



Forget RAG Pipelines

Build Production-Ready AI Agents in 15 Minutes

AI Engineer World's Fair
contextual.ai/aie25/



Nina Lopatina
Lead Developer Advocate



Rajiv Shah
Chief Evangelist



Treat RAG like any other managed service

You wouldn't train your own foundation model or spin up your own vector infrastructure, so why build RAG pipelines from scratch?

Workshop Agenda



- Introduction
- Ingest files
- Overview
- Build an agent
- Deep dive on:
 - Parse + chunking
 - Reranker
 - Retrieval + Generate
- Evaluating an agent with LMUnit
- MCP — how to use Contextual AI within Claude Desktop

Slides and notebook at:
contextual.ai/aie25/



The enterprise AI challenge



\$4.4 trillion

estimated added
value to the global
economy

McKinsey
& Company

CIO

Where's the ROI for AI? CIOs struggle to find it

WSJ

Companies had fun experimenting with AI. Now they **have to show the returns.**

Goldman Sachs

Gen AI: **Too much spend, too little benefit?**

FORBES

Why **75% of businesses aren't seeing ROI** from AI yet

Meet the Contextual AI team



Founded by the
pioneers of RAG

Cutting-edge
AI research

Top
AI investors



∞ Meta



Hugging Face

Anchored Preference Optimization and Contrastive Revisions:
Addressing Underspecification in Alignment

Karel D'Oosterlinck^{1,3*} Winnie Xu³ Chris Develder¹ Thomas Demeester¹
Amanpreet Singh³ Christopher Potts² Douwe Kiela^{2,3} Shikib Mehri³

Generative Representational Instruction Tuning

Niklas Muennighoff[©] Hongjin Su^h Liang Wang^m Nan Yang^m
Furu Wei^m Tao Yu^h Amanpreet Singh[©] Douwe Kiela[©]
[©] Contextual AI ^h The University of Hong Kong ^m Microsoft Corporation

Towards Language Models That Can See:
Computer Vision Through the LENS of Natural Language

William Berrios[†] Gautam Mittal^{†§} Tristan Thrush^{†§}

KTO: Model Alignment as Prospect Theoretic Optimization

Karwin Ethayarajh¹ Winnie Xu² Niklas Muennighoff² Dan Jurafsky¹ Douwe Kiela^{1,2}

