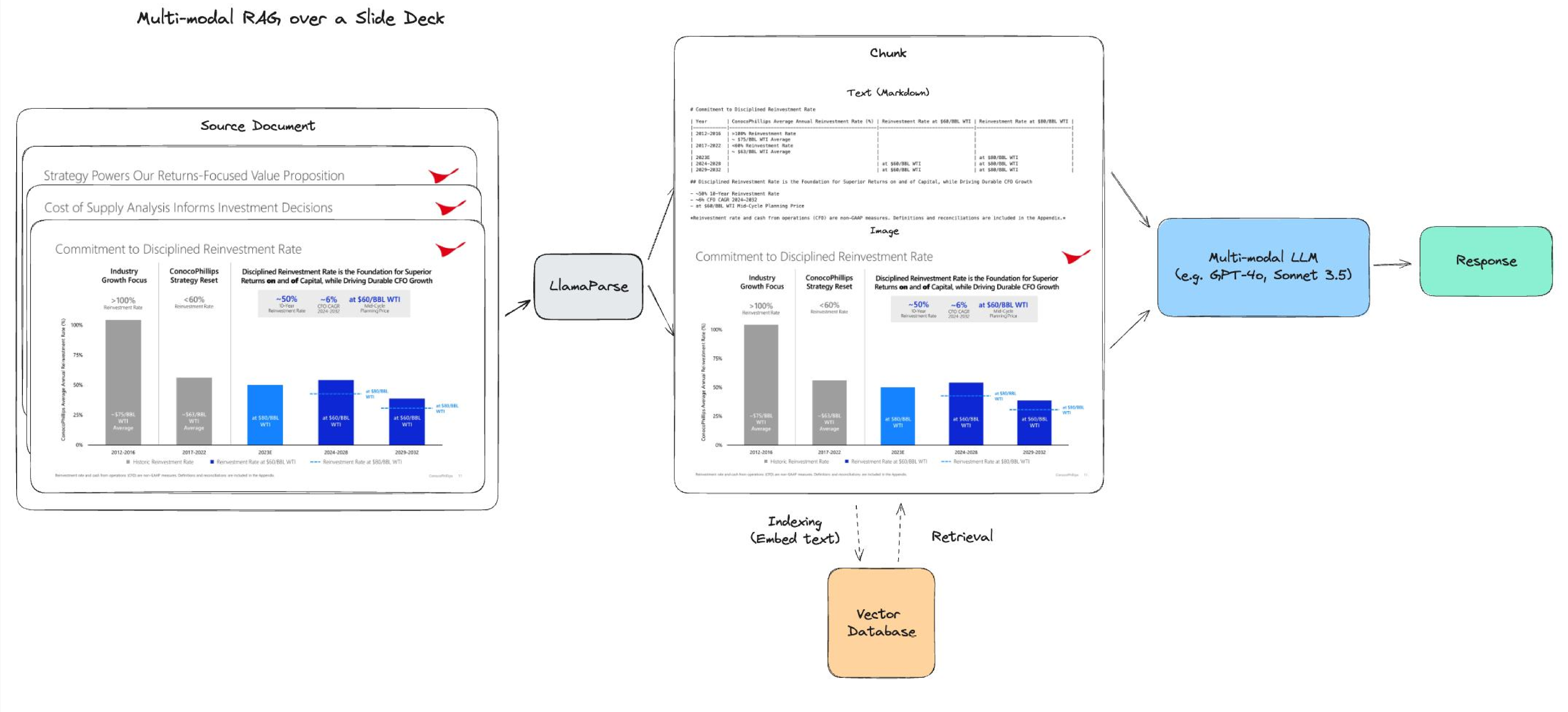
1. Retrieve text chunks by text embeddings
2. Feed in both text and image to multimodal LLM during synthesis.

Multimodal RAG Pipeline

Let's run through a demo example of building multimodal RAG over a complex slide deck!

The end result is you're able to ask questions over visual data in the document.

https://github.com/run-llama/llama_parse/blob/main/examples/multimodal/multimodal_rag_slide_deck.ipynb



LlamaCloud: An Enterprise RAG Platform

A production-ready RAG platform that allows developers to easily connect their unstructured data sources to LLM agent systems.

✅ **Instant Time-to-Value** for building knowledge assistants

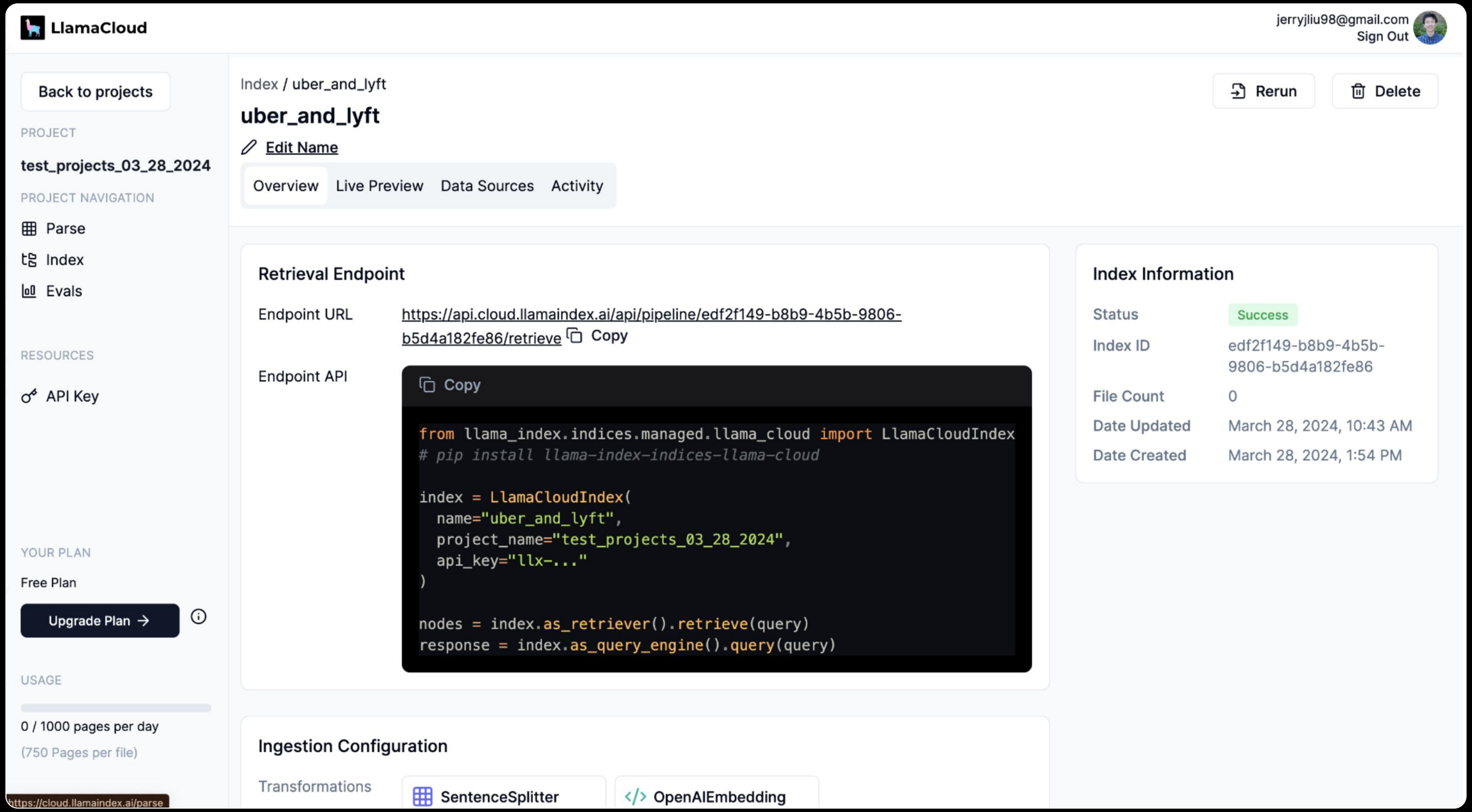
- Out-of-the-box advanced RAG capabilities
- Free up developer time to rapidly iterate on higher-level agent use cases

