From RAG to Knowledge Assistants

Jerry Liu
LlamaIndex

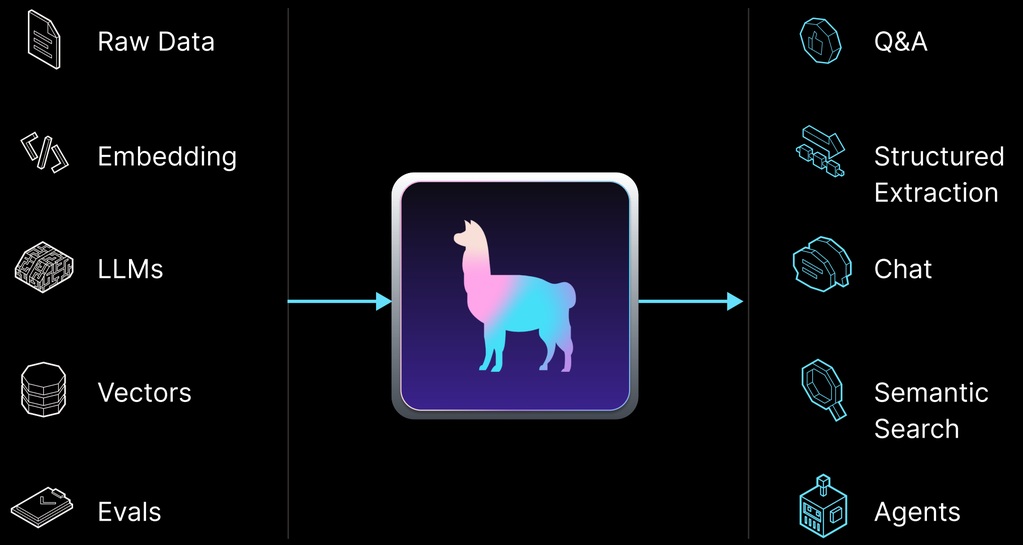
LlamaIndex

LlamaIndex is the fastest way to build production-quality **LLM applications over enterprise data**.

Products:

- **LlamaCloud:** Enterprise RAG platform
- **Open-Source:** Agent orchestration framework

Backed By: Greylock



Leading developer platform and community for GenAI:

- 200k+ followers on LinkedIn
- 70k+ followers on Twitter
- 34k+ stars
- 2M+ monthly downloads
- 20k+ Discord members
- 600+ integrations

Fortune 500 Companies to Startups

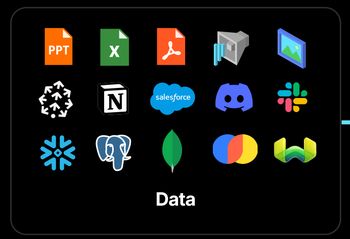
Red Hat
springworks
OPENBB
Jasper
replit
Softtek
Berkeley UNIVERSITY OF CALIFORNIA
Weights & Biases
DATASTAX
Clearbit
T Systems
GUNDERSON DETTMER
adyen
Uber

LLM Agents over any Data Type

LlamaCloud

Enterprise RAG platform to rapidly connect unstructured enterprise data to LLM agents.

LlamaCloud



Ingestion

Index

Value Proposition

✅ *Immediate* production deployment of multimodal RAG-based knowledge assistants

✅ Best-in-class performance over complex documents, w/ tables, charts, diagrams, images, and more.

✅ Necessary building block in a broader agentic workflow

Open-Source

Leading developer framework for orchestrating **single and multi-agent workflows** over your data.

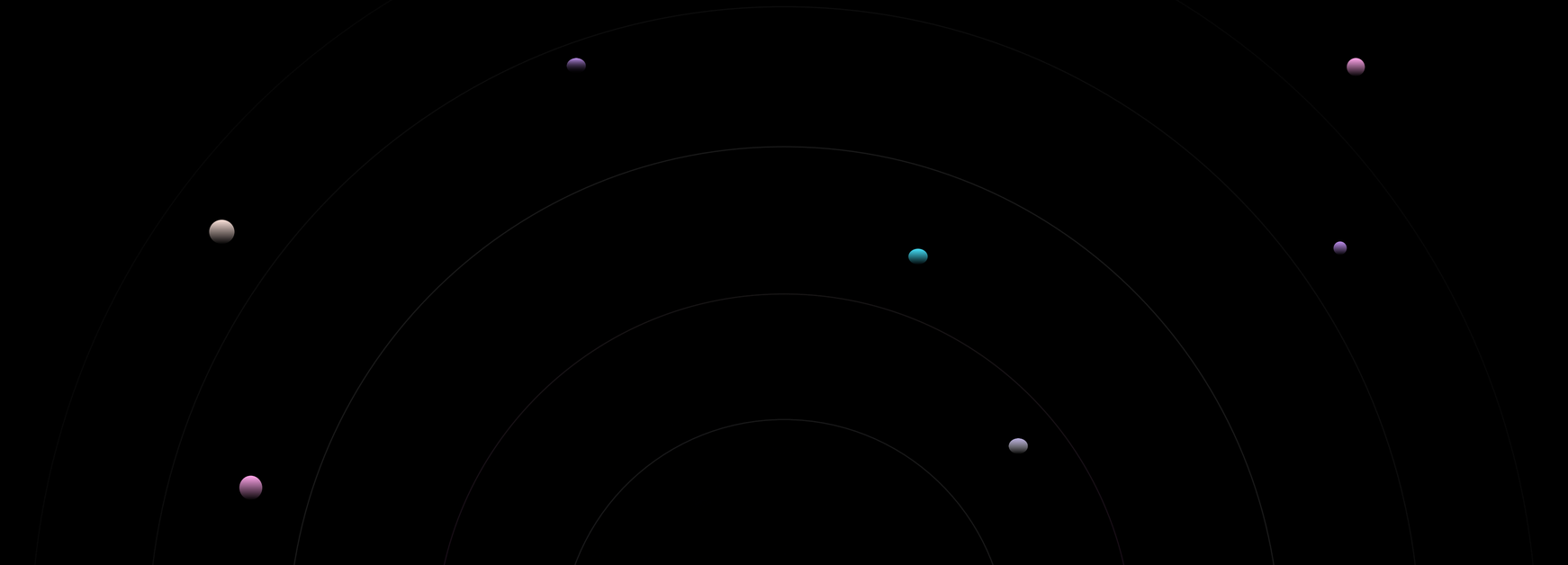
LlamaIndex

Orchestration

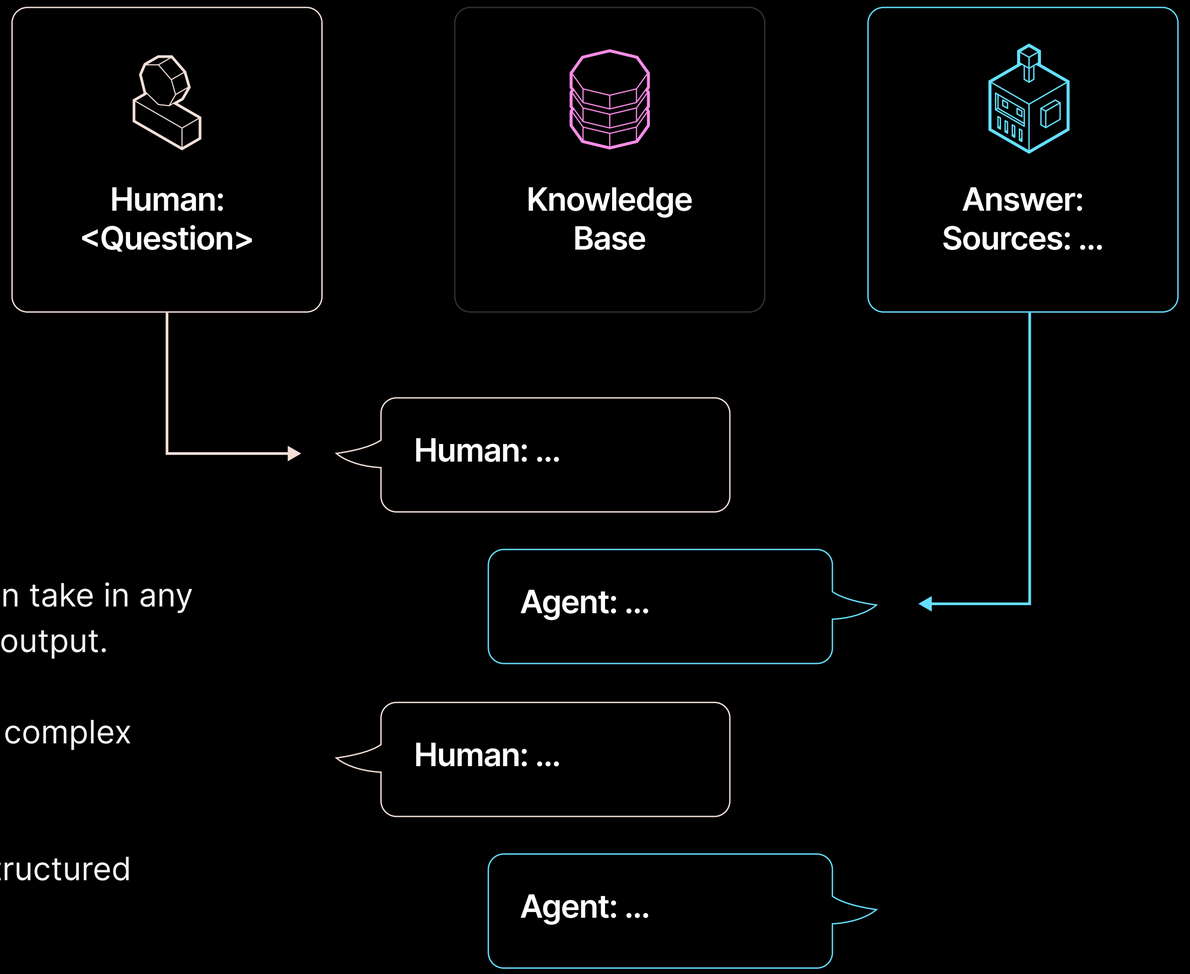
Production LLM Agents

- Knowledge Assistants
- Report Generation
- Business Process Automation
- and more!

Prototyping is Easy.
Productionization is Hard.



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

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

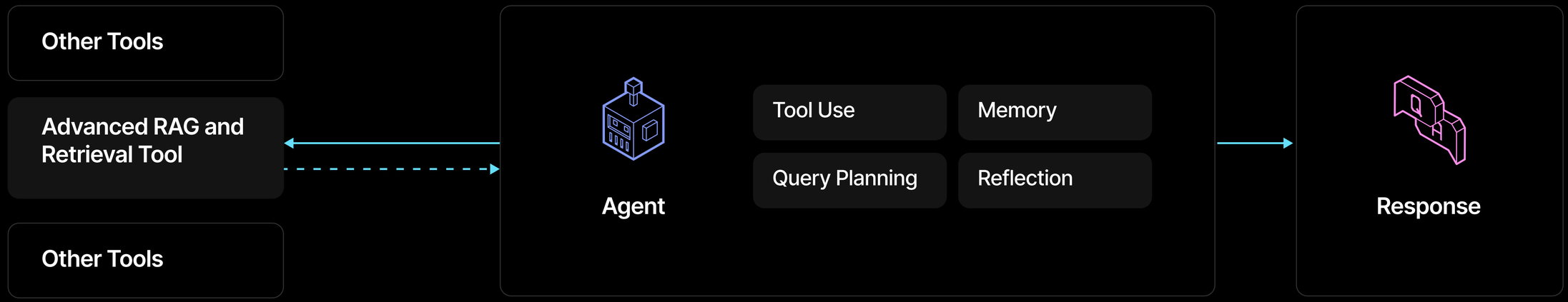
How do we aim to build a production-ready knowledge assistant?