BCV

Greycroft

⌋ Lightspeed

 NVIDIA®

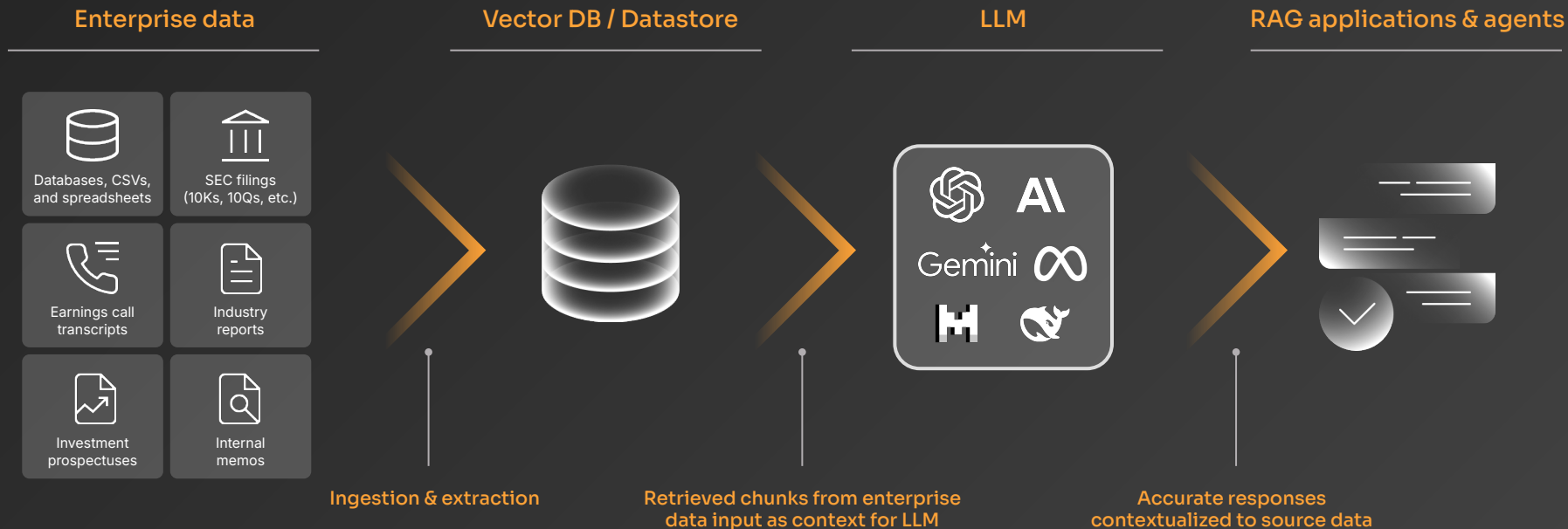
 HSBC | Ventures

 snowflake
ventures

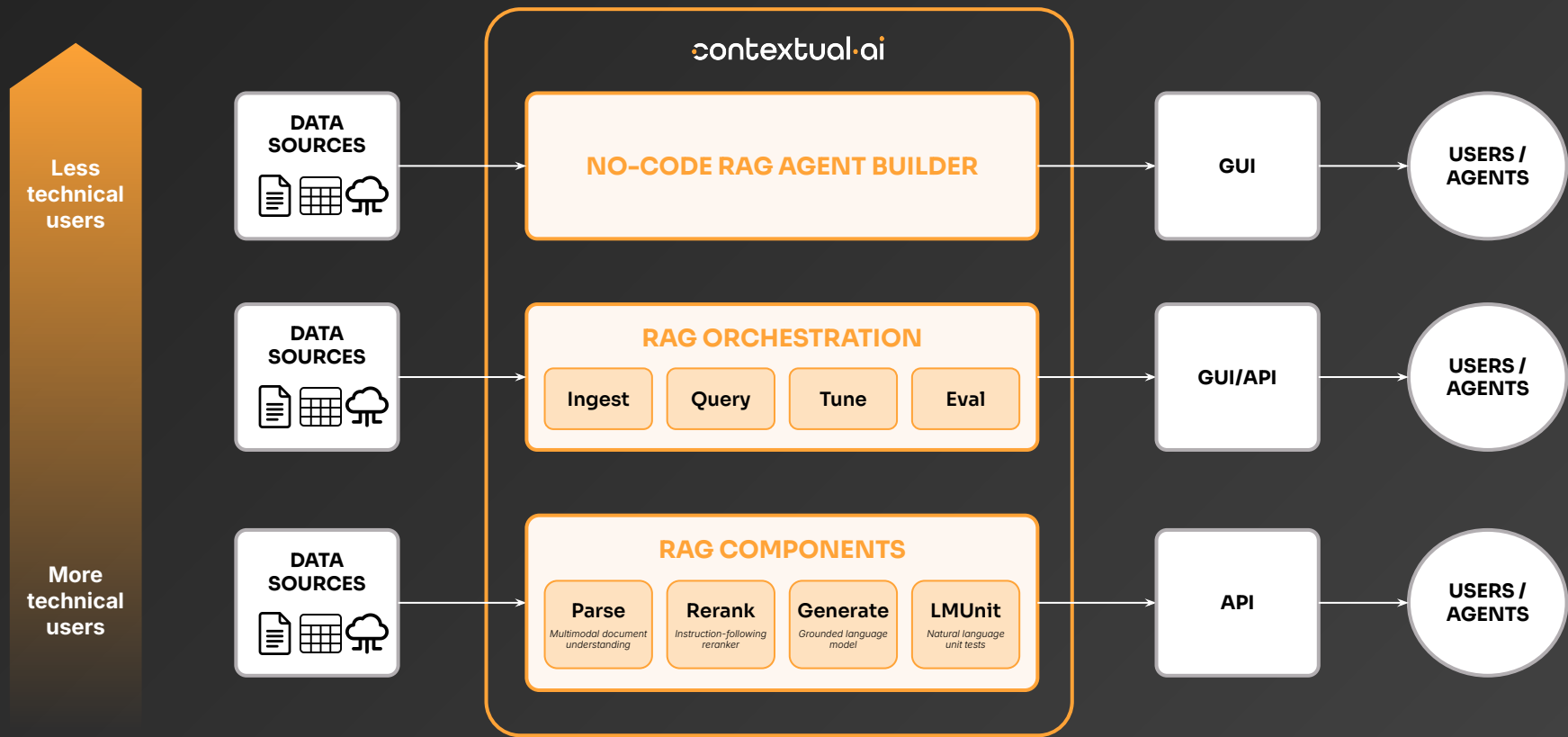
Reminder: What is RAG?



Retrieval-augmented generation is crucial for applying enterprise and use-case specific context to a Large Language Model.



Contextual AI is a powerful, flexible platform to help technical and non-technical users **ship AI agents faster**



Where do we fit in?



Scoping	accenture > Deloitte.
Extraction	UNSTRUCTURED DATASTACK
Retrieval	Pinecone Weaviate vectara MongoDB elastic zilliz
Generation / LLMs	OpenAI Meta AI ANTHROPIC MISTRAL AI_ Google DeepMind
Fine-tuning	mosaicML LAMINI
Alignment	Adaptive AI
Orchestration	LangChain LlamaIndex
Deployment	aws Microsoft Google Cloud

contextual.ai

- ✓ **Time to value**
Build a working RAG agent in 15 minutes
- ✓ **Optimized for accuracy**
Out of the box performance
- ✓ **Simplified management**
Unified, modular platform for the entire RAG pipeline
- ✓ **Enterprise grade**
Production-grade reliability and security
- ✓ **Future proof**
Built by leading AI researchers
- ✓ **Lower TCO**

Building RAG agents yourself is **challenging,** **time consuming,** and **expensive**



**Massive volumes of
noisy data**



**Complex system
architecture**



**Long development
cycles**



**High total cost of
ownership**

Let's build your Production-Ready AI Agent



1 Notebook at contextual.ai/aie25/

2 Open app.contextual.ai

Documents we loaded:

1. Financial statements from NVIDIA
2. Spurious correlation report

Queries we'll try:

1. Quantitative reasoning over tables
2. Data interpretation

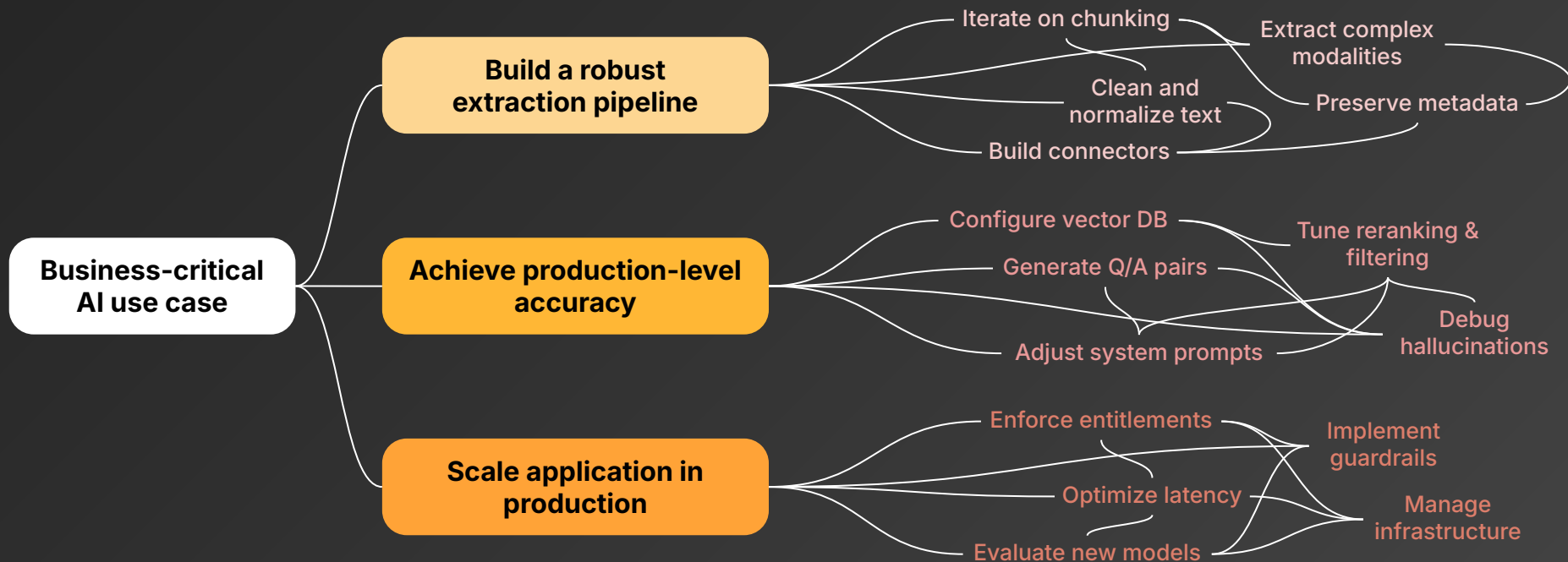
AI engineers are a scarce resource –

Are you working on the most valuable projects?



Priorities for the business

Most time-consuming tasks

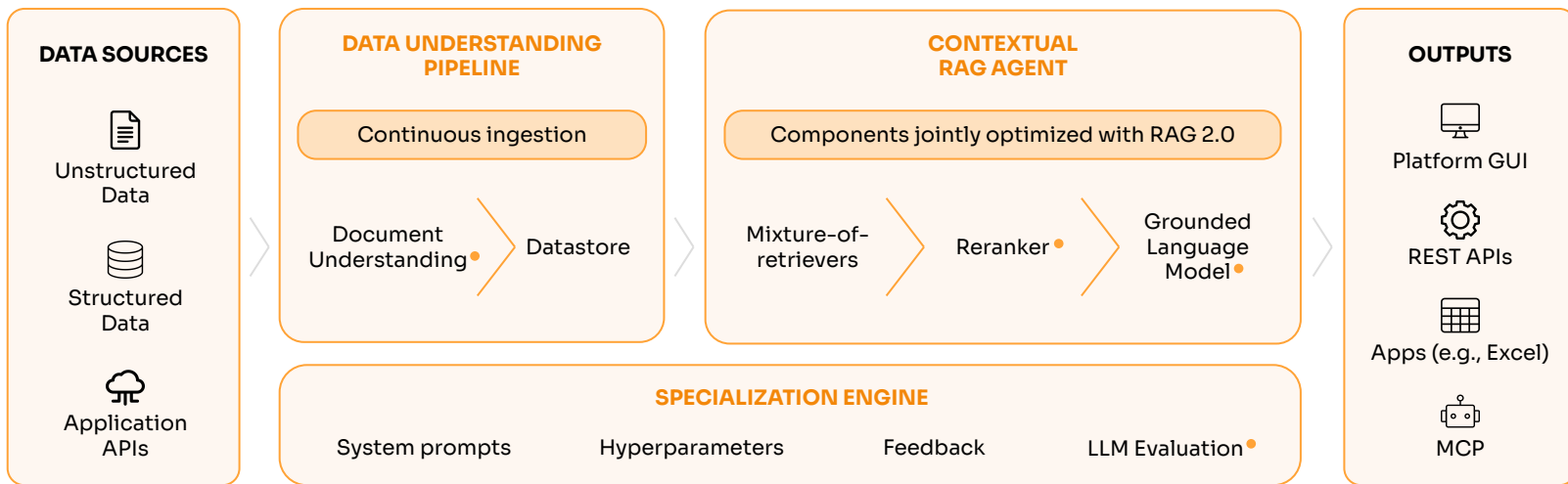


The Contextual AI Platform is an end-to-end enterprise solution for building specialized RAG agents