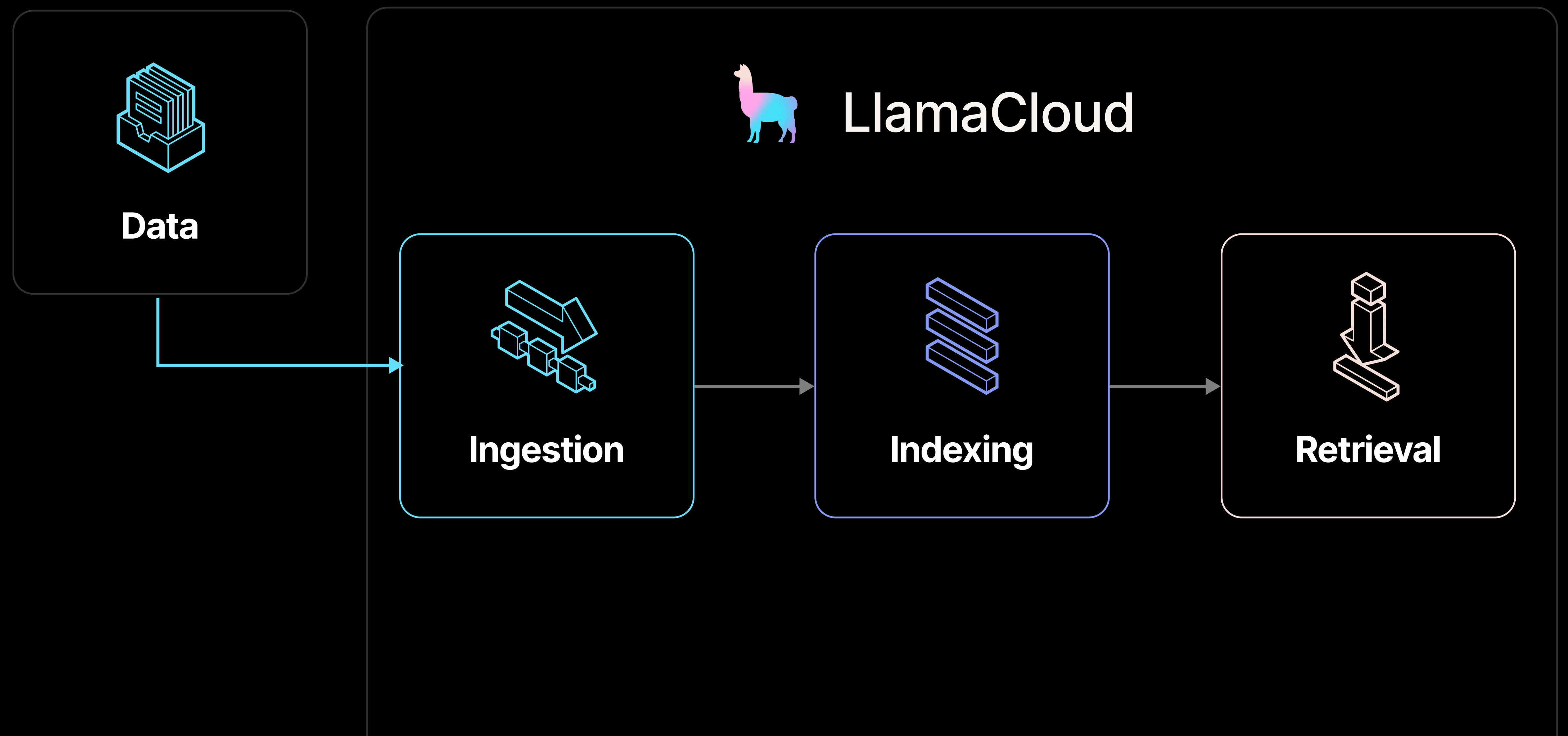
✅ **State-of-the-Performance** leads to increased satisfaction and reduced compliance risk