A Better Knowledge Assistant

1. **High-quality data and retrieval**
2. Agentic reasoning over complex inputs
3. Agentic decision-making and output generation
4. Towards a scalable, full-stack application

A Better Knowledge Assistant

1. High-quality data and retrieval
2. **Agentic reasoning over complex inputs**
3. Agentic decision-making and output generation
4. Towards a scalable, full-stack application

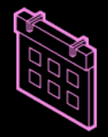


A Better Knowledge Assistant

1. High-quality data and retrieval
2. Agentic reasoning over complex inputs
3. **Agentic decision-making and output generation**
4. Towards a scalable, full-stack application

Report Generation

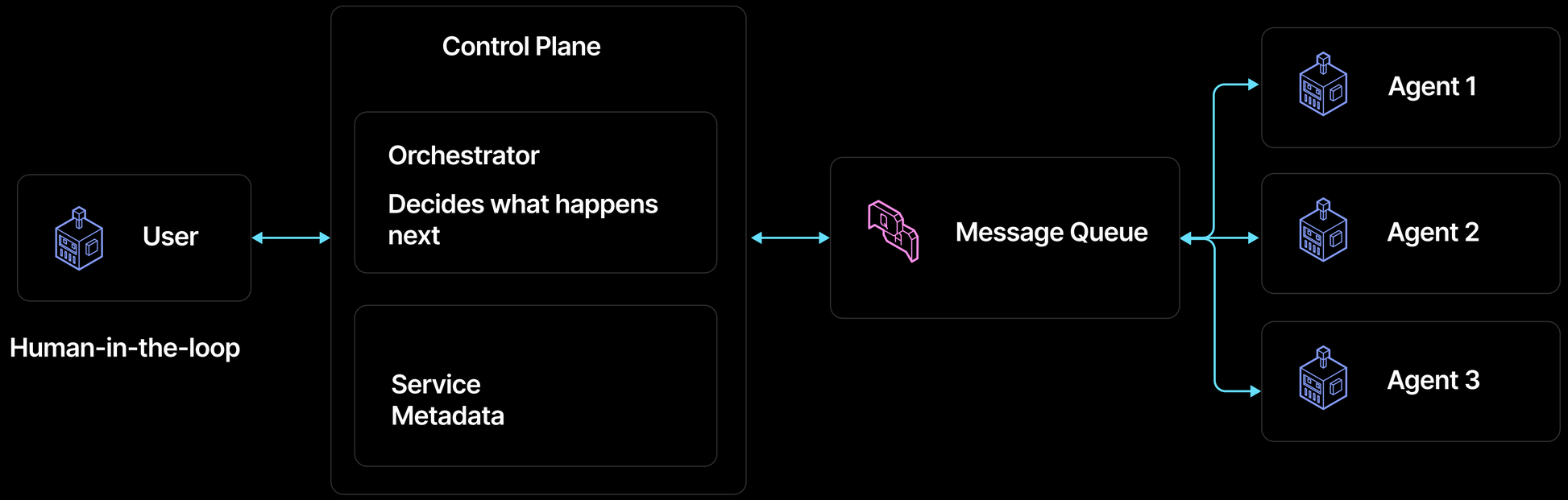
Data Analysis



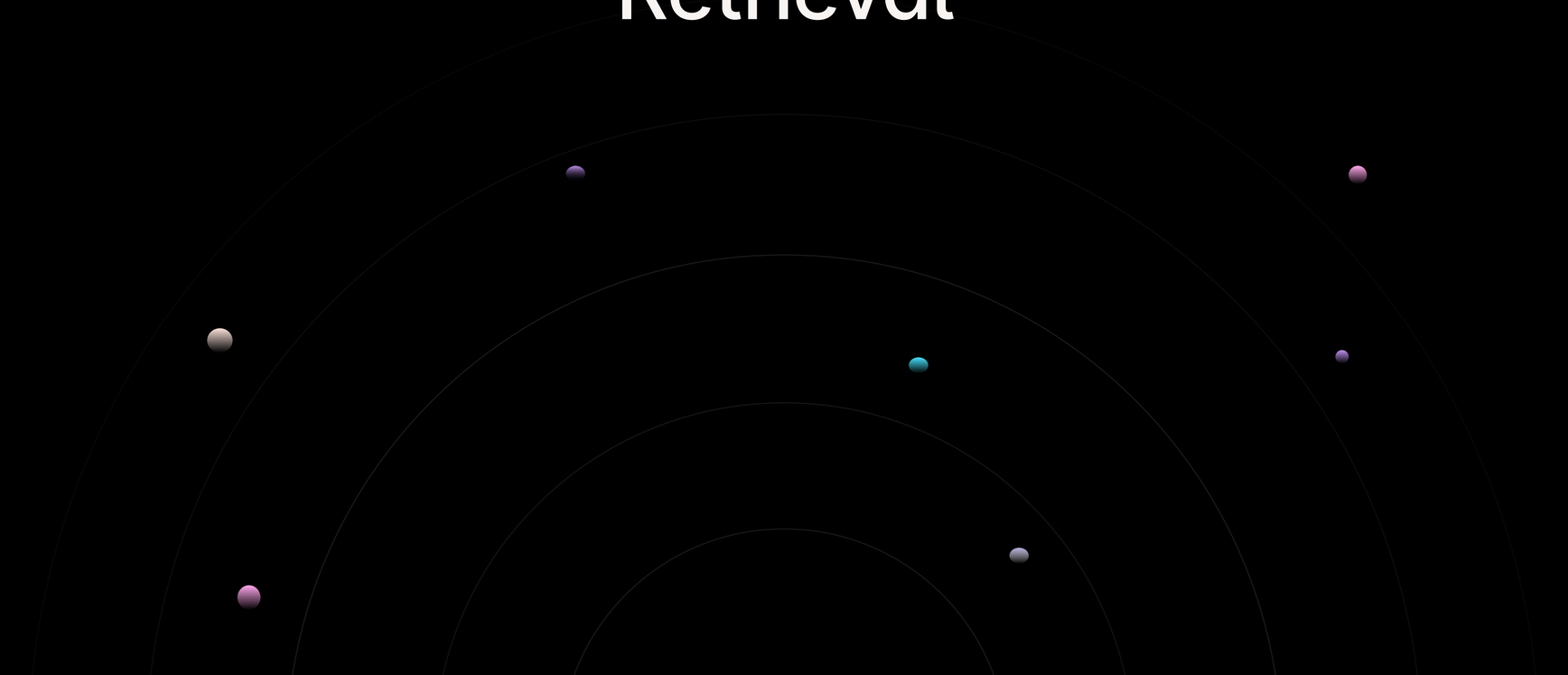
Action-Taking

A Better Knowledge Assistant

1. High-quality data and retrieval
2. Agentic reasoning over complex inputs
3. Agentic decision-making and output generation
4. **Towards a scalable, full-stack application**



High-Quality Data and Retrieval



Case Study: Complex Documents

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

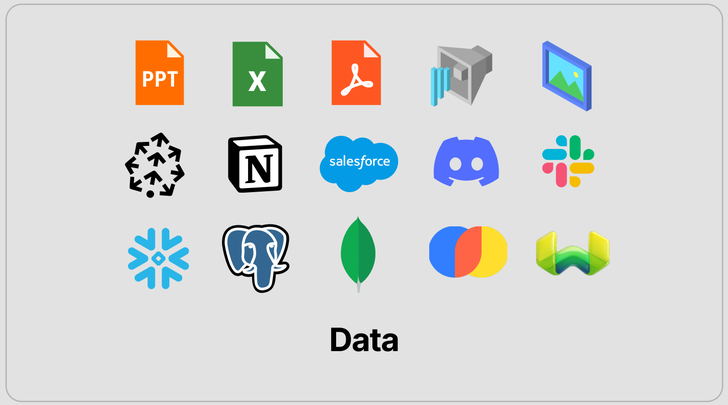
- Embedded Tables, Charts, Images
- Irregular Layouts
- Headers/Footers

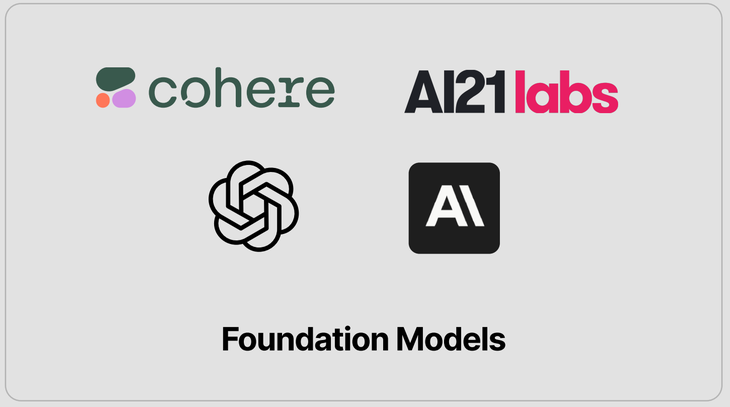
Users want to ask research questions over this data:

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

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

Developers

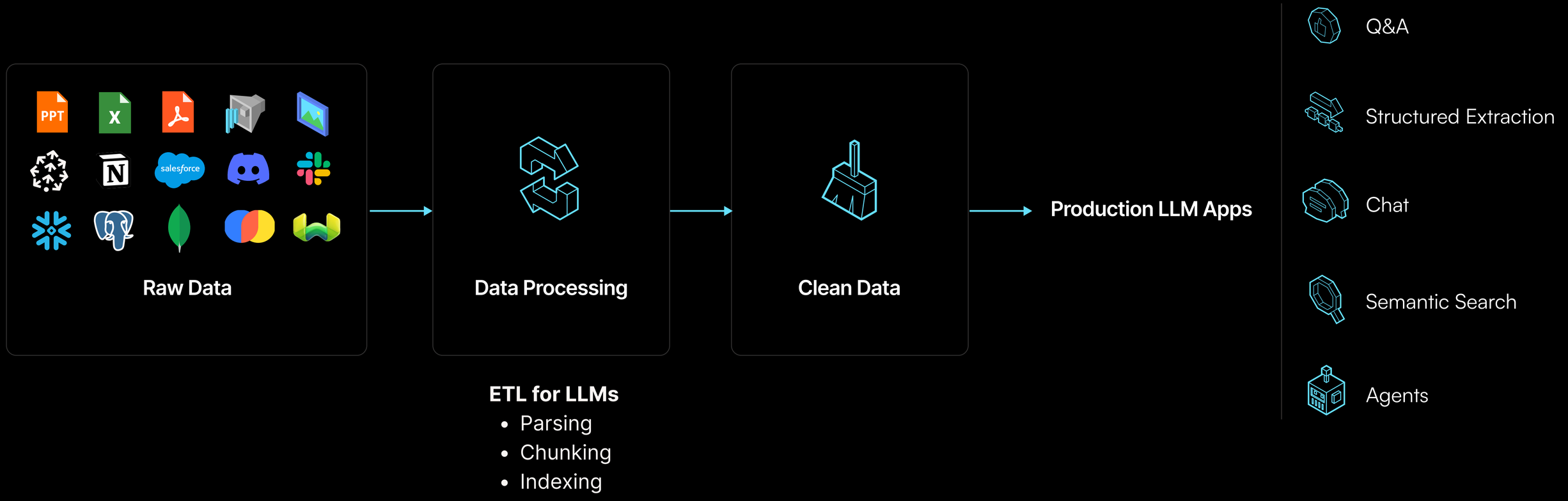




Any LLM App is only as Good as your Data

Garbage in = garbage out

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



ETL for LLMs

- Parsing
- Chunking
- Indexing

LlamaParse

“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

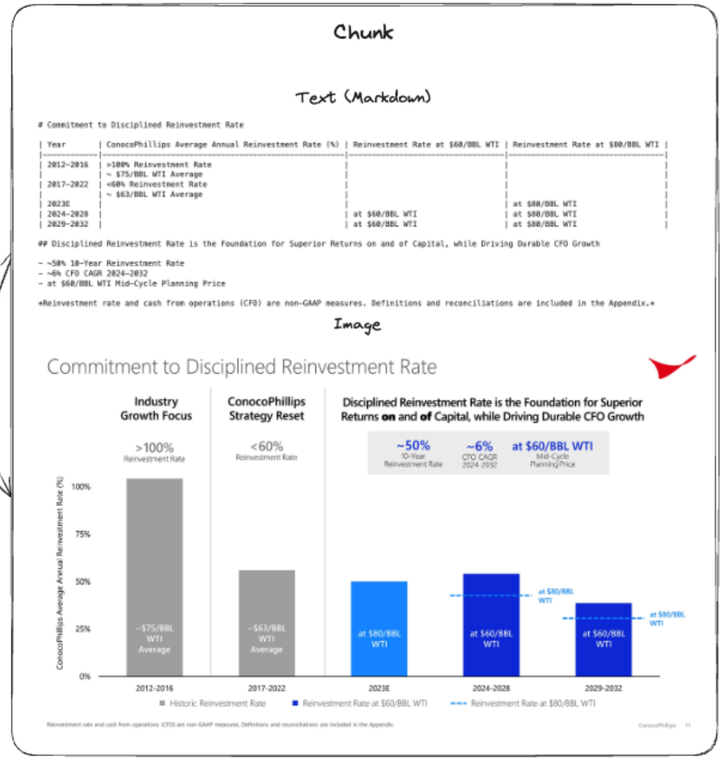
Earnings Decks (PPTX)

Chunk

Text (Markdown)

```
# Commitment to Disciplined Reinvestment Rate

| Year      | ConocoPhillips Average Annual Reinvestment Rate (%) | Reinvestment Rate at $60/BBL WTI | Reinvestment Rate at $80/BBL WTI |
|-----------|-----------------------------------------------------|----------------------------------|----------------------------------|
| 2012–2016 | >100% Reinvestment Rate                             |                                  |                                  |
|           | ~ $75/BBL WTI Average                               |                                  |                                  |
| 2017–2022 | <60% Reinvestment Rate                              |                                  |                                  |
|           | ~ $63/BBL WTI Average                               |                                  |                                  |
| 2023E     |                                                     |                                  | at $80/BBL WTI                   |
| 2024–2028 |                                                     | at $60/BBL WTI                   | at $80/BBL WTI                   |
| 2029–2032 |                                                     | at $60/BBL WTI                   | at $80/BBL WTI                   |

## Disciplined Reinvestment Rate is the Foundation for Superior Returns on and of Capital, while Driving Durable CFO Growth

- ~58% 10-Year Reinvestment Rate
- ~6% CFO CAGR 2024–2032
- at $60/BBL WTI Mid-Cycle Planning Price

*Reinvestment rate and cash from operations (CFO) are non-GAAP measures. Definitions and reconciliations are included in the Appendix.*
```