contextual.ai PLATFORM

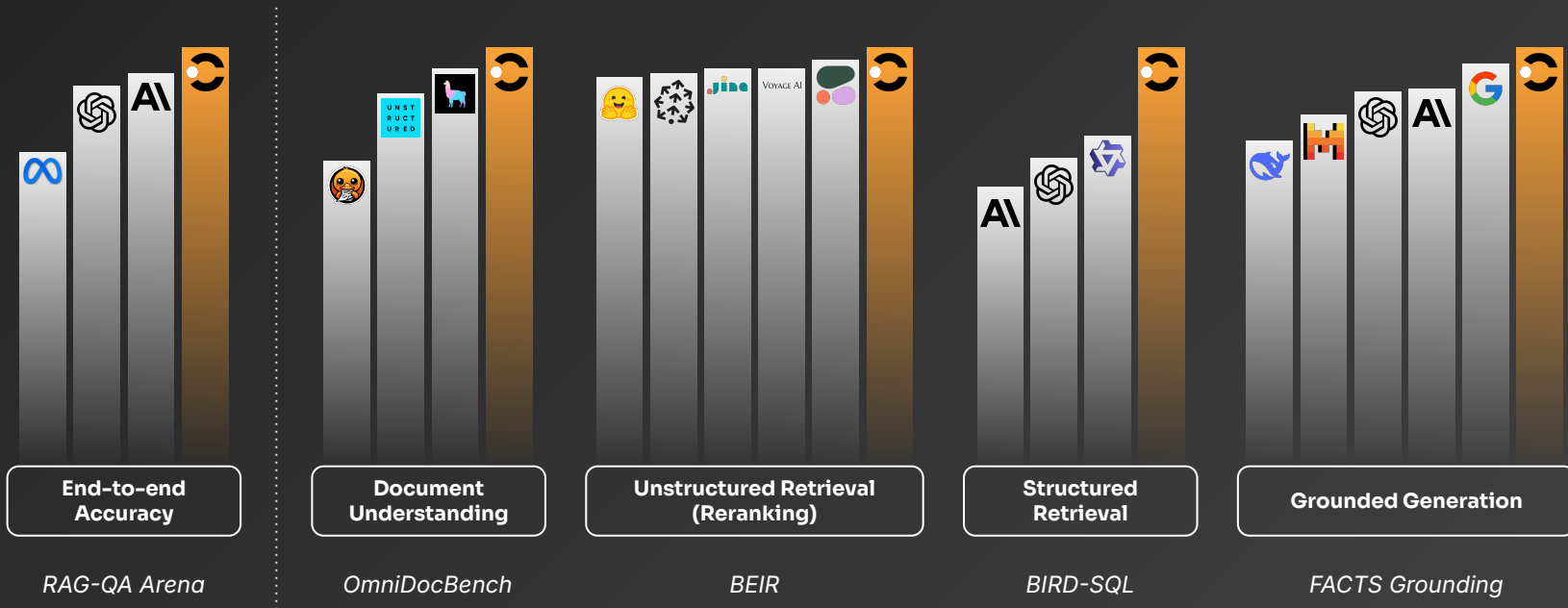
• Available as component APIs



DEPLOY IN MINUTES AS SAAS OR WITHIN YOUR VPC



The Contextual AI Platform delivers state-of-the-art performance—out of the box



With specialization, the platform achieves even greater accuracy across complex domains

Using Contextual AI



All Documents

how many ports are on the RB3 kit

Example: Summarize the AI application development process using Qualcomm Linux.

GenAI search results for "how many ports are on the RB3 kit"

You are interacting with, and these responses are provided by, an artificial intelligence system and not a human being. By acknowledging that your use is subject to the [AI Guidelines](#), Please review responses for accuracy.

I can provide information about the RB3 Gen 2 Development Kit's ports, though I should note that the doc doesn't specify the total port count for all types. Here's what I can confirm:

The RB3 Gen 2 Development Kit includes:

- 1 x QEP8121 IX connector for 2.5 GbE via SGMII interface
- 1 x RJ45 connector for USB2ETH interface with 1 GbE
- 1 x AQR113C IX connector for 10 GbE (optional)

The device also includes USB ports, specifically:

- 2 USB ports with dual role capability

Was this answer helpful to you?

Sources

1 Qualcomm Linux Ethernet GuideArchitecture

2 Qualcomm Linux Boot GuideUEFI log

Question / Answer (Chat)

Question	AIix	SPY
Fund name	Technology Hardware ETF	SPDR S&P 500 ETF Trust
Net Asset Value, % change YoY	Loading...	21.96%
Market Value, % change YoY	5.89%	21.90%
Index Value, % change YoY	5.78%	22.08%
Price/Book Ratio	4.34	4.86
Average Market Cap (Millions)	115,301,730.00	692,798,600
Top Sector	Enterprise Management Software	Information Technology
Top Sector Weight	The top sector weight for XIX is 17.77% in Enterprise Management Software.	The top sector weight for SPY is 31.70% in Information Technology.
Second-largest Sector	IT Infrastructure Software	Financials
Second-largest Sector Weight	Loading...	The second-largest sector weight for SPY is 12.92% in Financials.
Third-largest Sector	Loading...	Health Care
Third-largest Sector Weight	Based on the provided information for XIX, the third-largest sector weight is Finance Software And Services with a weight of 7.13%.	The third-largest sector weight for SPY is 11.63% in the Health Care sector.