✅ **Reduced maintenance cost** once application is deployed

✅ **Enterprise-ready security** like access controls

Signup: https://cloud.llamaindex.ai/

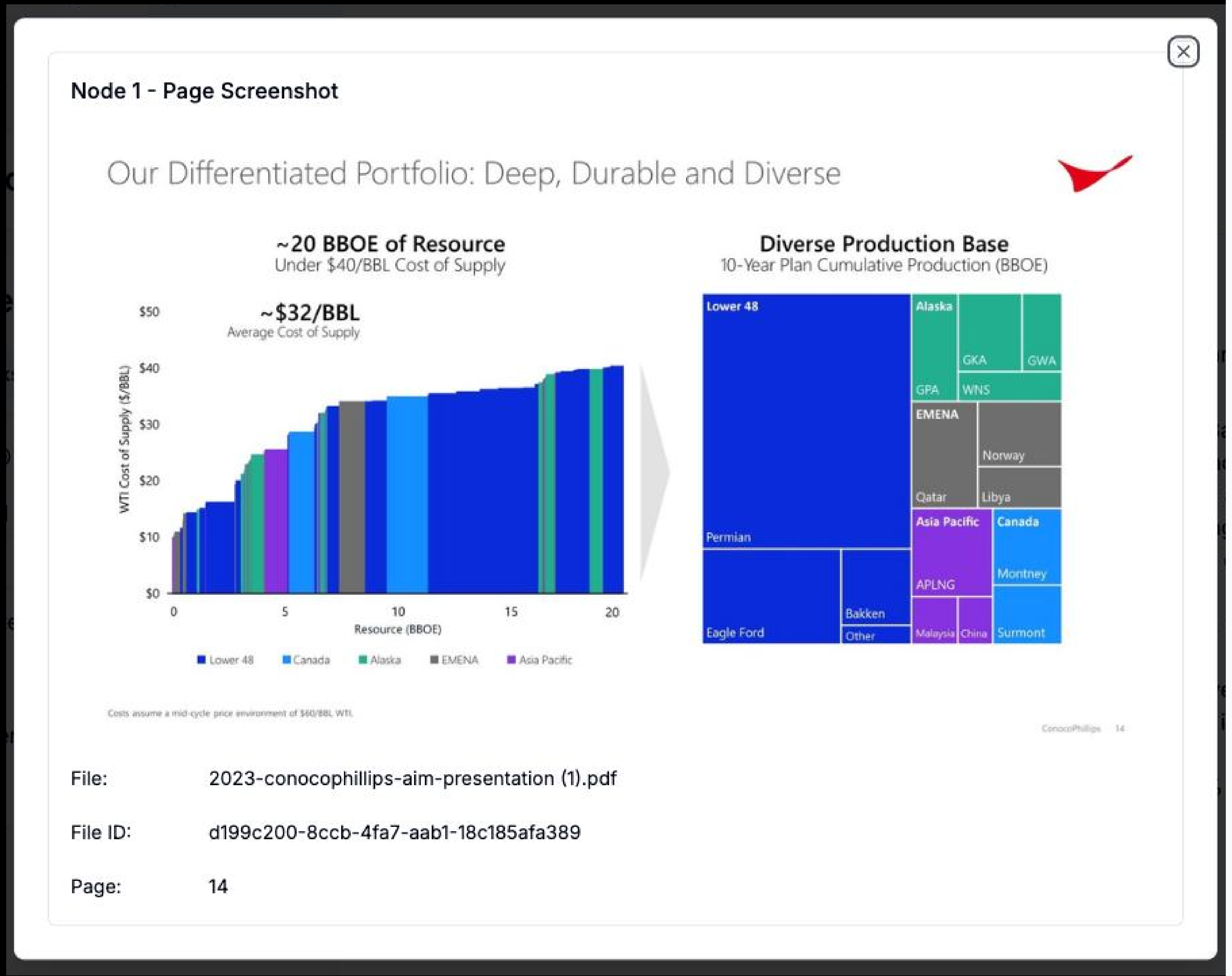


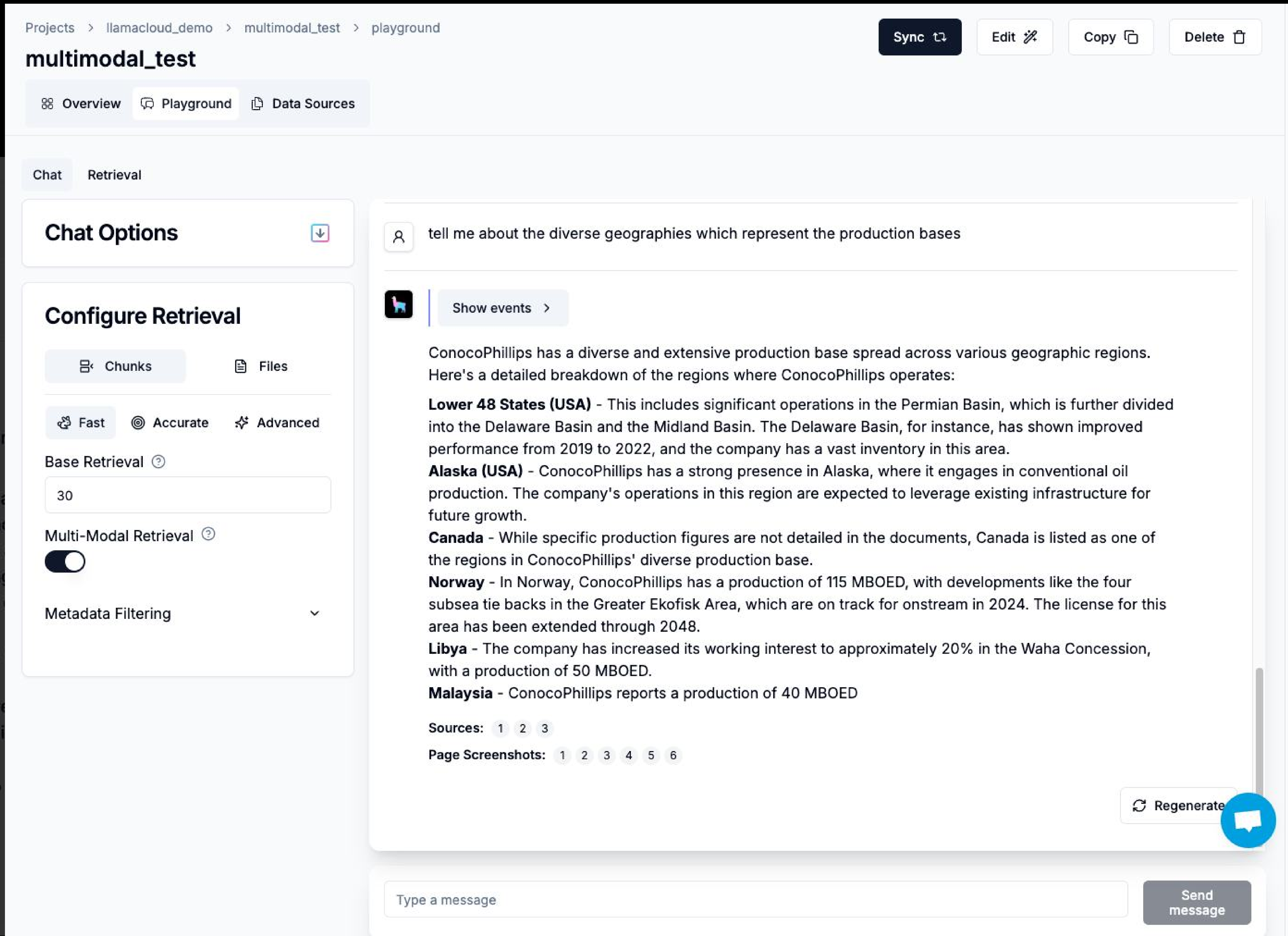


E2E Multimodal RAG Capabilities

Setup multimodal indexing and retrieval in minutes

Signup here: https://cloud.llamaindex.ai/





LlamaIndex


Playground
Data Sources
Node 1 - Page Screenshot
Our Differentiated Portfolio: Deep, Durable and Diverse
~20 BBOE of Resource
Under $40/BBL Cost of Supply
~$32/BBL
Average Cost of Supply
WTI Cost of Supply ($/BBL)
Resource (BBOE)
Lower 48
Canada
Alaska
EMENA
Asia Pacific
Diverse Production Base
10-Year Plan Cumulative Production (BBOE)
Lower 48
Permian
Eagle Ford
Bakken
Other
Alaska
GKA
GWA
GPA
WNS
EMENA
Norway
Qatar
Libya
Asia Pacific
APLNG
Malaysia
China
Canada
Montney
Surmont
Costs assume a mid-cycle price environment of $60/BBL WTI.
File: 2023-conocophillips-aim-presentation (1).pdf
File ID: d199c200-8ccb-4fa7-aab1-18c185afa389
Page: 14
Projects > llamacloud_demo > multimodal_test > playground
multimodal_test
Sync
Edit
Copy
Delete
Overview
Playground
Data Sources
Chat
Retrieval
Chat Options
Configure Retrieval
Chunks
Files
Fast
Accurate
Advanced
Base Retrieval
30
Multi-Modal Retrieval
Metadata Filtering
tell me about the diverse geographies which represent the production bases
Show events
ConocoPhillips has a diverse and extensive production base spread across various geographic regions. Here's a detailed breakdown of the regions where ConocoPhillips operates:
Lower 48 States (USA) - This includes significant operations in the Permian Basin, which is further divided into the Delaware Basin and the Midland Basin. The Delaware Basin, for instance, has shown improved performance from 2019 to 2022, and the company has a vast inventory in this area.
Alaska (USA) - ConocoPhillips has a strong presence in Alaska, where it engages in conventional oil production. The company's operations in this region are expected to leverage existing infrastructure for future growth.
Canada - While specific production figures are not detailed in the documents, Canada is listed as one of the regions in ConocoPhillips' diverse production base.
Norway - In Norway, ConocoPhillips has a production of 115 MBOED, with developments like the four subsea tie backs in the Greater Ekofisk Area, which are on track for onstream in 2024. The license for this area has been extended through 2048.
Libya - The company has increased its working interest to approximately 20% in the Waha Concession, with a production of 50 MBOED.
Malaysia - ConocoPhillips reports a production of 40 MBOED
Sources: 1 2 3
Page Screenshots: 1 2 3 4 5 6
Regenerate
Type a message
Send message

LlamaIndex

Report Generation

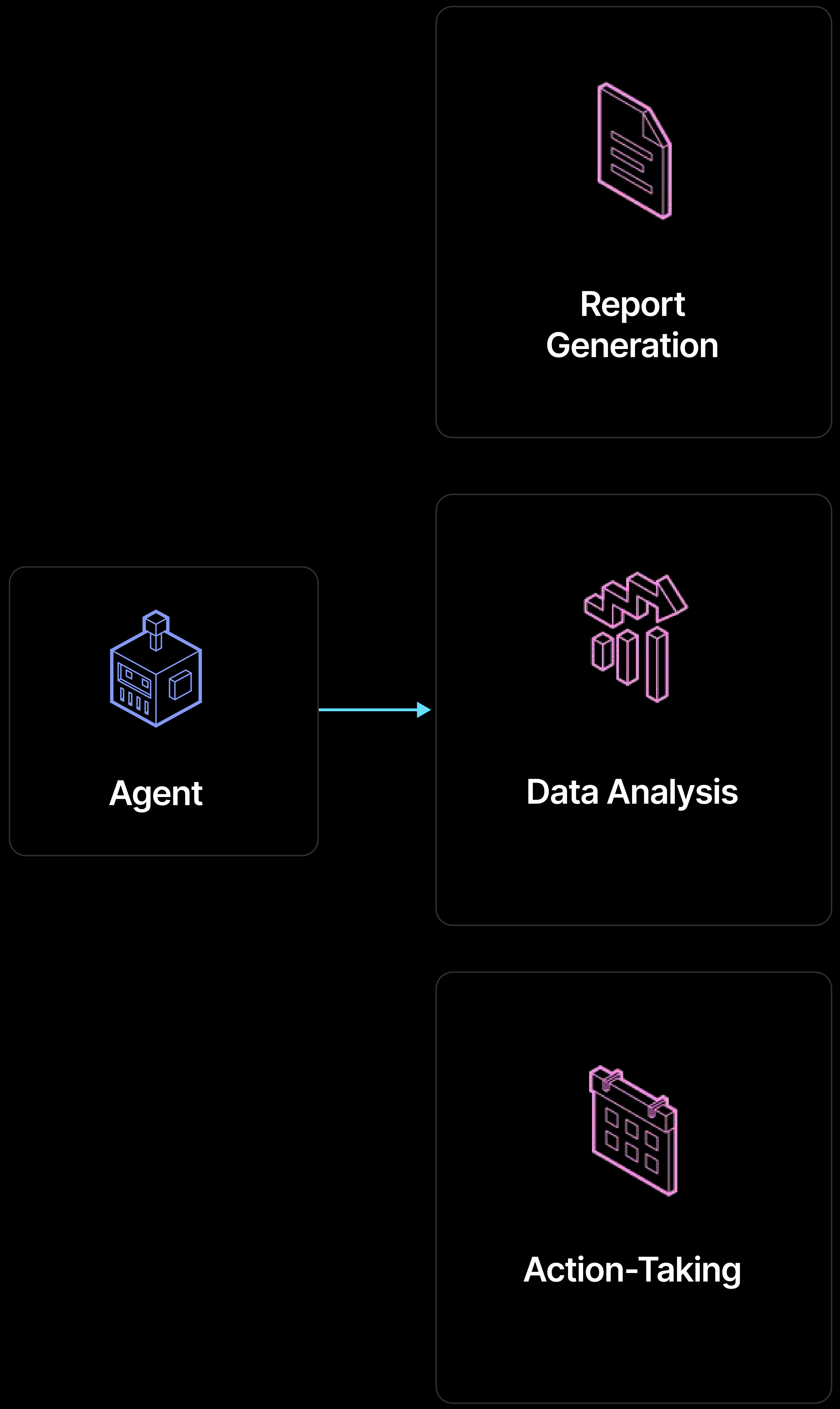
Automating Decision Making

Agents should have the capability to not only generate chatbot responses, but also

1. Produce knowledge work
2. Take actions

📈 Action-taking and Output Generation potentially lead to **much greater ROI** in terms of time savings and capability improvement

Solution 💡: Structured Outputs and Function Calling



Multimodal Report Generation

Generate interleaving text-and-image responses with the help of **structured outputs.**

https://github.com/run-llama/llama_parse/blob/main/examples/multimodal/multimodal_report_generation.ipynb

Output Schema

```
class TextBlock(BaseModel):
    text: str

class ImageBlock(BaseModel):
    file_path: str

class ReportOutput(BaseModel):
    blocks: ListBlock | ImageBlock]
```