Image



Indexing (Embed text)

Retrieval

Vector Database

10K Reports (Text + Tables)

```
Cash flows from financing activities for Net
uery)
********")

 Engine**********")

)
etriever Query Engine**********")

financing activities for Netflix is not provi
**********
 flows from financing activities were $700,00
mber 31, 2021.

 Query Engine**********
financing activities for the year ended Decem
```

(407,729)	(524,585)
—	(26,919)
(757,387)	(788,349)
(911,276)	—
(2,076,392)	(1,339,853)
—	—
—	—
(700,000)	(500,000)
35,746	174,414
—	(600,022)
—	(224,168)

Net cash provided by (used in) financing activities

Financials (Excel)

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	

Accident Claims (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
approaching the intersection with Sunset Blvd. As I entered the intersection, a blue 2018 Ford Escape (License Plate: 8ABC456) traveling eastbound on Sunset Blvd ran a red light and collided with the front passenger side of my

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

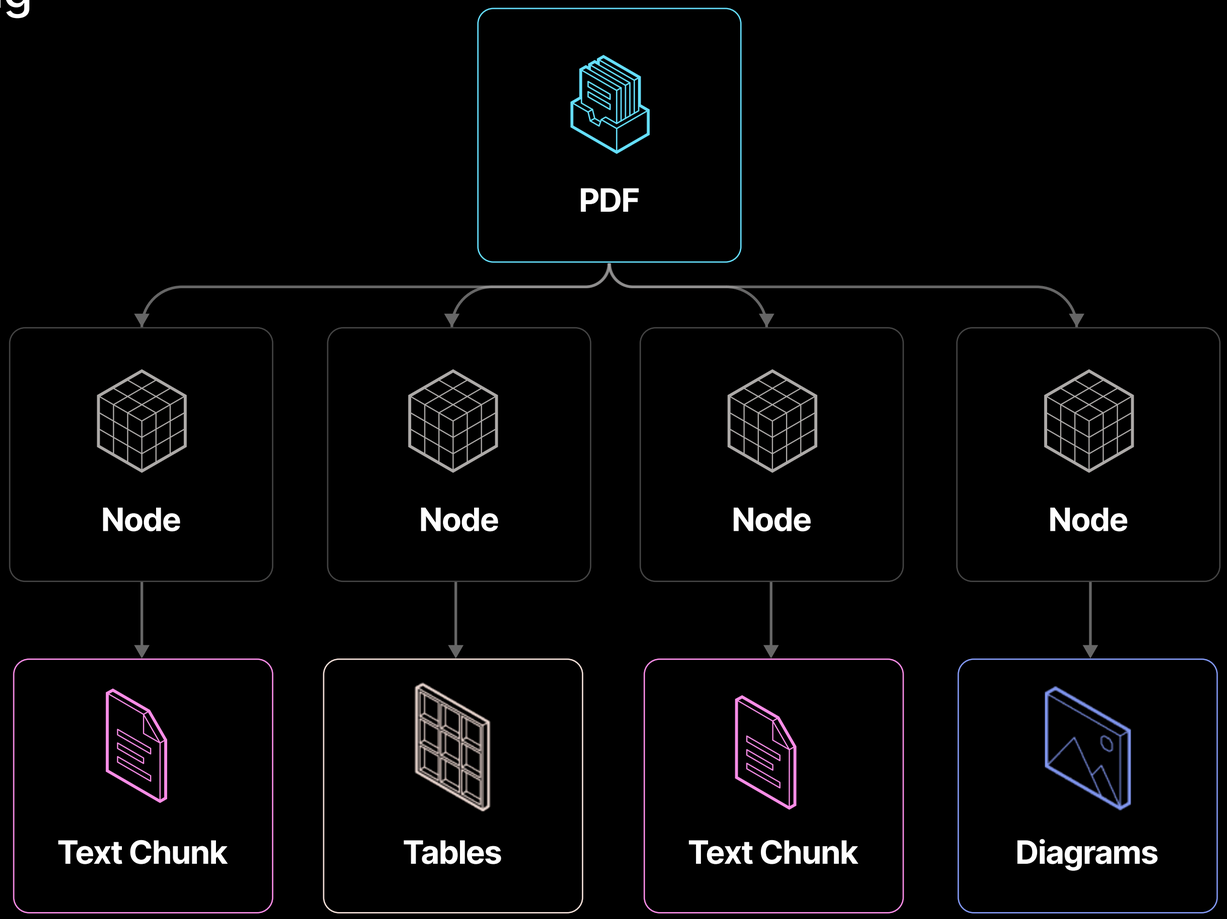
IMPACT

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**



LlamaCloud: An Enterprise RAG Platform

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

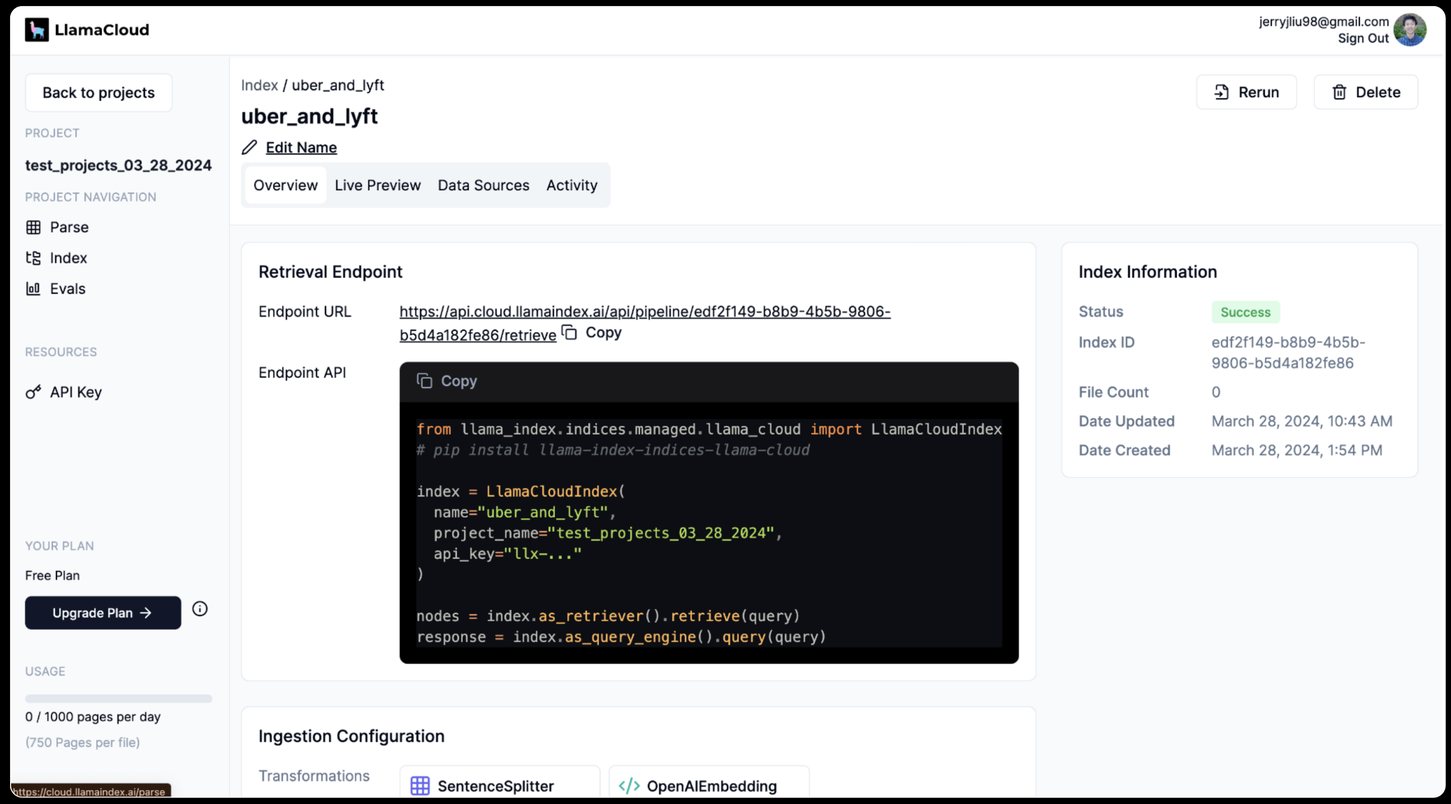
✅ **Instant Time-to-Value** for building knowledge assistants
- Out-of-the-box advanced RAG capabilities
- Free up developer time to rapidly iterate on higher-level agent use cases

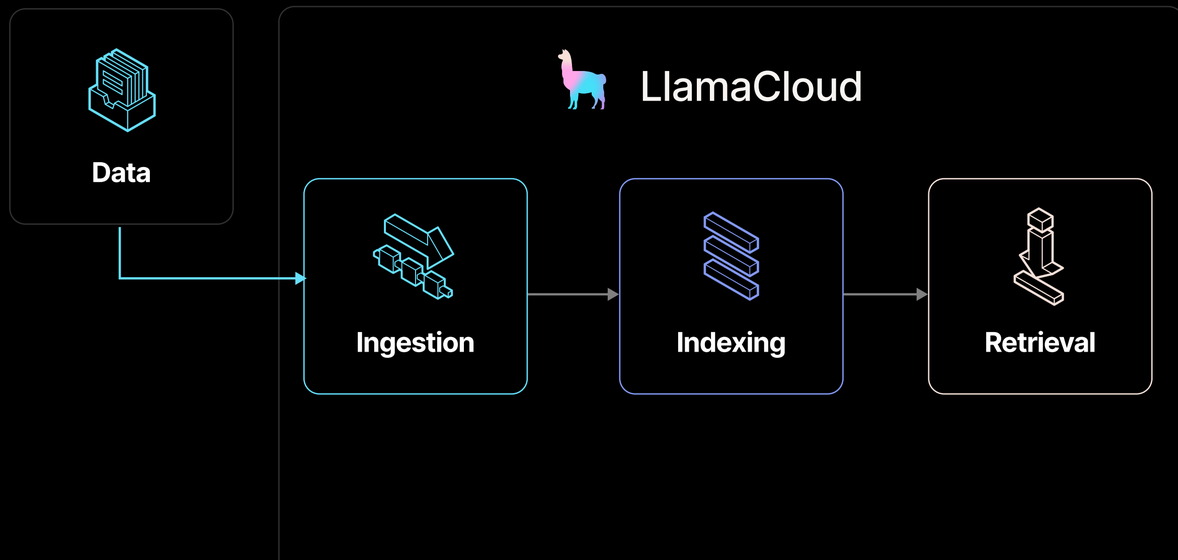
✅ **State-of-the-Performance** leads to increased satisfaction and reduced compliance risk