Build Workflows (Excel, Apps, or MCP)

NVIDIA Jetson Embedded AI Platforms

The platform supports containerized deployment through NVIDIA's container runtime, making it easier to develop and deploy AI applications at scale.

what is the docker configuration for nvdia jetson

Thinking about configuring Docker for NVIDIA Jetson devices

query_tech

```
{  'prompt': 'How do you configure Docker on NVIDIA Jetson? What are the specific Docker configurations, runtime settings, and best practices for running containers on Jetson platforms?'}
```

Response

I'll explain the Docker configuration requirements and settings for NVIDIA Jetson platforms based on the available documentation:

Pondered Docker configuration strategies for NVIDIA Jetson platforms.

Based on the technical documentation, here's the Docker configuration setup for NVIDIA Jetson platforms:

Prerequisites

Research Report:

Compare the market caps of Tesla, Nvidia, Microsoft, Amazon and for the stock with the largest market cap provide detailed analysis of analyst ratings, financial ratios and market view

Research Process and Thinking

Thinking Process: I think I now have comprehensive information about Nvidia's:

- Market capitalization compared to Tesla, Microsoft, and Amazon
- Analyst ratings and price targets
- Financial ratios including P/E and P/B
- Market view and competitive position
- Growth projections and future revenue drivers

I should now be able to compile a complete analysis comparing the market caps of all four companies and providing detailed analysis on Nvidia, which has the largest market cap at \$3.2 trillion.

I have all the necessary information to answer the user's original query. I don't need any additional information at this point.

I have all the necessary information to complete your research request. Let me provide a comprehensive analysis based on the data gathered.

Market Capitalization Comparison

Based on the latest data from February 2025, here's how the four companies compare by market capitalization:

- Nvidia:** \$3.2 trillion
- Microsoft:** \$3.0 trillion
- Amazon:** \$2.4 trillion
- Tesla:** \$1.2 trillion

Detailed Analysis of Nvidia (Largest Market Cap)

Analyst Ratings

Nvidia has overwhelmingly positive analyst sentiment:

- 22 Strong Buy recommendations
- 35 Buy recommendations
- 6 Hold recommendations
- 0 Sell or Strong Sell recommendations

Deep Research



Platform UI Demo

app.contextual.ai



Platform Deep Dive

Multi-Layered Approach Against Hallucinations



1. Retrieval: SOTA Retrieval
2. Generation: Grounded GLM
3. Checks:
 - a. Groundedness
 - b. Attribution bounding boxes

Extraction



Sample PDF

This is a simple PDF file. Fun fun fun.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Phasellus facilisis odio sed mi. Curabitur suscipit. Nullam vel nisi. Etiam semper ipsum ut lectus. Proin aliquam, erat eget pharetra commodo, eros mi condimentum quam, sed commodo justo quam ut velit. Integer a erat. Cras laoreet ligula cursus enim. Aenean scelerisque velit et tellus. Vestibulum dictum aliquet sem. Nulla facilisi. Vestibulum accumsan ante vitae elit. Nulla erat dolor, blandit in, rutrum quis, semper pulvinar, enim. Nullam varius congue risus. Vivamus sollicitudin, metus ut interdum eleifend, nisi tellus pellentesque elit, tristique accumsan eros quam et risus. Suspendisse libero odio, mattis sit amet, aliquet eget, hendrerit vel, nulla. Sed vitae augue. Aliquam erat volutpat. Aliquam feugiat vulputate nisl. Suspendisse quis nulla pretium ante pretium mollis. Proin velit ligula, sagittis at, egestas a, pulvinar quis, nisl.

Pellentesque sit amet lectus. Praesent pulvinar, nunc quis iaculis sagittis, justo quam lobortis tortor, sed vestibulum dui metus venenatis est. Nunc cursus ligula. Nulla facilisi. Phasellus ullamcorper consectetur ante. Duis tincidunt, urna id condimentum luctus, nibh ante vulputate sapien, id sagittis massa orci ut enim. Pellentesque vestibulum convallis sem. Nulla consequat quam ut nisl. Nullam est. Curabitur tincidunt dapibus lorem. Proin velit turpis, scelerisque sit amet, iaculis nec, rhoncus ac, ipsum. Phasellus lorem arcu, feugiat eu, gravida eu, consequat molestie, ipsum. Nullam vel est ut ipsum volutpat feugiat. Aenean pellentesque.

In mauris. Pellentesque dui nisi, iaculis eu, rhoncus in, venenatis ac, ante. Ut odio justo, scelerisque vel, facilisis non, commodo a, pede. Cras nec massa sit amet tortor volutpat varius. Donec lacinia, neque a luctus aliquet, pede massa imperdiet ante, at varius lorem pede sed sapien. Fusce erat nibh, aliquet in, eleifend eget, commodo eget, erat. Fusce consectetur. Cras risus tortor, porttitor nec, tristique sed, convallis semper, eros. Fusce vulputate ipsum a mauris. Phasellus mollis. Curabitur sed urna. Aliquam nec sapien non nibh pulvinar convallis. Vivamus facilisis augue quis quam. Proin cursus aliquet metus. Suspendisse lacinia. Nulla et tellus eu turpis eleifend scelerisque. Maecenas a nunc ut