LlamaIndex

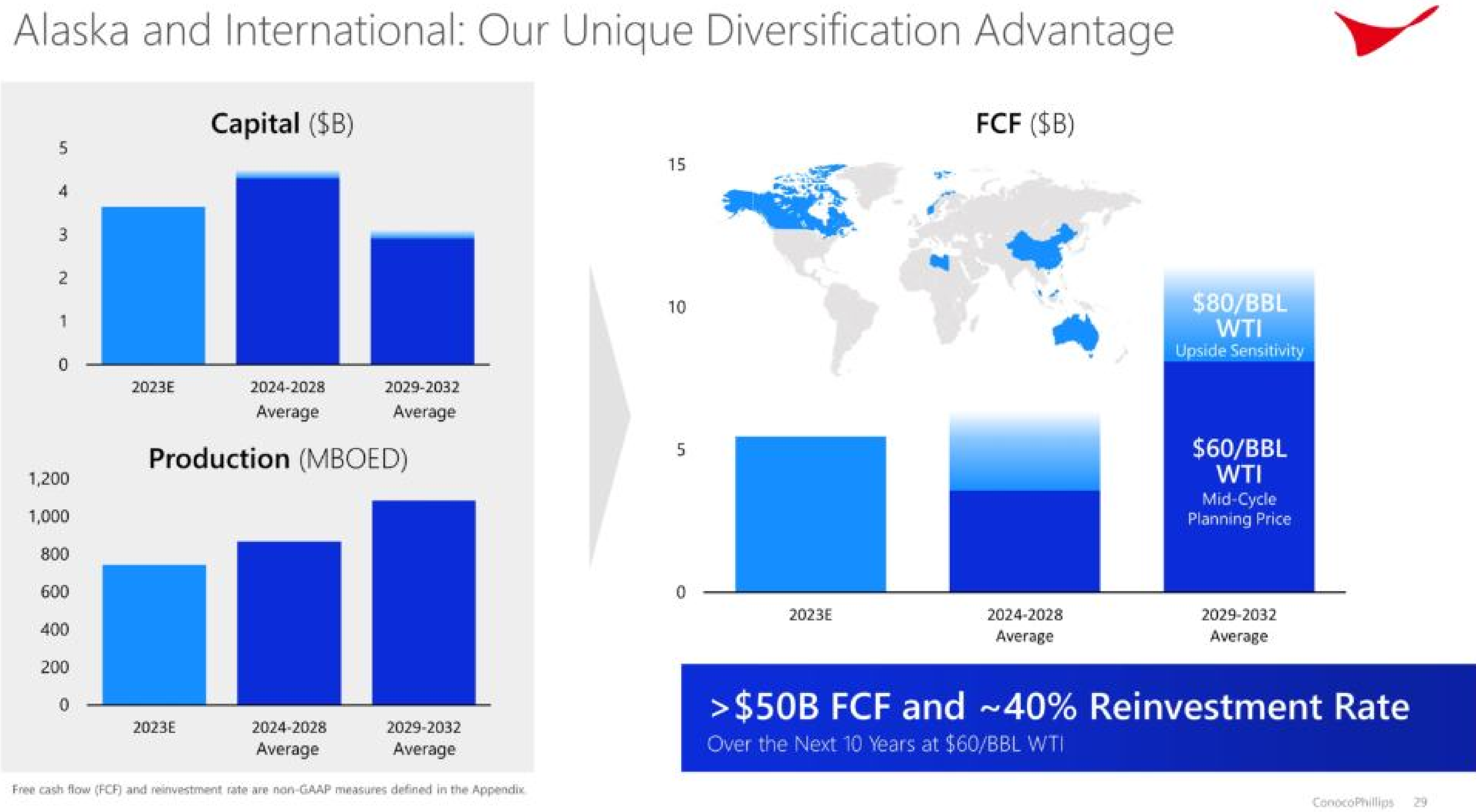
The financial performance of ConocoPhillips' Alaska/International segment and the Lower 48 segment can be summarized as follows:

Alaska/International Segment

- **Capital Expenditures**: Expected to average $3.7 billionin 2023,$4.4 billion from 2024-2028, and $3.0 billion from 2029-2032.
- **Production**: Projected to be around 750 MBOED in 2023, increasing to 870 MBOED on average from 2024-2028, and reaching 1080 MBOED on average from 2029-2032.
- **Free Cash Flow (FCF)**: Estimated at $5.5 billionin 2023, averaging$6.5 billion from 2024-2028, and $15.0 billion from 2029-2032.
- **Key Projects**: Includes significant investments in LNG, Surmont, Montney, and conventional international assets.

Lower 48 Segment

- **Capital Expenditures**: Expected to be $6.3 billionin 2023, averaging$6.5 billion from 2024-2028, and $8.1 billion from 2029-2032.
- **Production**: Projected to be around 1050 MBOED in 2023, increasing to 1220 MBOED on average from 2024-2028, and reaching 1530 MBOED on average from 2029-2032.
- **Free Cash Flow (FCF)**: Estimated at $7 billionin 2023, averaging$5.5 billion from 2024-2028, and $8 billion from 2029-2032.
- **Key Projects**: Focused on the Permian Basin, Eagle Ford, and Bakken, with significant investments in technology and emissions reductions.

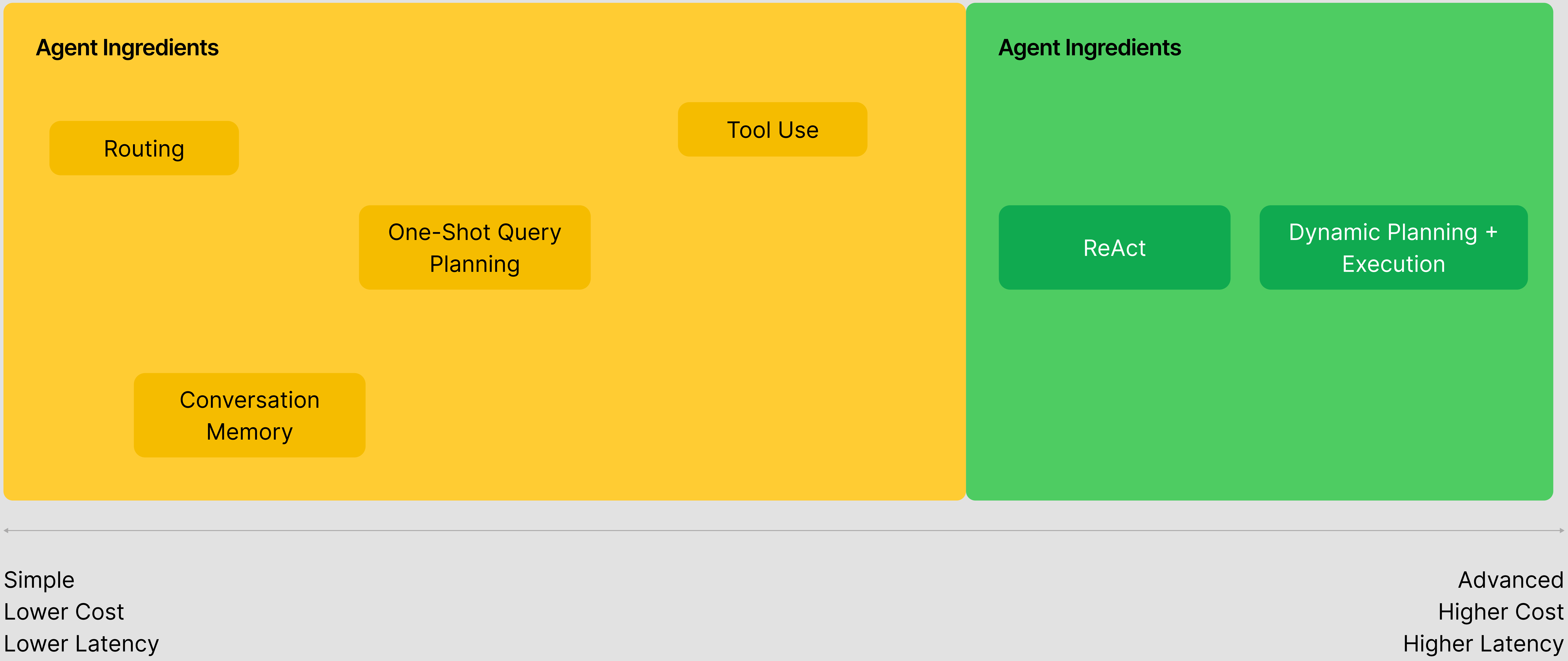


Agentic Reasoning over Complex Inputs

Naive RAG works well for pointed questions, but fails on more complex tasks.

Summarization Questions: “Give me a summary of the entire <company> 10K annual report”
Comparison Questions: “Compare the open-source contributions of candidate A and candidate B”
Multi-part Questions: “Tell me about the pro-X arguments in article A, and tell me about the pro-Y arguments in article B, make a table based on our internal style guide, then generate your own conclusion based on these facts.”
Research Tasks: “I want to create a research survey on current supervised fine-tuning techniques. Can you help?”