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

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

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

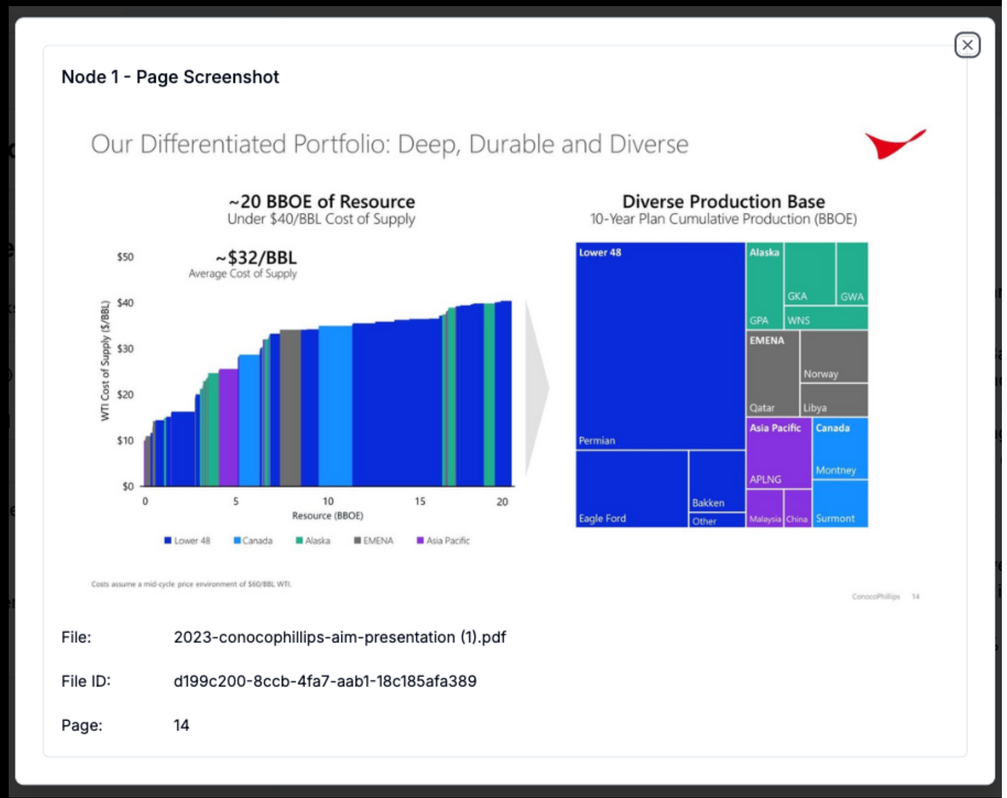


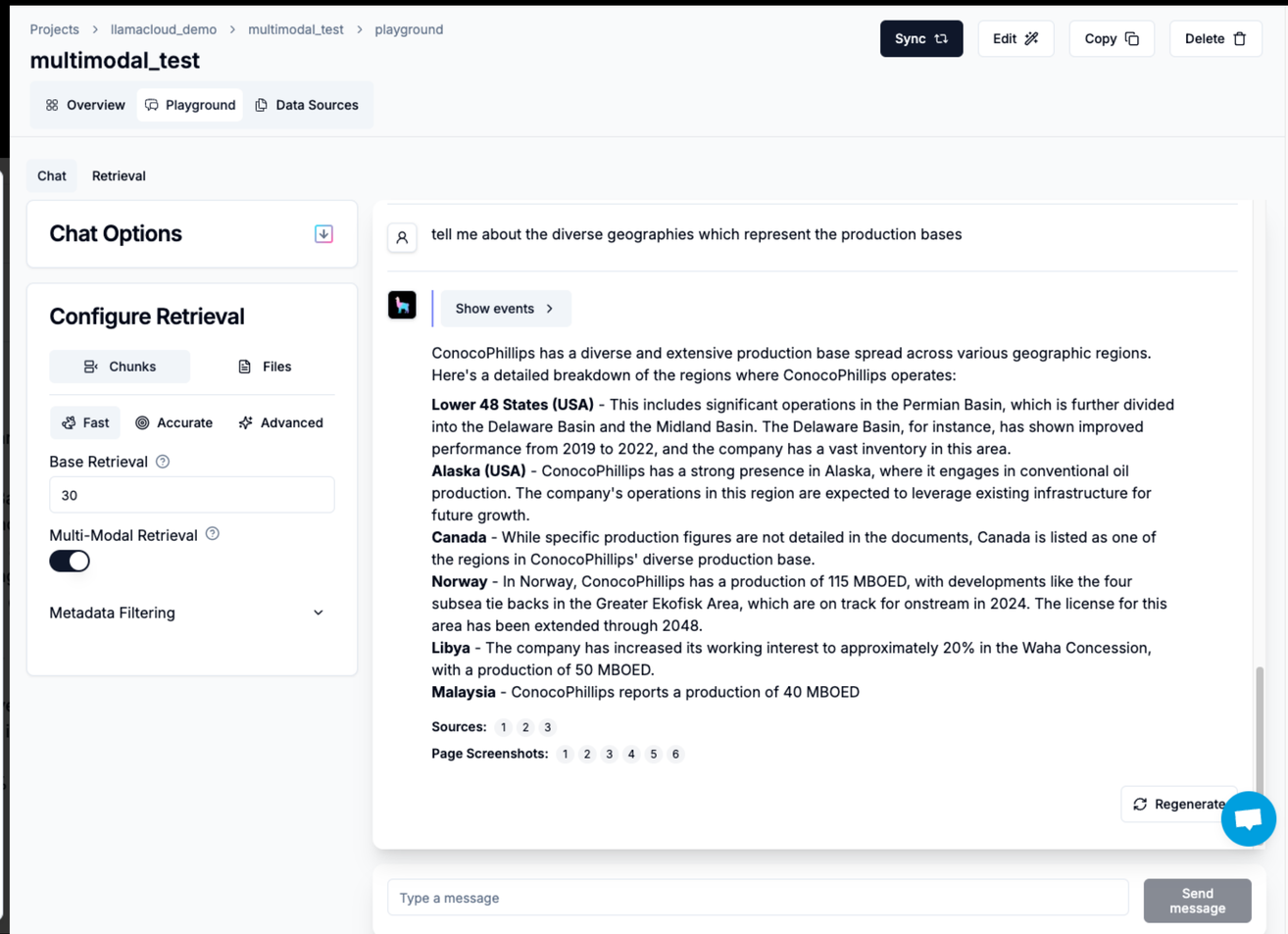


E2E Multimodal RAG Capabilities

Setup multimodal indexing and retrieval in minutes

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





LlamaIndex

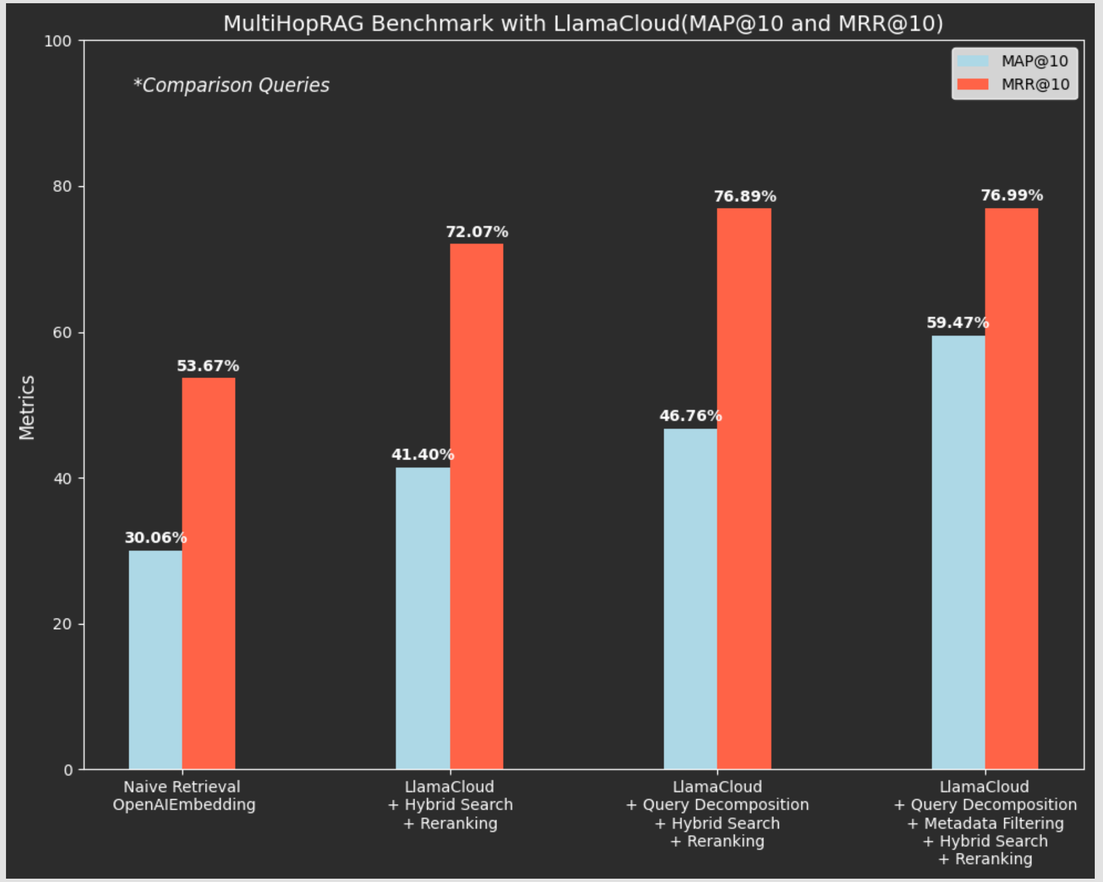
LlamaCloud Dramatically Accelerates Time-to-Value with High Performance

LlamaCloud helps users dramatically accelerate their RAG **time to production for multiple use cases.**

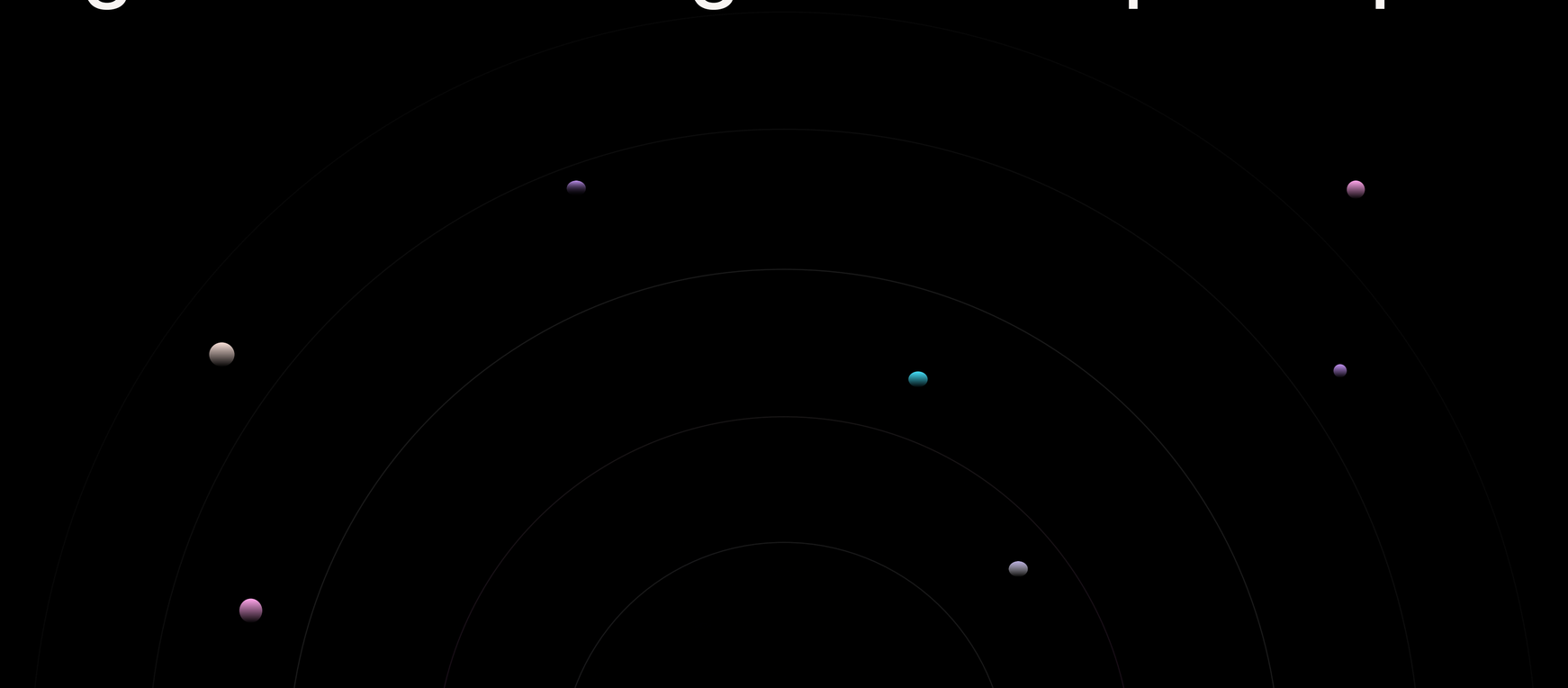
Vs. Pure DIY (open-source, roll-your-own)
Vs. Pure Buy

Metric	In-House Solution	LlamaCloud
Dev Time for Data Sources	2 weeks	~5-30 mins
Dev Time for Parsing	2-3 weeks	~5-30 mins
Dev Time for Chunking	2-3 weeks	~5-30 mins

LlamaCloud enables users to get **state-of-the-art quality over their data.**



Agentic Reasoning over Complex Inputs



Complex Inputs

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

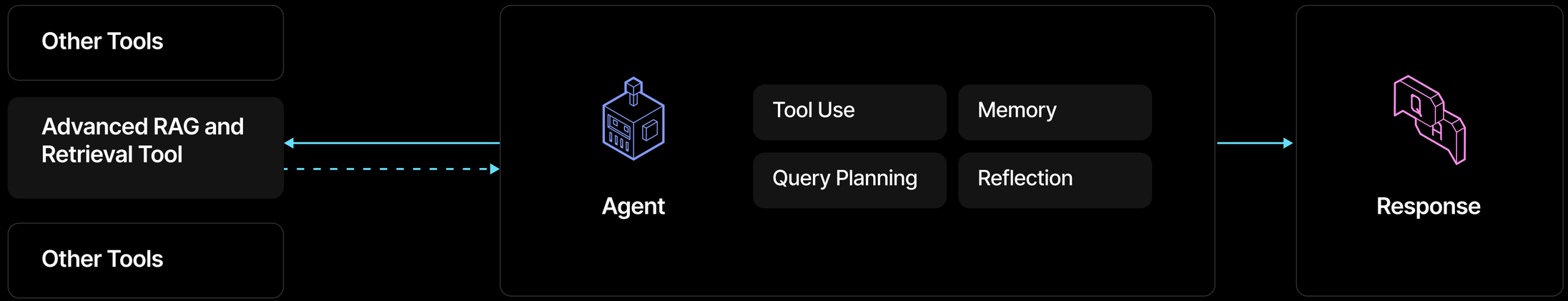
Summarization Questions: “Give me a summary of the entire <company> 10K annual report”
Comparison Questions: “Compare the open-source contributions of candidate A and candidate B”
Multi-part Questions: “Tell me about the pro-X arguments in article A, and tell me about the pro-Y arguments in article B, 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?”