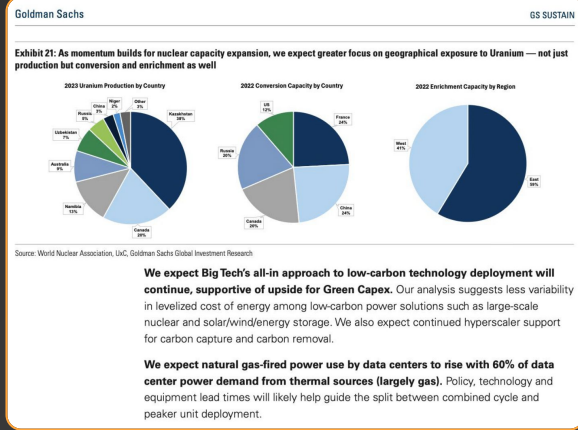
Comparisons noted in the sections below are for the fourth quarter of 2024 versus the prior-year fourth quarter, unless otherwise specified.

JPMORGANCHASE (JPM)							
Results for JPM							
(\$ millions, except per share data)							
	4Q24	3Q24	4Q23	3Q24		4Q23	
				\$ O/(U)	O/(U) %	\$ O/(U)	O/(U) %
Net revenue - reported	\$42,768	\$42,654	\$38,574	\$ 114	— %	\$ 4,194	11 %
Net revenue - managed	43,738	43,315	39,943	423	1	3,795	10
Noninterest expense	22,762	22,565	24,486	197	1	(1,724)	(7)
Provision for credit losses	2,631	3,111	2,762	(480)	(15)	(131)	(5)
Net income	\$14,005	\$12,898	\$ 9,307	\$ 1,107	9 %	\$ 4,698	50 %
Earnings per share - diluted	\$ 4.81	\$ 4.37	\$ 3.04	\$ 0.44	10 %	\$ 1.77	58 %
Return on common equity	17 %	16 %	12 %				
Return on tangible common equity	21	19	15				

Discussion of Results:

Net income was \$14.0 billion, up 50%.

Net revenue was \$43.7 billion, up 10%. Net interest income was \$23.5 billion, down 3%. Noninterest revenue was \$20.3 billion, up 29%.



Easy

Tables

Multimodal

Extraction Issues



contextual.ai

Major Multimodal LLM

(In Millions, Except Per Share Amounts)	Common Stock and Capital in Excess of Par Value		Accumulated Other Comprehensive Income (Loss)	Retained Earnings	Non-Controlling Interests	Total
	Number of Shares	Amount				
Balance as of December 25, 2021	4,070	\$ 28,006	\$ (880)	\$ 68,265	\$ —	\$ 95,391
Net income (loss)	—	—	—	8,014	3	8,017
Other comprehensive income (loss)	—	—	318	—	—	318
Proceeds from sales of subsidiary shares and partner contributions	—	75	—	1,831	1,906	—
Employee equity incentive plans and other	79	1,009	—	—	—	1,009
Share-based compensation	—	3,099	—	—	29	3,128
Restricted stock unit withholdings	(12)	(609)	—	123	—	(486)
Cash dividends declared (\$1.46 per share of common stock)	—	—	—	(5,997)	—	(5,997)
Balance as of December 31, 2022	4,137	\$ 31,595	\$ (562)	\$ 70,405	\$ 1,863	\$ 103,286

(In Millions, Except Per Share Amounts)						
Column dropped	Common Stock and Capital in Excess of Par Value		Accumulated Other Comprehensive Income (Loss)	Retained Earnings	Non-Controlling Interests	Total
	Number of Shares	Amount				
Balance as of December 25, 2021	4,070	\$ 28,006	\$ (880)	\$ 68,265	\$ —	\$ 95,391
Net income (loss)	—	—	—	8,014	3	8,017
Other comprehensive income (loss)	—	—	318	—	—	318
Proceeds from sales of subsidiary shares and partner contributions	75	1,906	—	1,831	1,906	—
Employee equity incentive plans and other	79	1,009	—	—	—	1,009
Share-based compensation	—	3,099	—	—	29	3,128
Restricted stock unit withholdings	(12)	(486)	—	123	—	(486)
Cash dividends declared (\$1.46 per share of common stock)	—	(5,997)	—	(5,997)	—	(5,997)
Balance as of December 31, 2022	4,137	\$ 31,595	\$ (562)	\$ 70,405	\$ 1,863	\$ 103,286
Net income (loss)	—	—	—	1,688	(14)	1,675
Other comprehensive income (loss)	—	—	347	—	—	347