From Simple to Advanced Agents

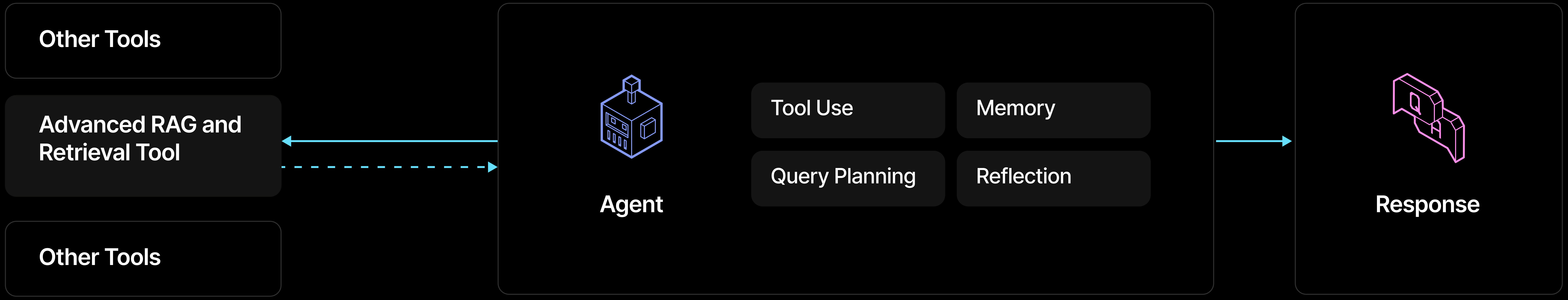


Agentic RAG

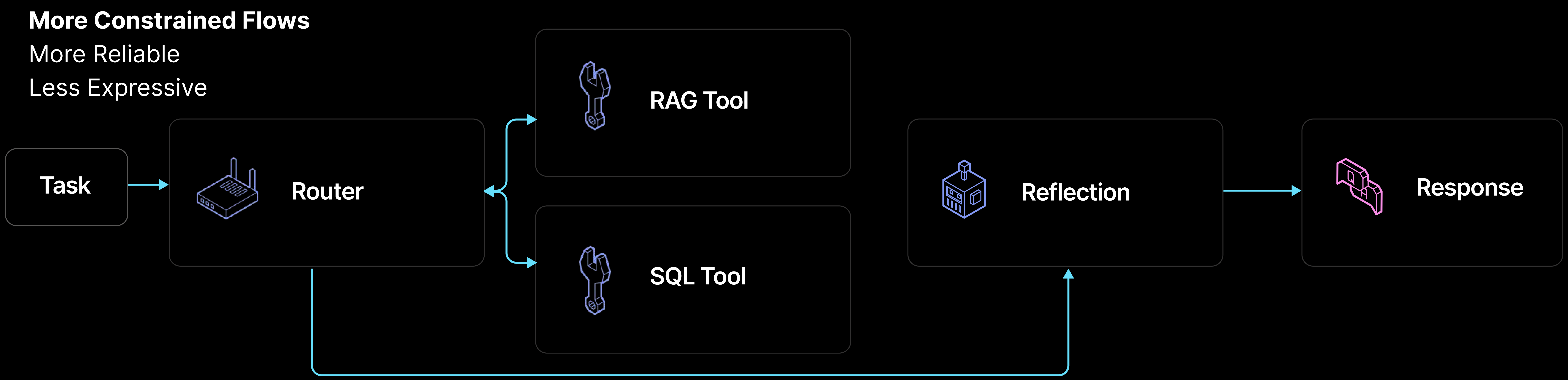
Every data interface is a tool

Use agent reasoning loops (sequential, DAG, tree) to tackle complex tasks

End Result: Build personalized QA systems capable of handling complex questions!

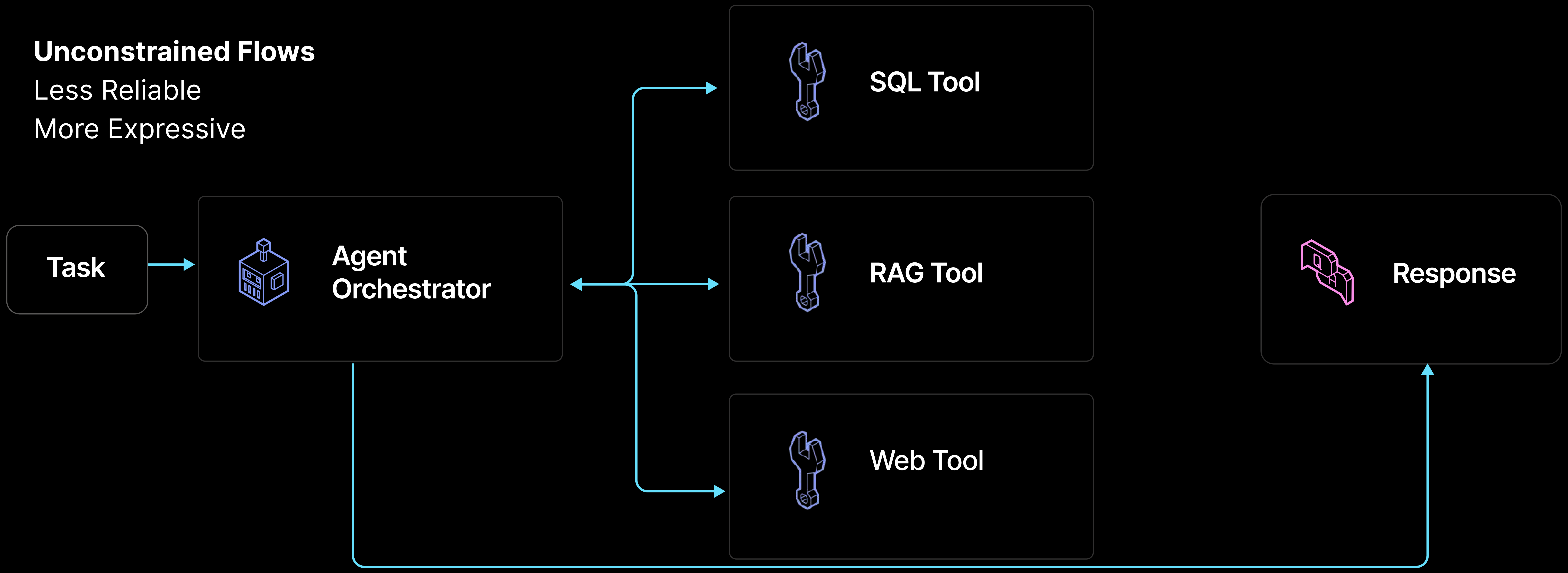


Unconstrained vs. Constrained Flows



Unconstrained vs. Constrained Flows

Unconstrained Flows
Less Reliable
More Expressive

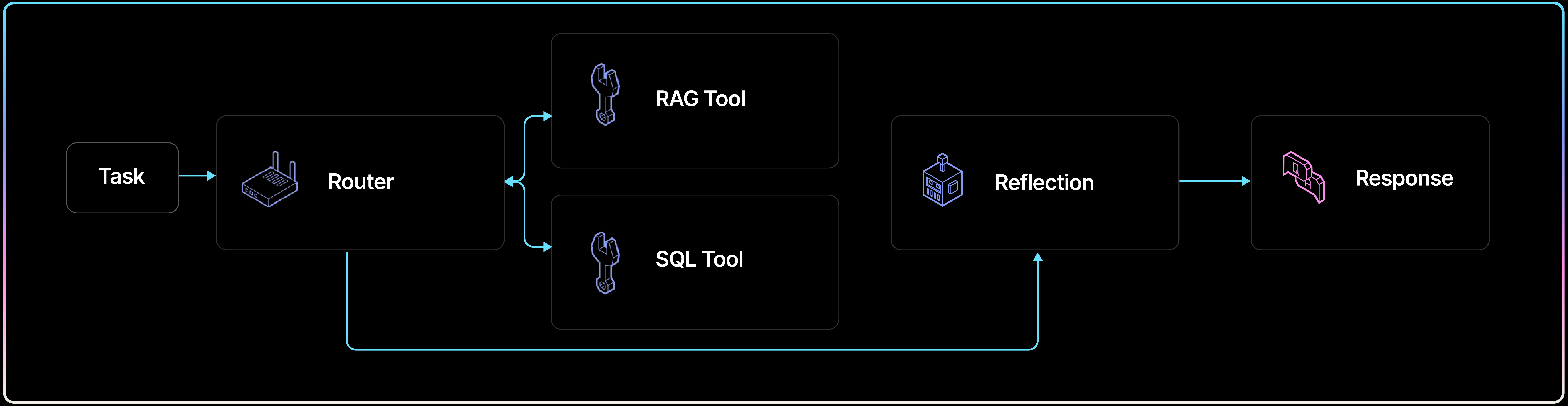


Agentic Orchestration Foundations

We believe an agent orchestration framework should have the following properties

- ✅ **Event-Driven:** Model each step as listening to input events and emitting output events
- ✅ **Composable:** Piece together granular workflows into higher-level workflows
- ✅ **Flexible:** Write logic through LLM calls or through plain Python
- ✅ **Code-first:** Express orchestration logic through code. Easy to read and easy to extend.
- ✅ **Debuggable and Observable:** Step through and observe states
- ✅ **Easily Deployable to Production:** Translate notebook code into services that run in production.