Agentic RAG

Every data interface is a tool

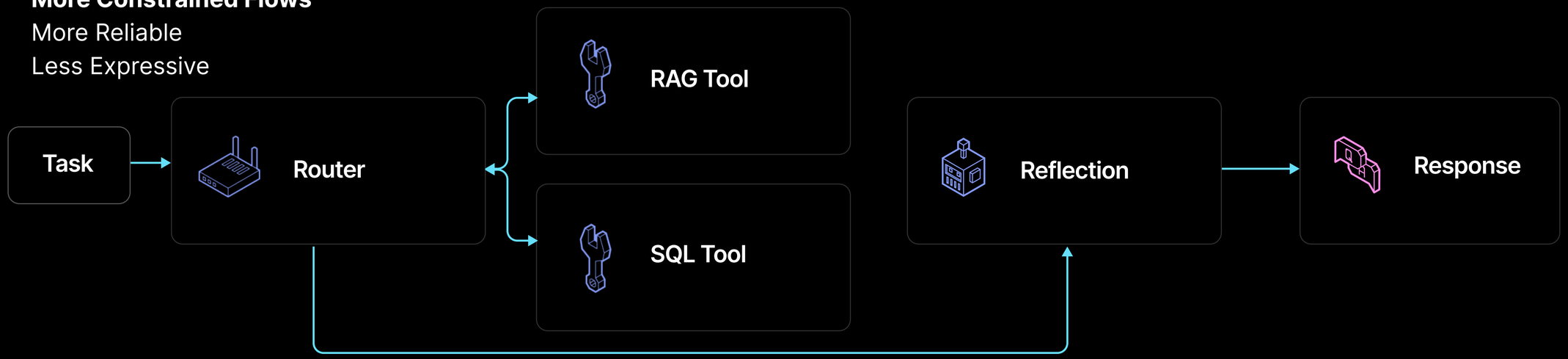
Use agent reasoning loops (sequential, DAG, tree) to tackle complex tasks

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



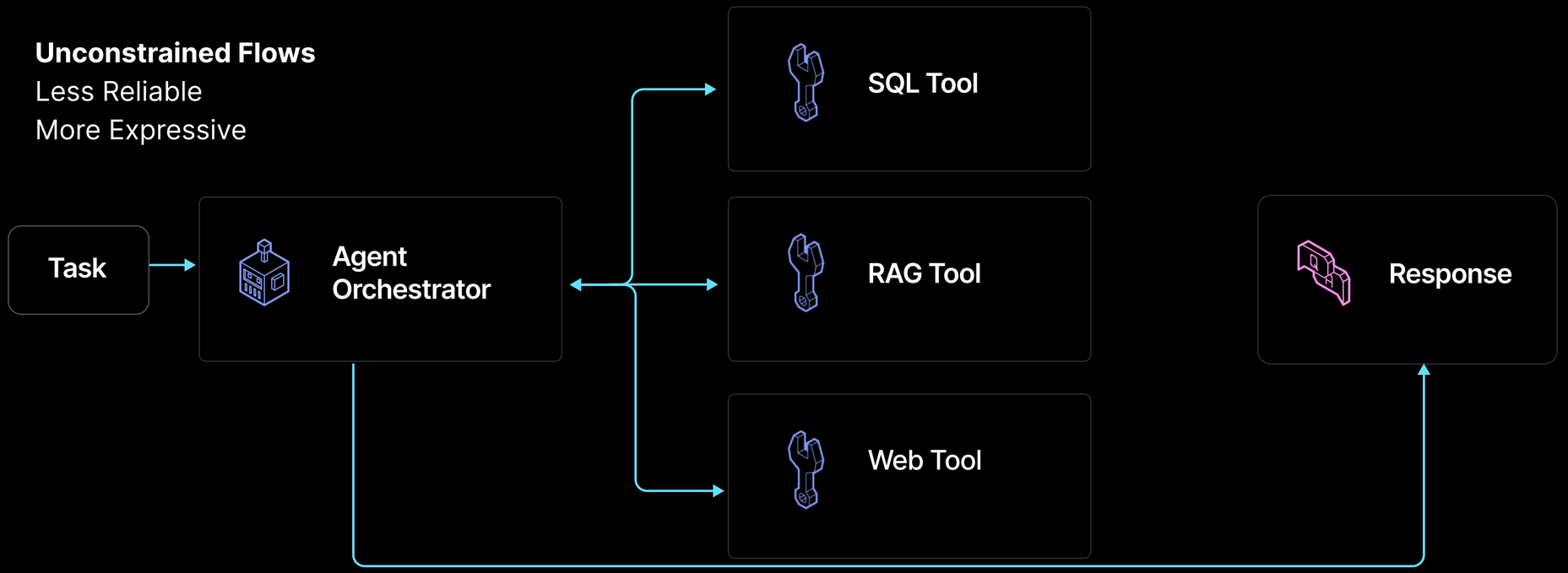
Unconstrained vs. Constrained Flows

More Constrained Flows
More Reliable
Less Expressive



Unconstrained vs. Constrained Flows

Unconstrained Flows
Less Reliable
More Expressive

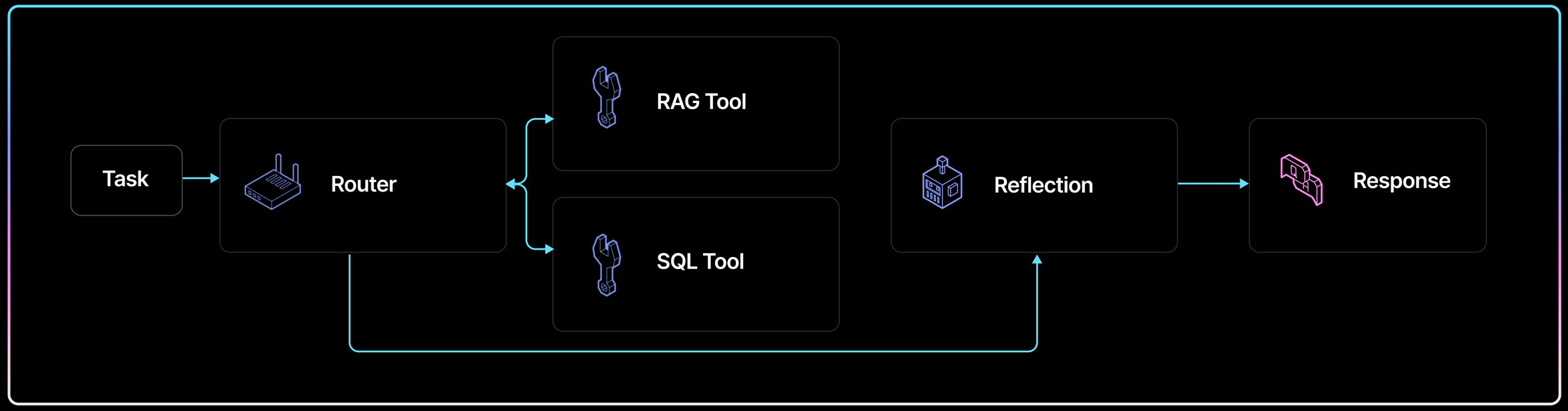


Agentic Orchestration Foundations

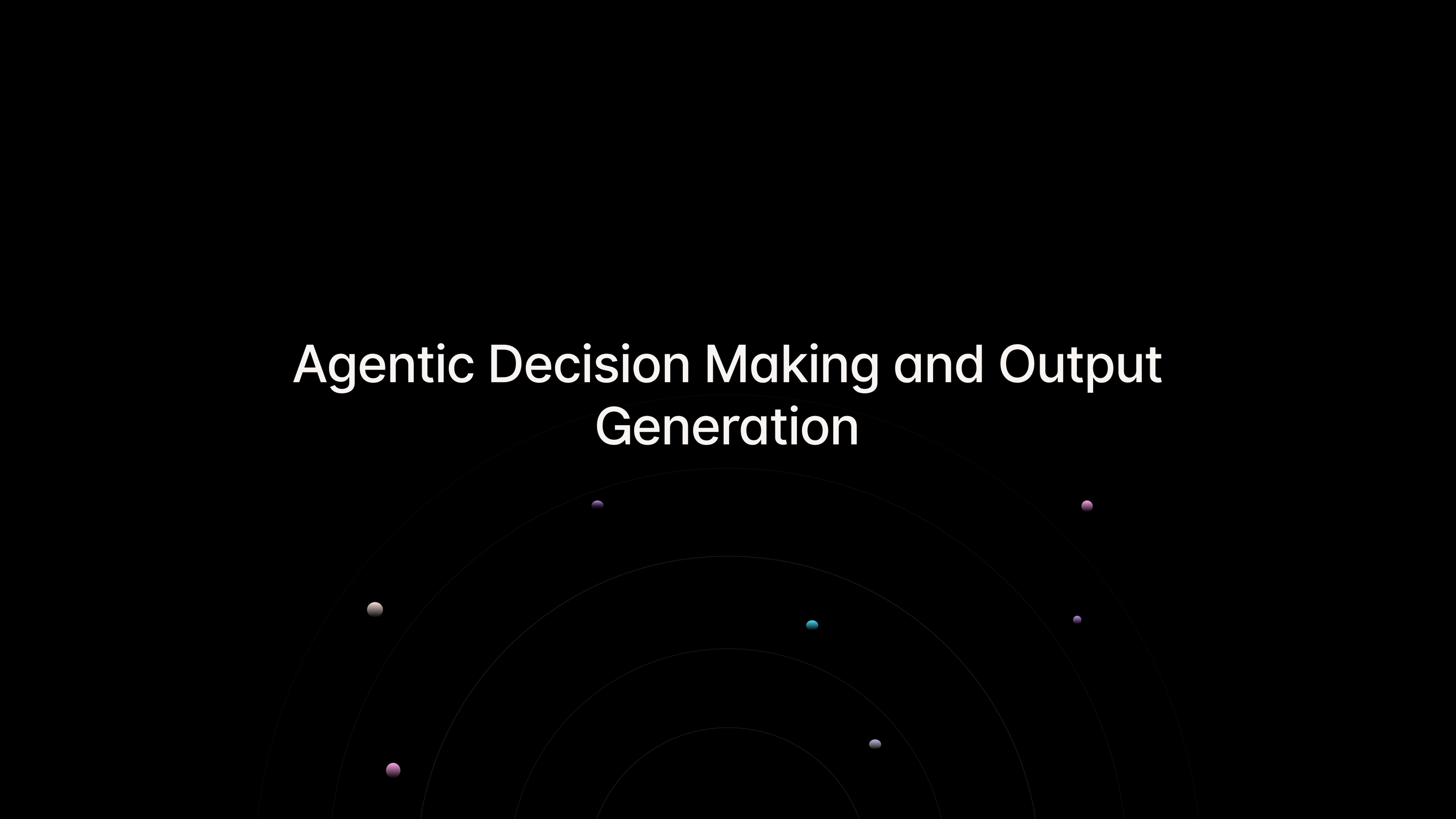
We believe an agent orchestration framework should have the following properties