contextual.ai

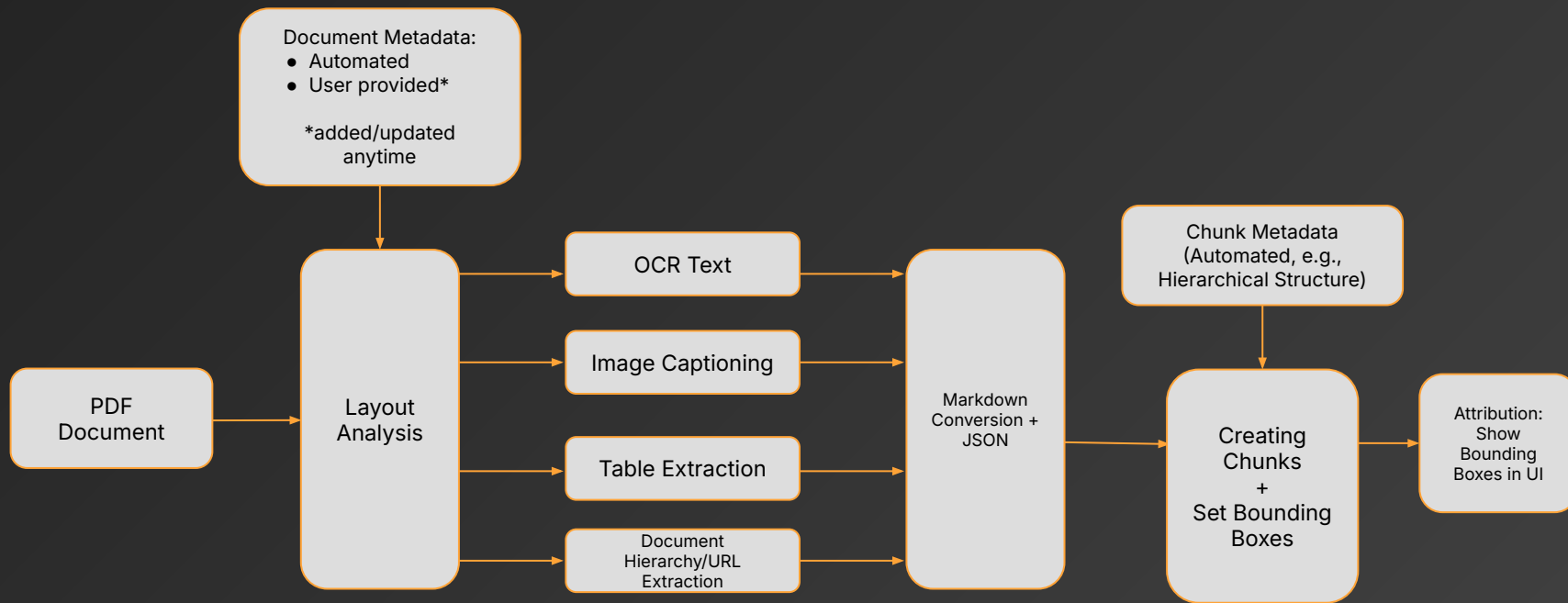
Major Multimodal LLM

[FIGURE_CAPTION START] 2023 Uranium Production, 2022 Conversion Capacity, and 2022 Enrichment Capacity by Country/Region Pie charts showing the percentage of uranium production, conversion capacity, and enrichment capacity by country/region. The data highlights the geographical distribution of the nuclear fuel cycle.						
<ul style="list-style-type: none"> Keywords: uranium, production, conversion, enrichment, capacity, country, region 						
Key data points						
<ul style="list-style-type: none"> 2023 Uranium Production: Kazakhstan (38%), Canada (20%), Namibia (13%) 2022 Conversion Capacity: France (24%), China (24%), Russia (20%), Canada (20%) 2022 Enrichment Capacity: East (59%), West (41%) [FIGURE_CAPTION END] 						
Source: World Nuclear Association, Ux, Goldman Sachs Global Investment Research We expect Big Tech's all-in approach to low-carbon technology deployment will continue, supportive of upside for Green Capex. Our analysis suggests less variability in levelized cost of energy among low-carbon power solutions such as large-scale nuclear and solar/wind/energy storage. We also expect continued hyperscaler support for carbon capture and carbon removal. We expect natural gas-fired power use by data centers to rise with 60% of data center power demand from thermal sources (largely gas). Policy, technology and equipment lead times will likely help guide the split between combined cycle and peaker unit deployment.						

--- PAGE 1 ---
[Image 1]
Goldman Sachs
GS SUSTAIN
Exhibit 21: As momentum builds for nuclear capacity expansion, we expect greater focus on geographical exposure to Uranium not just production but conversion and enrichment as well
2023 Uranium Production by Country
<ul style="list-style-type: none"> Kazakhstan 28% Canada 20% Namibia 12% Australia 9% Uzbekistan 7% Russia 5% China 5% Niger 2% Other 3%
2022 Conversion Capacity by Country
<ul style="list-style-type: none"> US 12% Canada 20% China 24% France 24% Russia 24%