LlamaIndex Workflows



Compared to Graph-based Approaches

Graph-based approaches (e.g. our deprecated Query Pipelines) can be cumbersome and non-Pythonic for complex agentic workflows.

```
def generate_response(context, query):
    prompt = f"Question: {query}\n\nContext: {context}\n\nAnswer:"
    response = llm.complete(prompt)
    return response.text

# Define the pipeline
pipeline = QueryPipeline()
pipeline.add_modules({
    "input": InputComponent(),
    "retriever": retriever,
    "reranker": reranker,
    "response_generator": FnComponent(fn=generate_response)
})

# Define the flow
pipeline.add_link("input", "retriever")
pipeline.add_link("retriever", "reranker")
pipeline.add_link("input", "response_generator", dest_key="query")
pipeline.add_link(
   "reranker", "response_generator", dest_key="context"
)

# Run the pipeline
response = pipeline.run("What is the capital of France?")
print(response)
```

- Orchestration logic baked into edges
- More lines of code, less readable
- Cumbersome to dynamically generate workflows based on runtime conditions

Compared to Graph-based Approaches

Graph-based approaches (e.g. our deprecated Query Pipelines) can be cumbersome and non-Pythonic for complex agentic workflows.
Compared to query pipelines, our workflows are more readable, and easier to maintain/scale.

```
def generate_response(context, query):
    prompt = f"Question: {query}\n\nContext: {context}\n\nAnswer:"
    response = llm.complete(prompt)
    return response.text

# Define the pipeline
pipeline = QueryPipeline()
pipeline.add_modules({
    "input": InputComponent(),
    "retriever": retriever,
    "reranker": reranker,
    "response_generator": FnComponent(fn=generate_response)
})

# Define the flow
pipeline.add_link("input", "retriever")
pipeline.add_link("retriever", "reranker")
pipeline.add_link("input", "response_generator", dest_key="query")
pipeline.add_link(
   "reranker", "response_generator", dest_key="context"
)

# Run the pipeline
response = pipeline.run("What is the capital of France?")
print(response)
```

```
class RAGWorkflow(Workflow):
    def __init__(self):
        ...

    @step
    async def retrieve(self, query: str):
        return self.retriever.retrieve(query)

    @step
    async def rerank(self, retrieved_nodes):
        return self.reranker.postprocess_nodes(retrieved_nodes)

    @step
    async def generate_response(self, query: str, context):
        prompt = f"Question: {query}\n\nContext:
{context}\n\nAnswer:"
        response = await self.llm.complete(prompt)
        return response.text

    @step
    async def run_workflow(self, query: str):
        retrieved_nodes = await self.retrieve(query)
        reranked_nodes = await self.rerank(retrieved_nodes)
        response = await self.generate_response(query,
[node.get_content() for node in reranked_nodes])
        return response
```

Benefits and Risks

📈 Action-taking and Output Generation potentially lead to **much greater ROI** in terms of time savings and capability improvement

⚠️ LLMs need to achieve a greater degree of reliability

⚠️ Action-taking requires ample human-in-the-loop to build trust.

Multimodal Report Generation

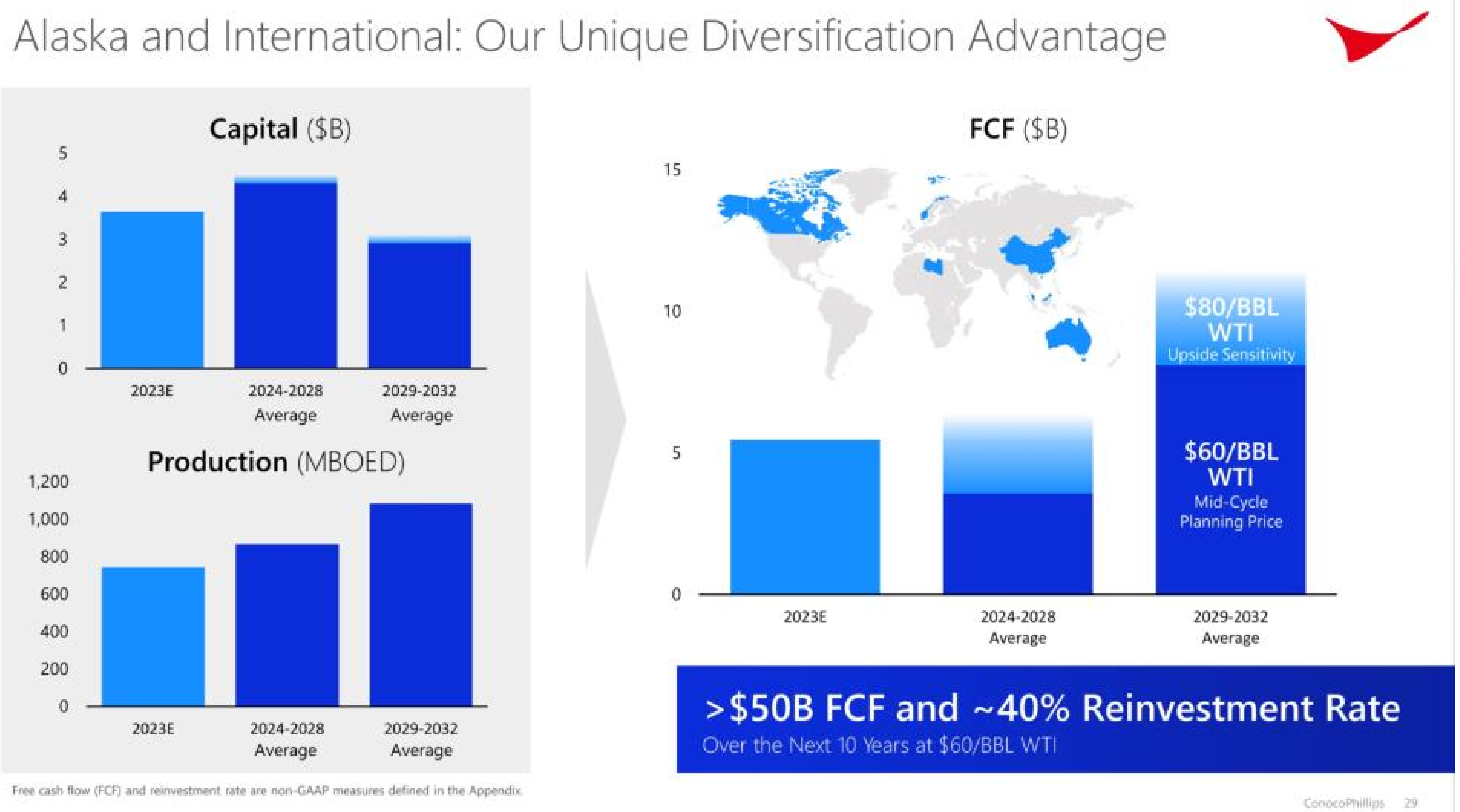
Generate interleaving text-and-image responses with the help of **structured outputs.**