✅ **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



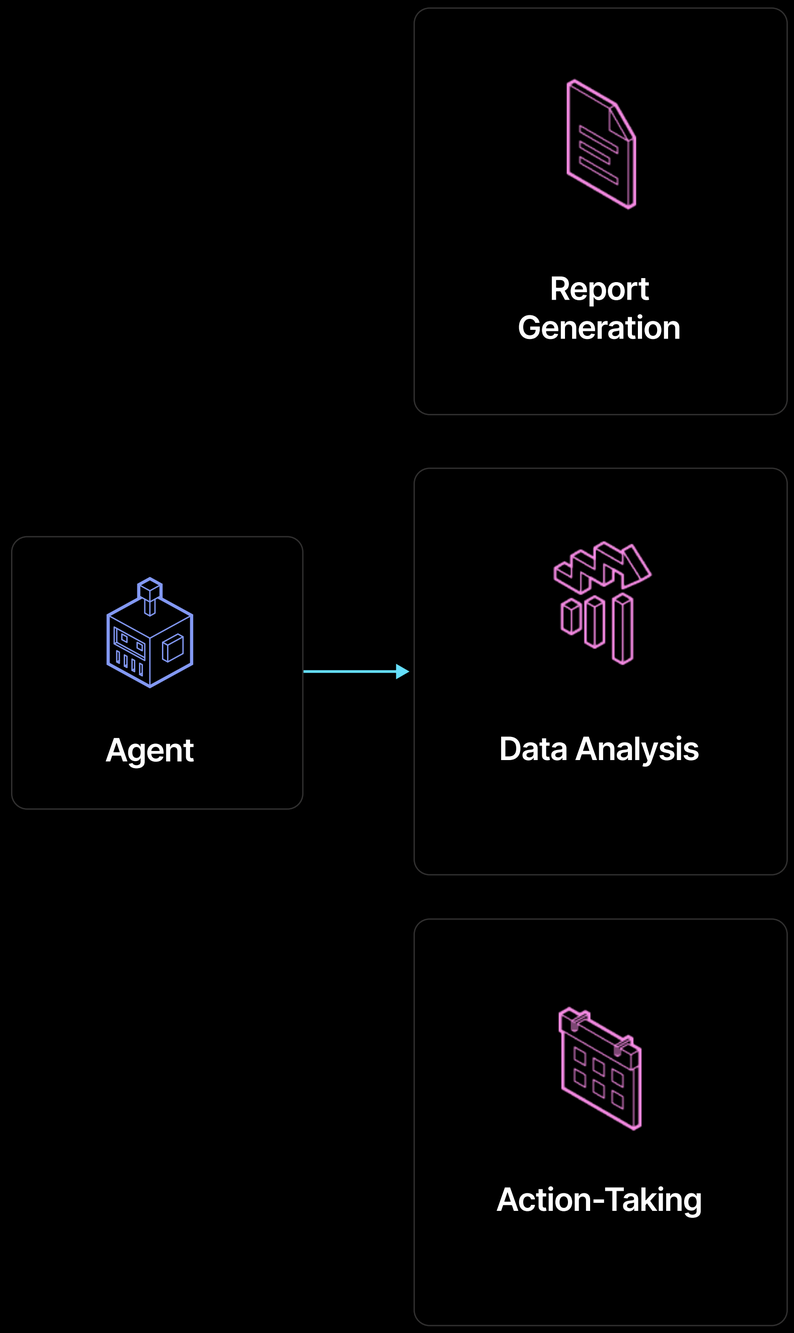
Agentic Decision Making and Output Generation



Automating Decision Making

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

1. Produce knowledge work
2. Take actions



Report Generation

Report generation is one of the top enterprise agent use cases, and the natural next step for agentic RAG.

Use Cases:

1. Generate a full research report or presentation (with text, tables, images).
2. Fill out an example form or questionnaire.
3. Fill out an Excel sheet

Top Use Case among our Customers

Fortune 500 Investment Firm

Global investment bank

Big 4 Consulting Firm

Leading Chinese Social Media Company

Global 2000 Construction Company

Example: Multimodal Report Generation

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

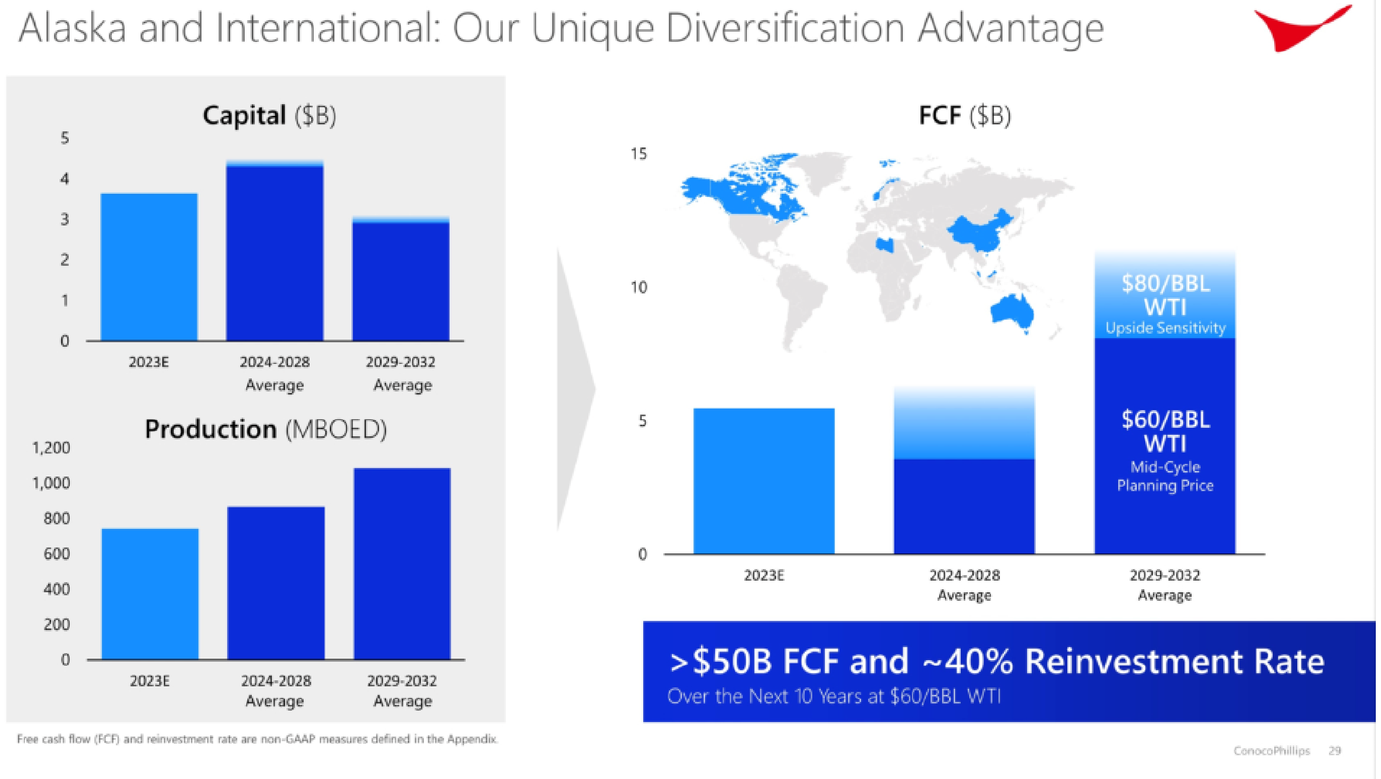
Example architecture: research and writer steps

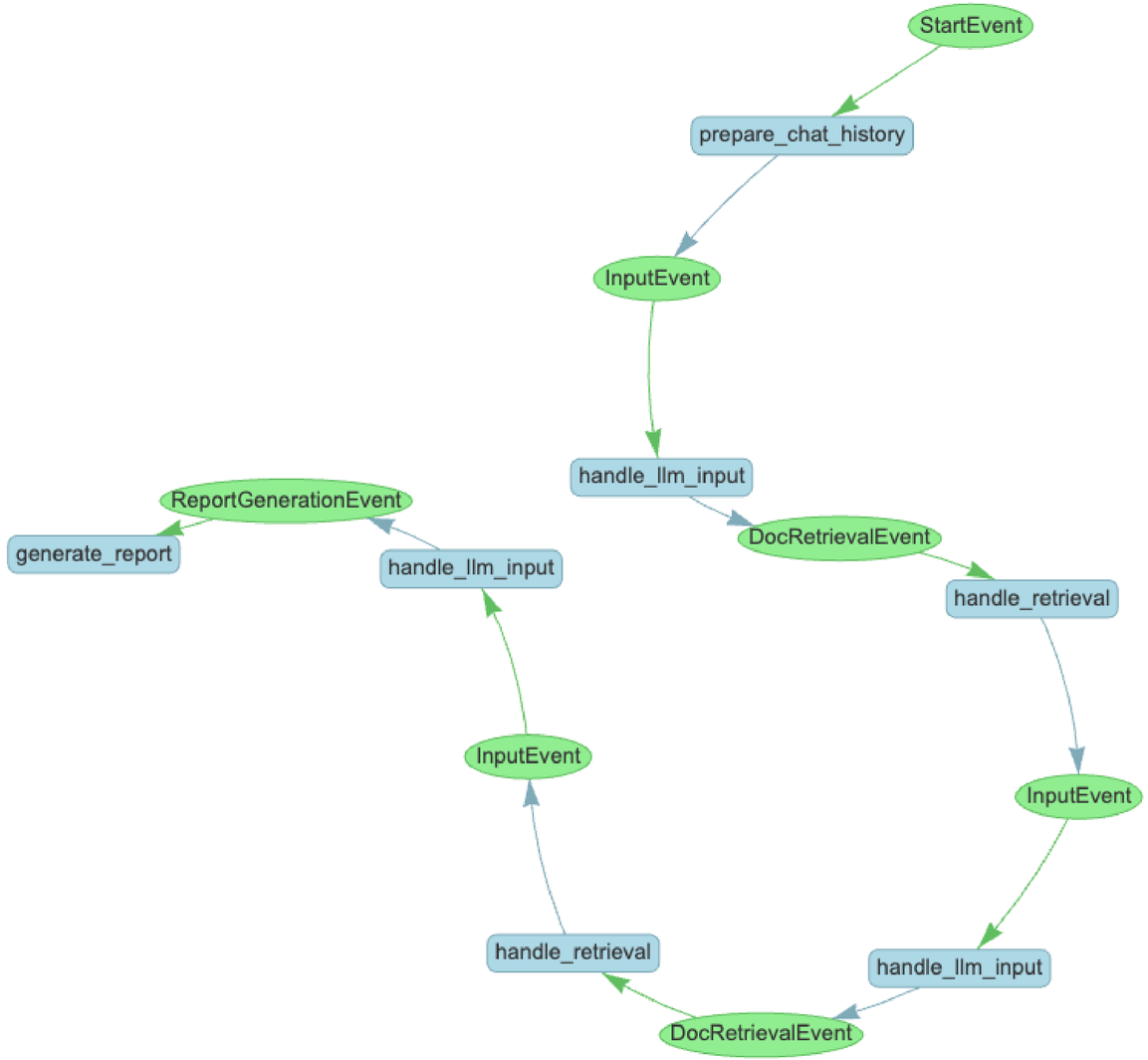
1. The **researcher** retrieves relevant chunks and documents, and puts them into a data cache.
2. The **writer** uses the data cache to generate a structured output of interleaving text and image blocks.

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

Lower 48 Segment

- **Capital Expenditures**: Expected to be $6.3 billion in 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 billion in 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.

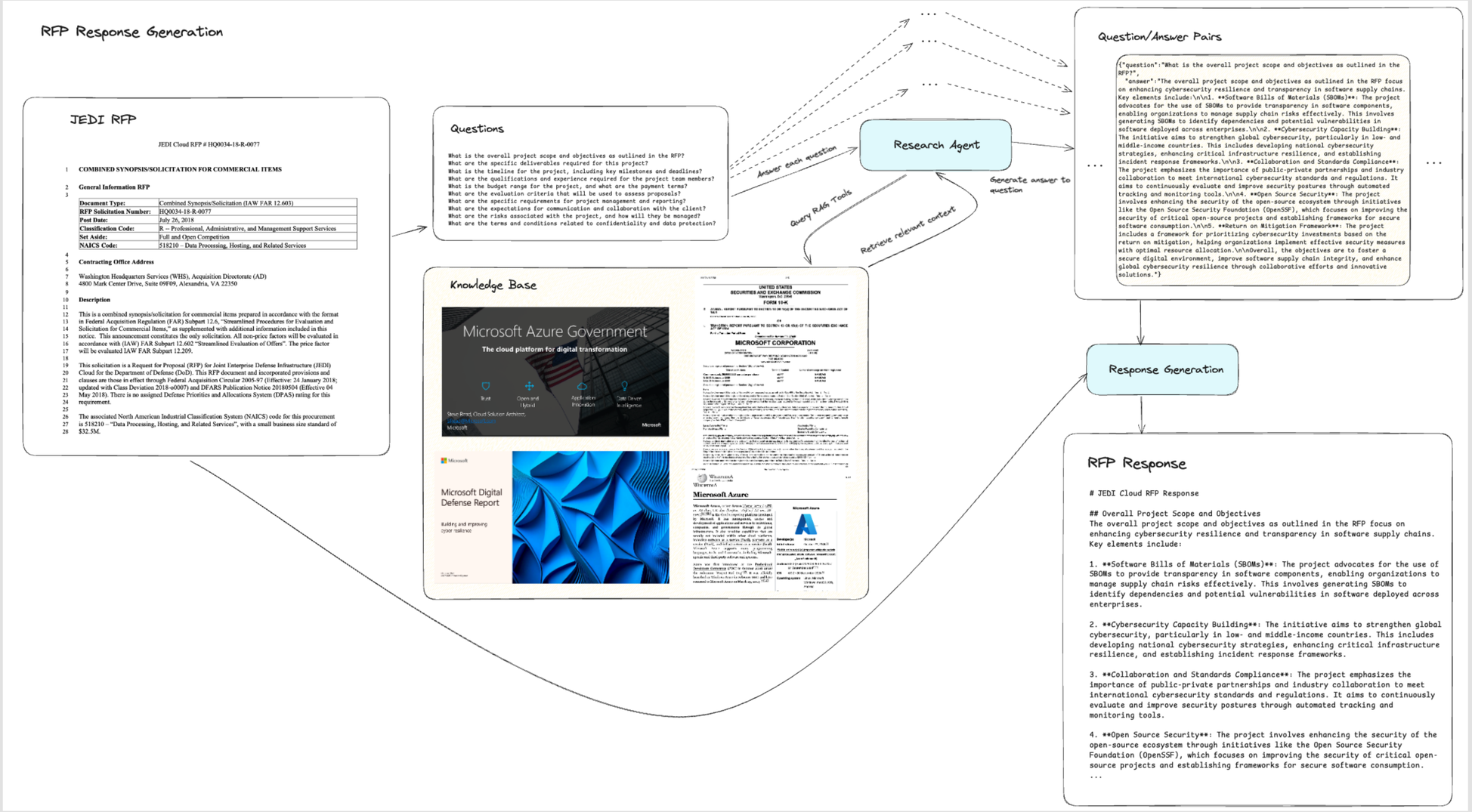




Example: RFP Response Generation

As a vendor, generate a response that adheres to guidelines outlined in a Request for Proposal (RFP).