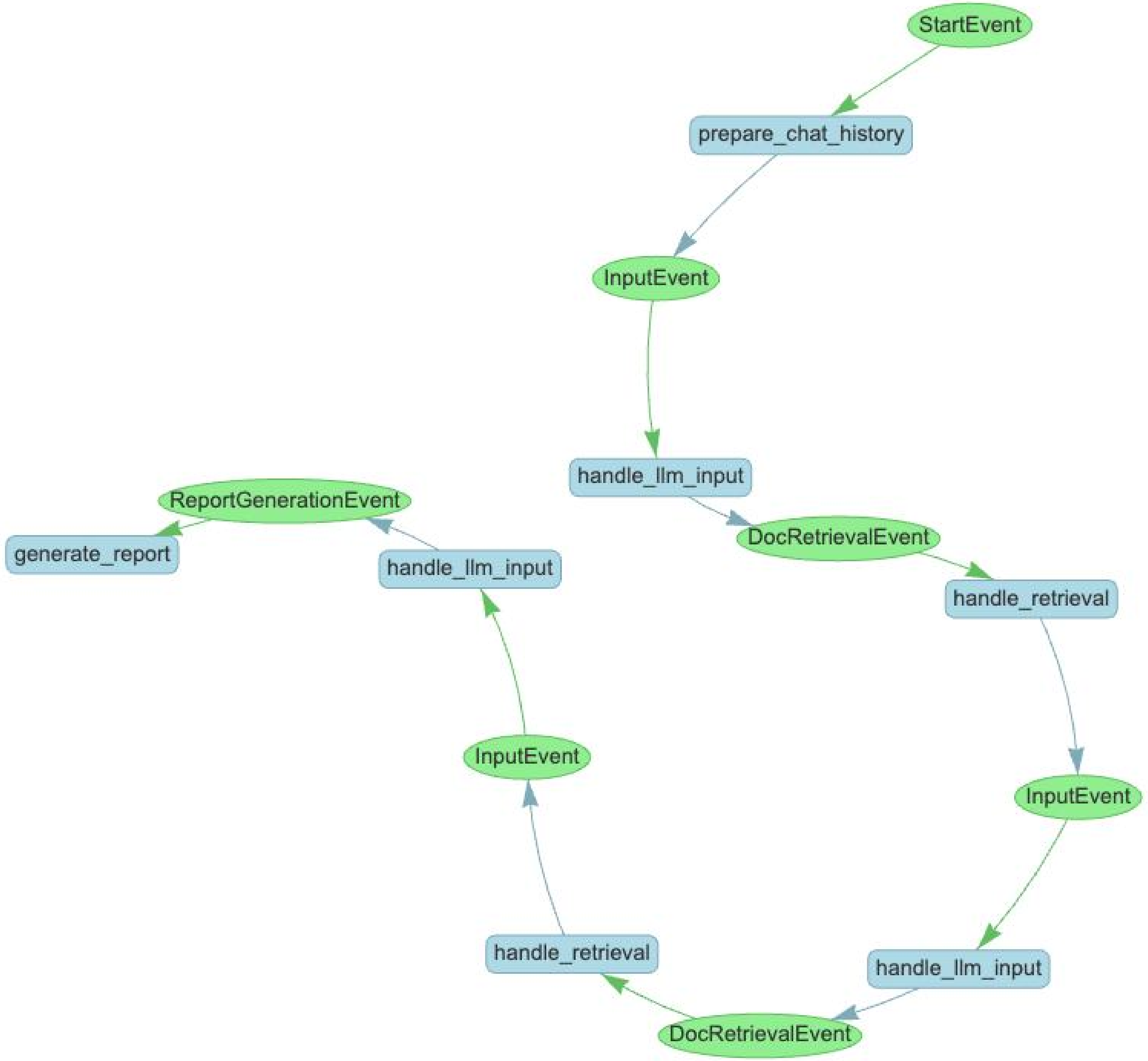
Example architecture: research and writer steps

1. The **researcher** retrieves relevant chunks and documents, and puts them into a data cache.
2. The **writer** uses the data cache to generate a structured output of interleaving text and image blocks.

Lower 48 Segment

- **Capital Expenditures**: Expected to be *6.3billionin2023, averaging*6.5 billion from 2024-2028, and $8.1 billion from 2029-2032.
- **Production**: Projected to be around 1050 MBOED in 2023, increasing to 1220 MBOED on average from 2024-2028, and reaching 1530 MBOED on average from 2029-2032.
- **Free Cash Flow (FCF)**: Estimated at *7billionin2023, averaging*5.5 billion from 2024-2028, and $8 billion from 2029-2032.
- **Key Projects**: Focused on the Permian Basin, Eagle Ford, and Bakken, with significant investments in technology and emissions reductions.