https://github.com/run-llama/llama_parse/blob/main/examples/report_generation/rfp_response/generate_rfp.ipynb



Example: Excel Form Filling

Parse and fill in a structured Excel template for financial comparisons

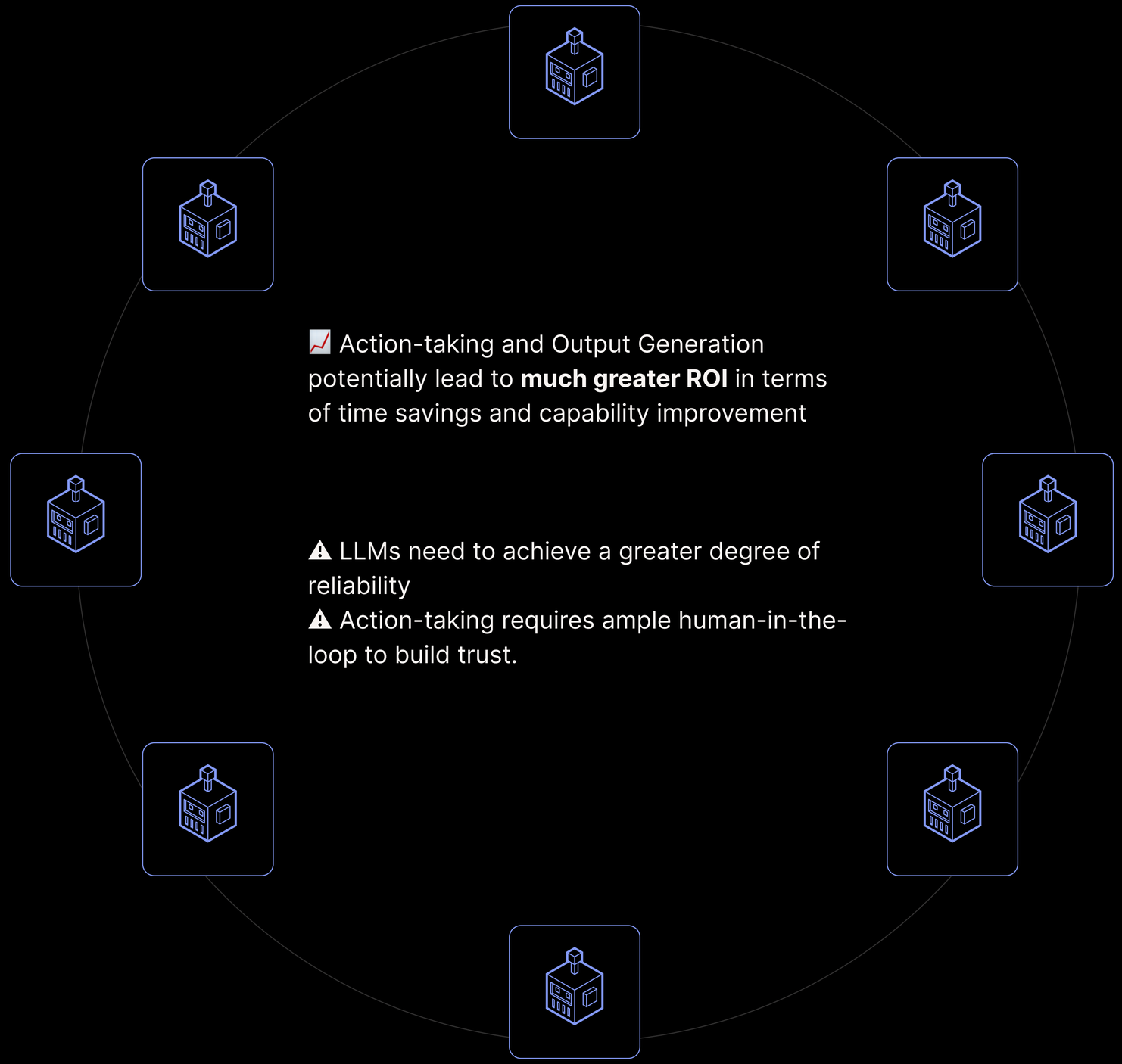
Example notebook: https://github.com/run-llama/llamacloud-demo/blob/main/examples/form_filling/Form_Filling_10K_SEC.ipynb

Parameter	2021			2022		
	Amazon (AMZN)	Microsoft (MSFT)	Apple (AAPL)	Amazon (AMZN)	Microsoft (MSFT)	Apple (AAPL)
1. Revenue						
2. Net Income						
3. Earnings Per Share (EPS)						
4. EBITDA						
5. Free Cash Flow						
6. Return on Equity (ROE)						
7. Return on Assets (ROA)						
8. Debt-to-Equity Ratio						
9. Current Ratio						
10. Gross Margin						
11. Operating Margin						
12. Net Profit Margin						
13. Inventory Turnover						
14. Accounts Receivable Turnover						
15. Capital Expenditures						
16. Research and Development Expenses						
17. Market Cap						
18. Price-to-Earnings (P/E) Ratio						
19. Dividend Yield						
20. Year-over-Year Growth Rate						

Parameter	2021			2022		
	Amazon (AMZN)	Microsoft (MSFT)	Apple (AAPL)	Amazon (AMZN)	Microsoft (MSFT)	Apple (AAPL)
1. Revenue						
2. Net Income						
3. Earnings Per Share (EPS)						
4. EBITDA						
5. Free Cash Flow						
6. Return on Equity (ROE)						
7. Return on Assets (ROA)						
8. Debt-to-Equity Ratio						
9. Current Ratio						
10. Gross Margin						
11. Operating Margin						
12. Net Profit Margin						
13. Inventory Turnover						
14. Accounts Receivable Turnover						
15. Capital Expenditures						
16. Research and Development Expenses						
17. Market Cap						
18. Price-to-Earnings (P/E) Ratio						
19. Dividend Yield						
20. Year-over-Year Growth Rate						

	Year	Company	Accounts Receivable Turnover	Capital Expenditures	Current Ratio	Debt-to-Equity Ratio	Dividend Yield	EBITDA	Earnings Per Share (EPS)	Free Cash Flow	Gross Margin	Inventory Turnover	Market Cap	Net Income	Net Profit Margin	Operating Margin	Price-to-Earnings (P/E) Ratio	Research and Development Expenses	Return on Assets (ROA)
0	2021	Amazon (AMZN)	5.6 times	58.3 billion	1.02	0.63	NaN	24,879	64.81	-$11,569	41.0%	4.8 times	$1.66 trillion	33,364	6.8%	13.7%	73.60	Not significant	5.9%
1	2021	Microsoft (MSFT)	7.5 times	$9.5 billion	1.79	0.30	1. 0.8%	$ 76,632 million	8.05	28.7 billion	$115.9 billion	2.5 times	$2.5 trillion	$61,271 million	19.7%	32%	34.50	20,716 million	10.9%
2	2021	Apple (AAPL)	6.2 times	$9,000 million	1.12	0.93	1. 0.0065	19,863	5.61	$ 73,000	$152,836 million	Not available	$2.46 trillion	94,680 million	21.7%	44.7%	28.11	$21,914 million	5.1%
3	2022	Amazon (AMZN)	6.4 times	$58.3 billion	1.10	1.39	NaN	$15,432 million	9.70	$-11,569 million	NaN	NaN	$1.47 trillion	(2,722)	NaN	16.9%	NaN	$73,213 million	6.9%
4	2022	Microsoft (MSFT)	6.1 times	$8.5 billion	1.78	0.59	2.0%	$107.895 billion	9.65	$58.7 billion	$135.620 billion	3.4 times	$1.87 trillion	72,738	19%	19%	38.60	$24,512 million	9.99%
5	2022	Apple (AAPL)	6.0 times	$42,117 million	2.78	0.68	0.0054	$145,787 million	6.15	$ 88,531	$170,782 million	1. 6.87	$2.9 trillion	99,803	21.9%	37.9%	24.60	$26,251 million	6.7%

Benefits and Risks

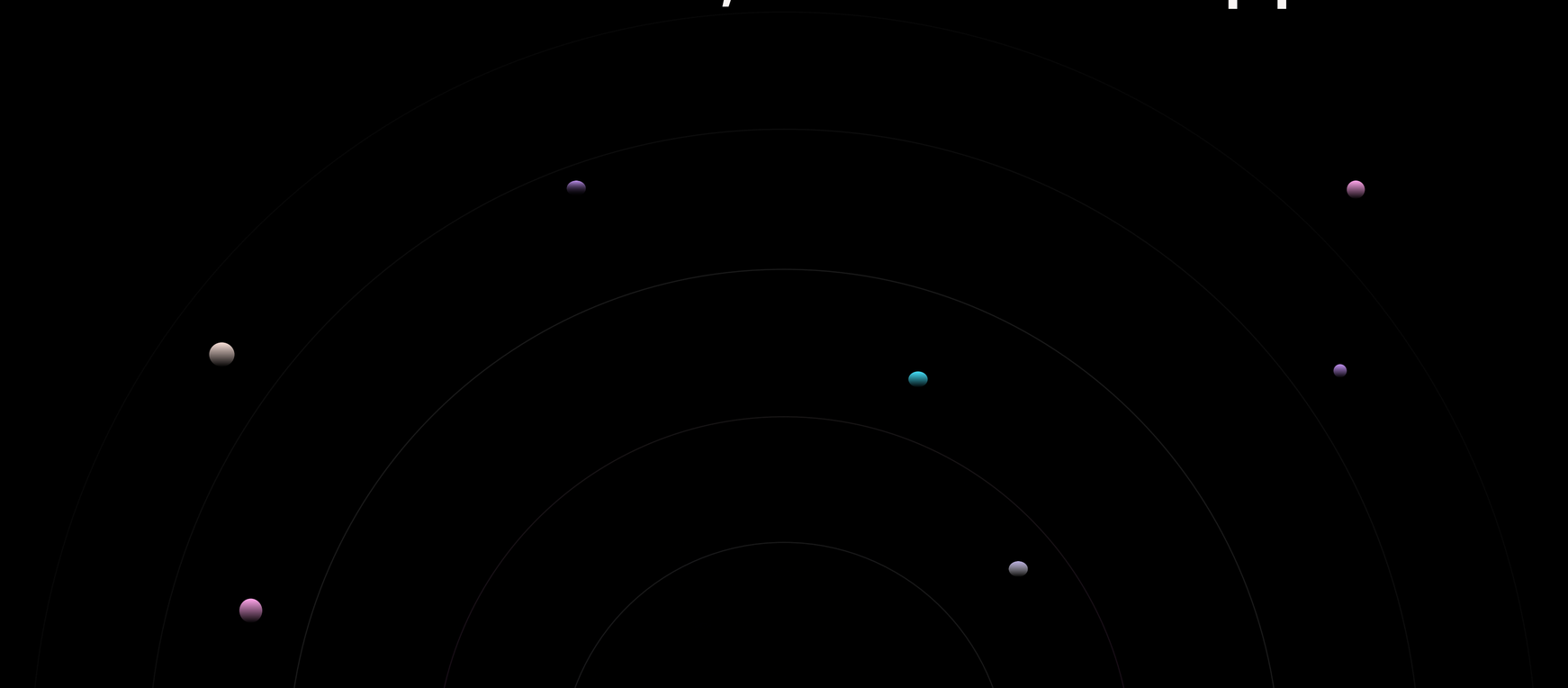


📈 Action-taking and Output Generation potentially lead to **much greater ROI** in terms of time savings and capability improvement