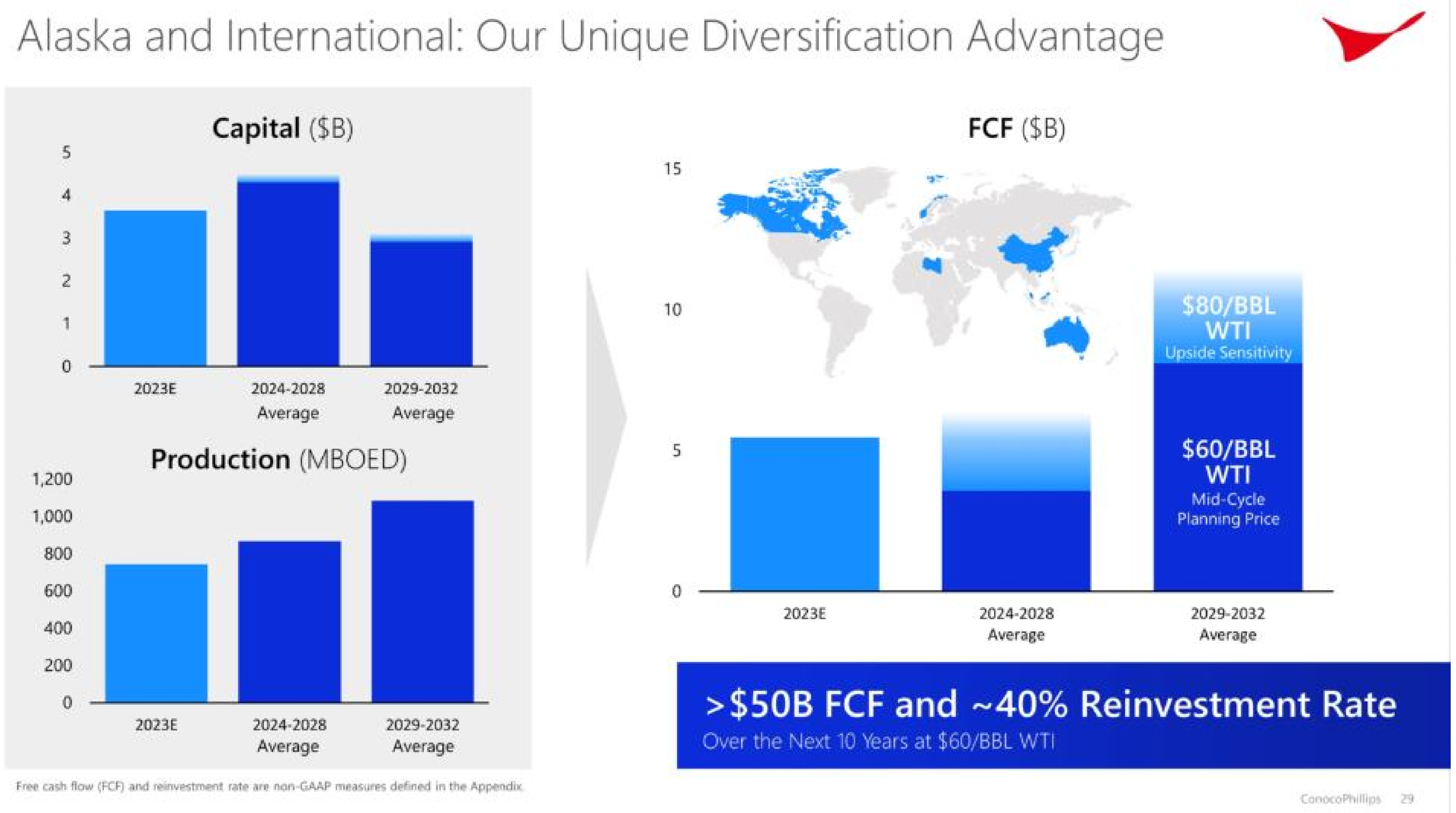
Multimodal Report Generation

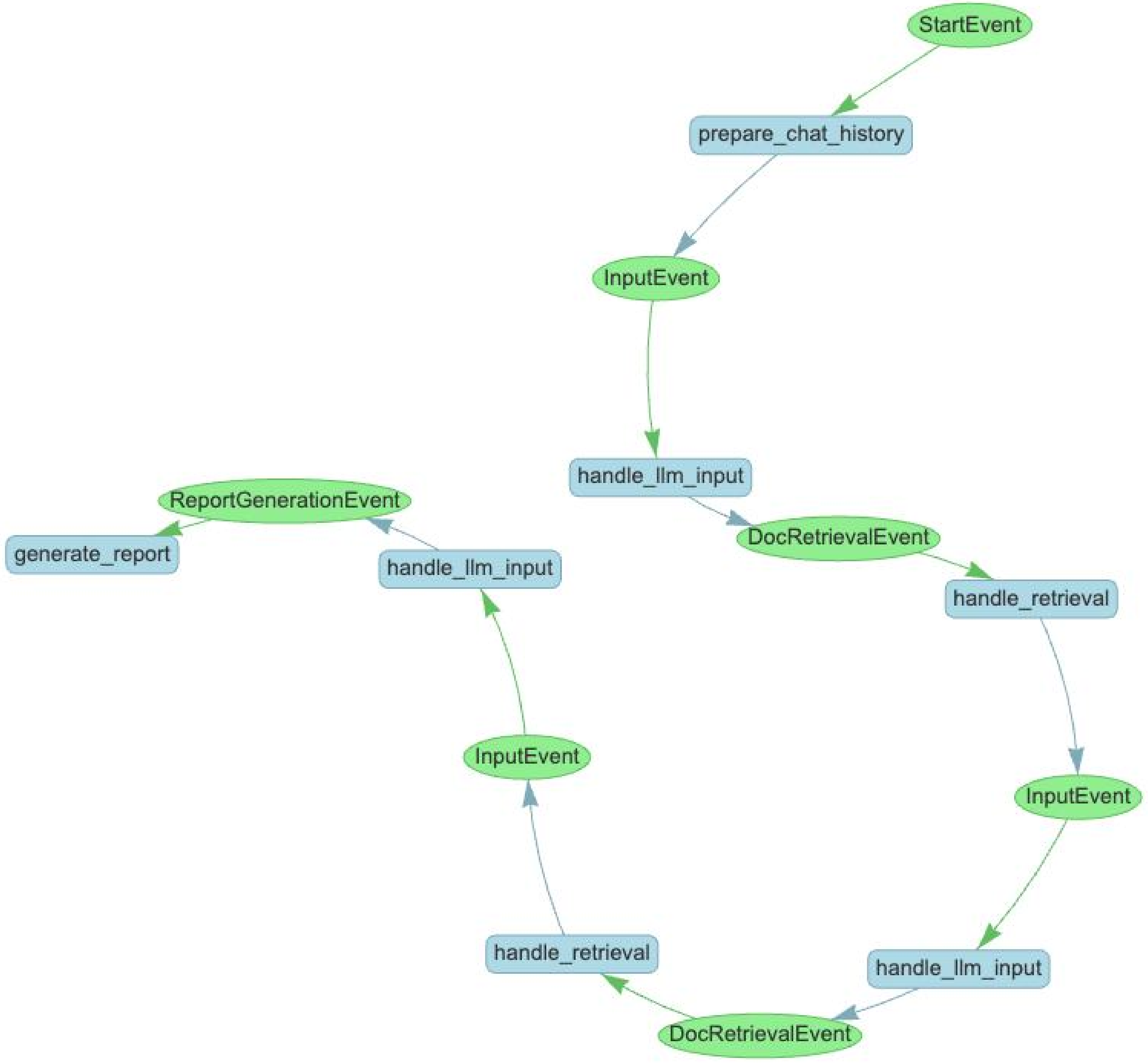
Generate interleaving text-and-image responses with the help of **structured outputs.**

https://github.com/run-llama/llama_parse/blob/main/examples/multimodal/multimodal_report_generation_agent.ipynb

Lower 48 Segment

- **Capital Expenditures**: Expected to be $6.3billionin2023, averaging$6.5 billion from 2024-2028, and $8.1 billion from 2029-2032.
- **Production**: Projected to be around 1050 MBOED in 2023, increasing to 1220 MBOED on average from 2024-2028, and reaching 1530 MBOED on average from 2029-2032.
- **Free Cash Flow (FCF)**: Estimated at $7billionin2023, averaging$5.5 billion from 2024-2028, and $8 billion from 2029-2032.
- **Key Projects**: Focused on the Permian Basin, Eagle Ford, and Bakken, with significant investments in technology and emissions reductions.





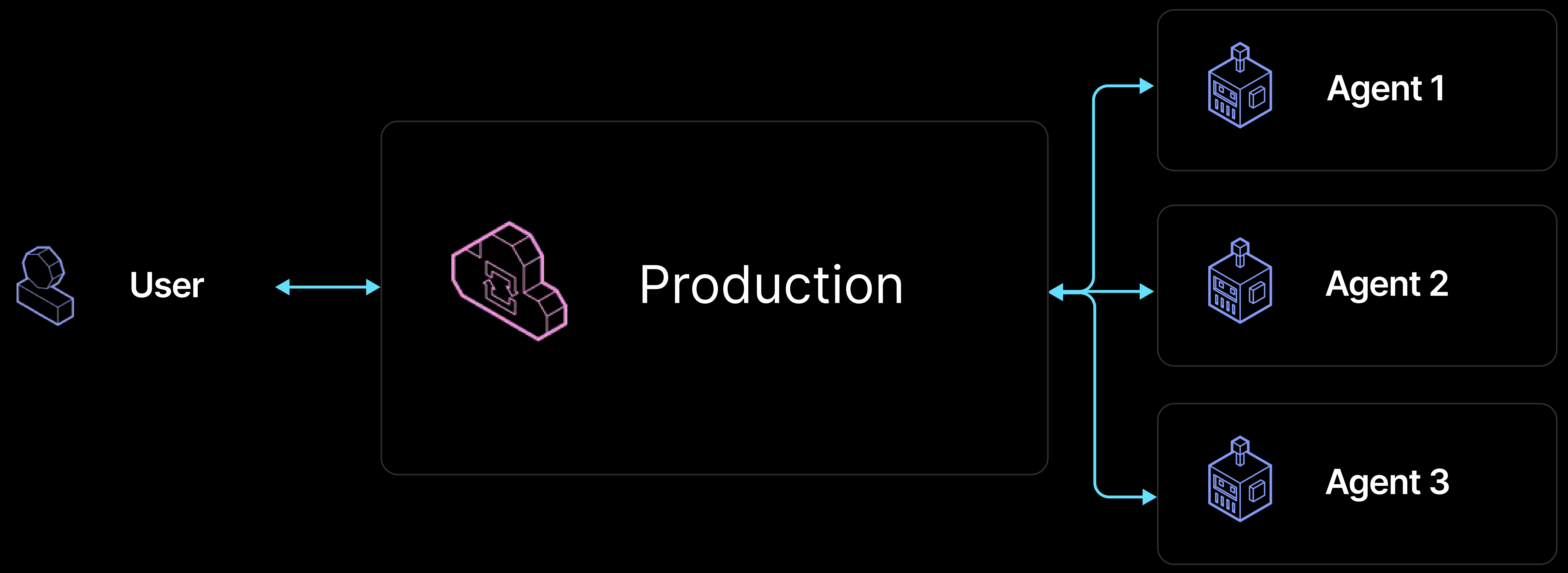
Towards a Scalable, Full-Stack Application

Running Agents in Production

You need the right architecture and infra components to serve complex, agentic workflows to end-users as a production application.

Requirements:

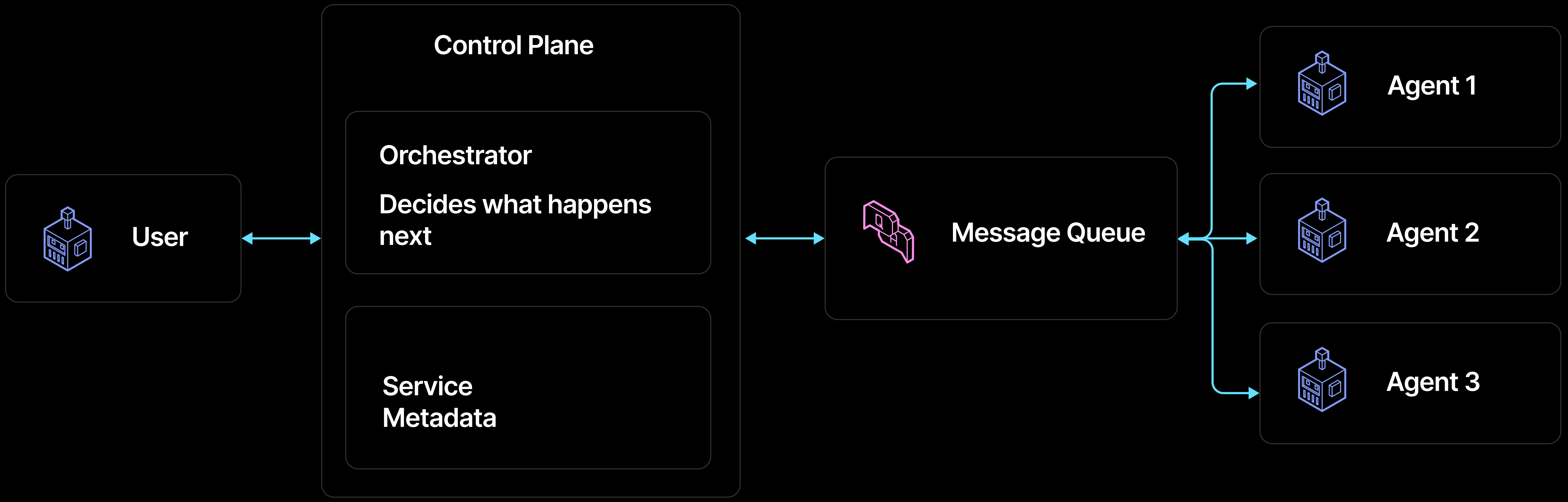
1. Encapsulation and re-use
2. Standardized communication interfaces between agents and with the client.
3. Scalability in number of users and number of agents
4. Human-in-the-loop for the end-user
5. Debugging and observability tools for the developer



llama-deploy

Deploy agentic workflows as **microservices**.

- Model every agent workflow as a service API
- All agent communication occurs via a central message queue
- Distributed tool-execution
- Human-in-the-loop as a service
- Easy deployment with docker-compose and Kubernetes



- https://github.com/run-llama/llama_deploy

Thank you!

LlamaIndex
September 23, 2024