⚠️ LLMs need to achieve a greater degree of reliability

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

Towards a Scalable, Full-Stack Application

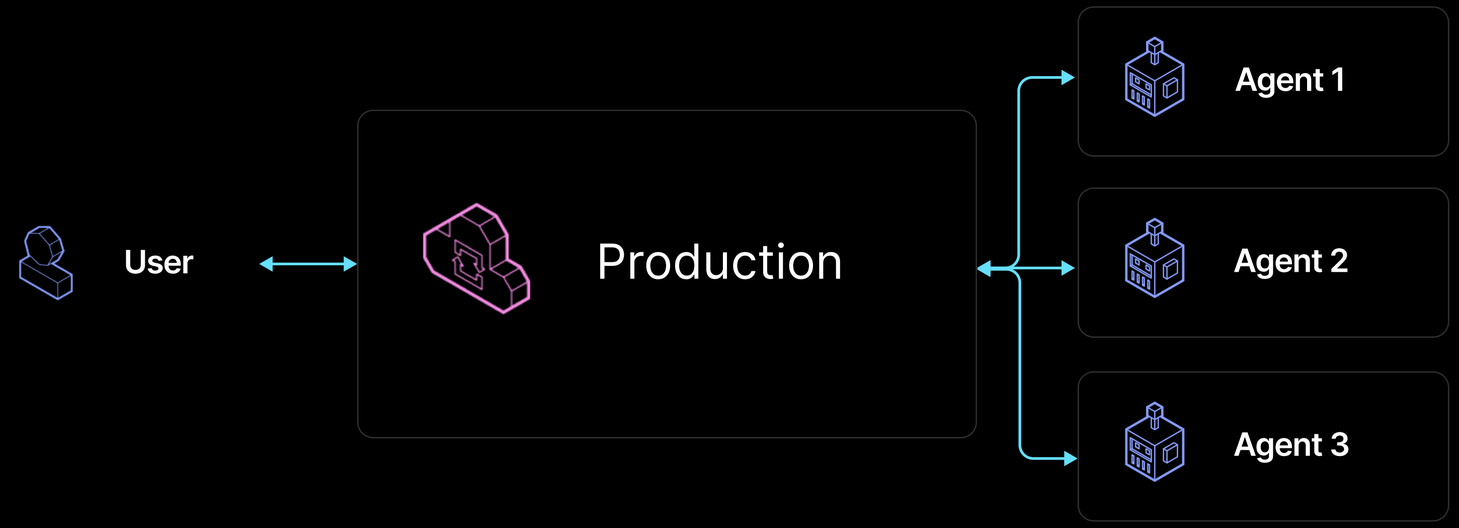


Running Agents in Production

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

Requirements:

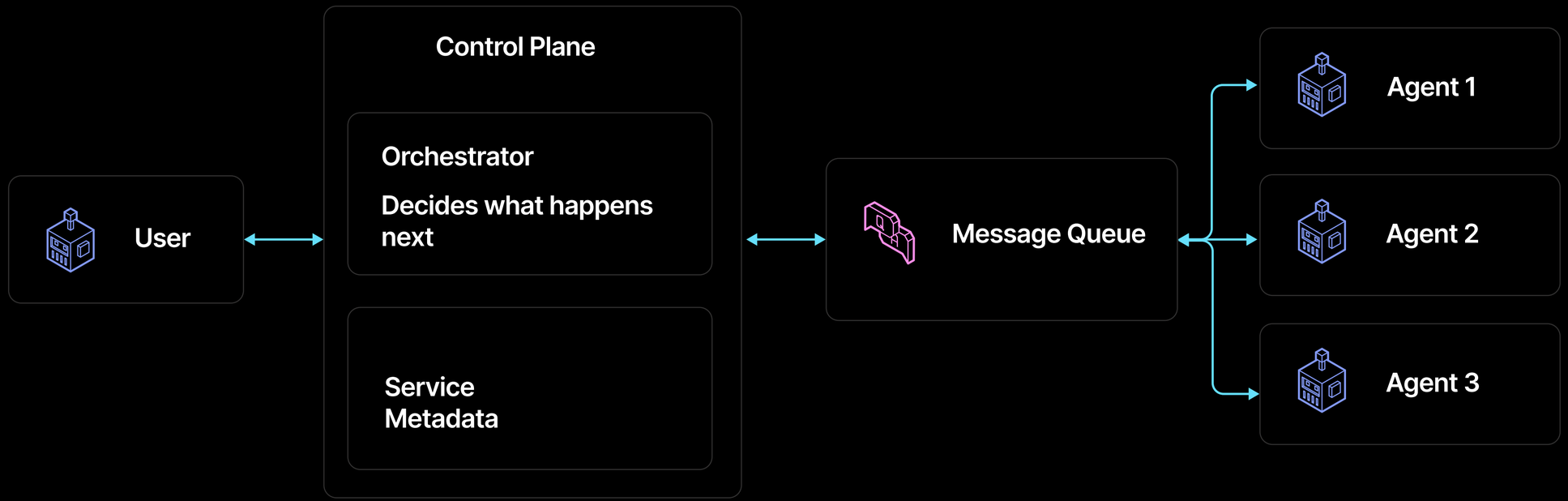
1. Encapsulation and re-use
2. Standardized communication interfaces between agents and with the client.
3. Scalability in number of users and number of agents
4. Human-in-the-loop for the end-user
5. Debugging and observability tools for the developer



llama-deploy

Deploy agentic workflows as **microservices**.

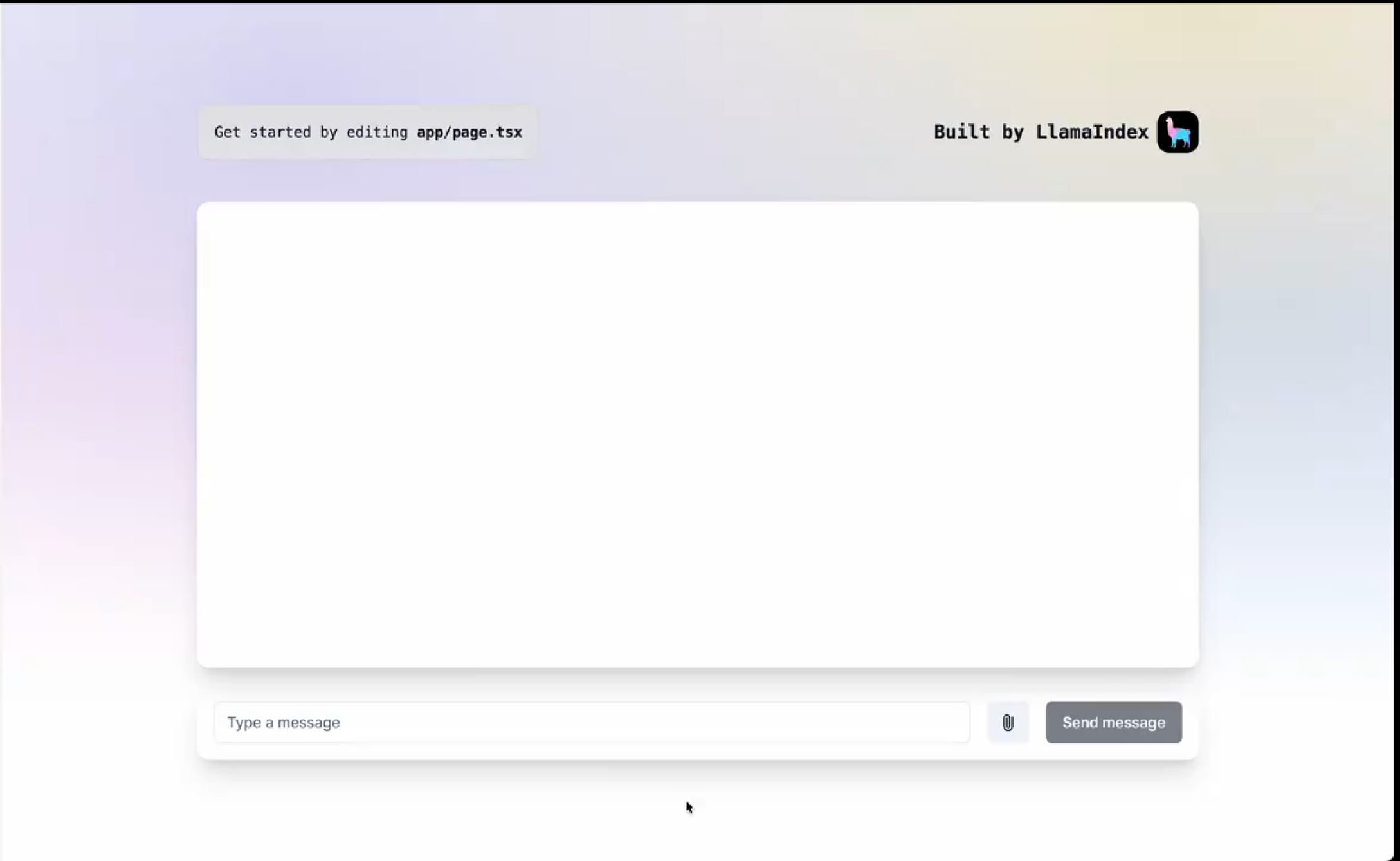
- Model every agent workflow as a service API
- All agent communication occurs via a central message queue
- Distributed tool-execution
- Human-in-the-loop as a service
- Easy deployment with docker-compose and Kubernetes



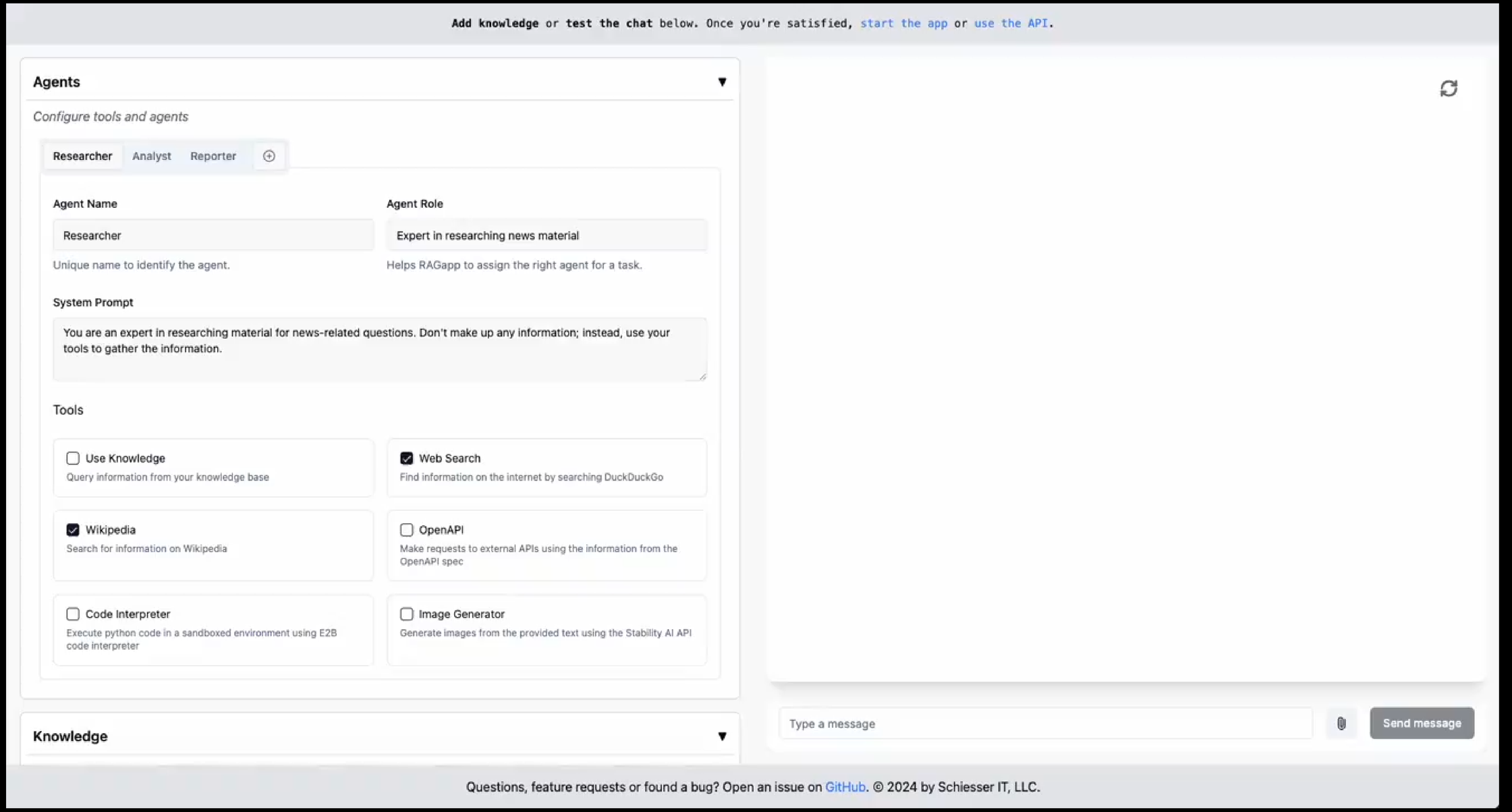
- https://github.com/run-llama/llama_deploy

create-llama

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



RAGapp



https://github.com/ragapp/ragapp

Thank you!

Get in Touch

LlamaCloud Signup