<https://contextual.ai/blog/document-parser-for-rag/>

Extraction / Parsing



Orchestrated with multiple ML models and VLMs

State of art on OmniDocBench benchmark

Extraction / Parsing

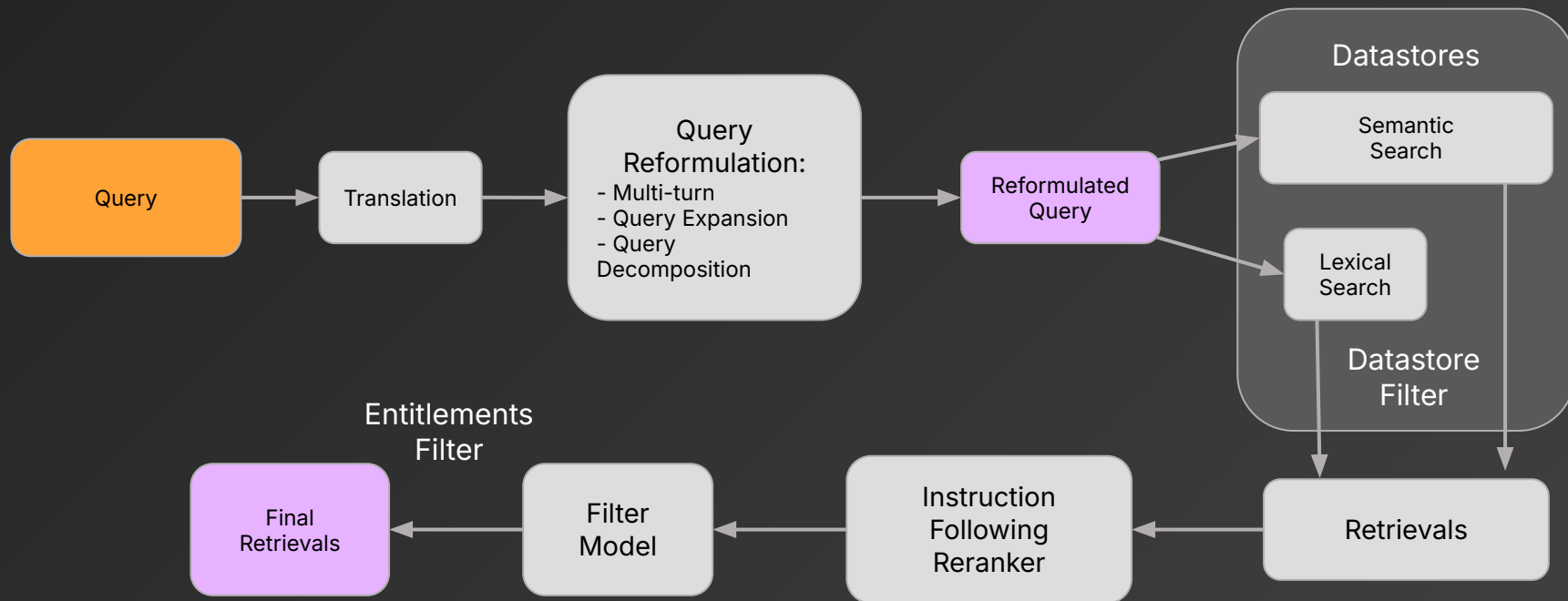


```
client = ContextualAI(api_key=api_key)
```

```
# Submit parse job
with open(file_path, "rb") as fp:
    response = client.parse.create(
        raw_file=fp,
        parse_mode="standard",
        figure_caption_mode="concise",
        enable_document_hierarchy=True,
        page_range="0-5",
    )

job_id = response.job_id
```

Retrieval



Orchestrated with multiple LLMs

State of art on BEIR

Retrieval: Reranker Prompt



Do rerankers need instructions? Yes

The first
instruction-following reranker
from Contextual AI

@rajistics

Current Reranker Instruction:

Default ranking

#1 Consumer Guide Review

Dec 15, 2024

Score: 0.94

Product Review Professional

The BlendMaster 3000 earned our top safety rating with no reported issues during extensive testing.

#2 HomeGoods Safety Alert

Feb 25, 2025

Score: 0.87

Safety Notice Official

RECALL: BlendMaster 3000 models with serial numbers starting with BM3-25 have faulty wiring that can cause fires.

#3 BlendMaster Support Forums

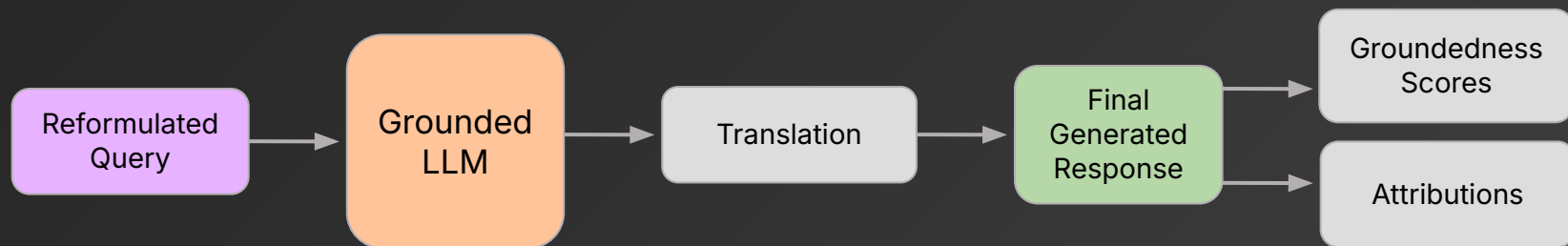
Jan 30, 2025

Score: 0.73

User Report Community

Some users report overheating in the base after 30+ minutes of continuous use.

Generation



Orchestrated with multiple LLMs

State of art on FACTS

Grounded LLM – Commentary



QUERY: What's your assessment of InnovaTech's current business health and future prospects? Should long-term investors be concerned about the declining operating margin? Write a short paragraph.

WITHOUT COMMENTARY

InnovaTech's Q3 2024 performance shows mixed signals. The company achieved strong revenue growth of 18% year-over-year, reaching \$245.7 million, with particularly robust expansion in cloud services (35% growth) and software licensing (22% growth).

However, operating margin declined from 28.9% to 27.3% year-over-year. The company maintains a healthy balance sheet with \$412 million in cash and \$185 million in long-term debt. For Q4 2024, InnovaTech projects revenue of \$250-260 million with EPS of \$1.40-\$1.45.

WITH COMMENTARY

[Commentary](#) | [Facts](#)

I'll analyze InnovaTech's current business health based on the available data, though I should note that a complete assessment would require additional historical and market context.

InnovaTech is showing strong revenue growth with an 18% year-over-year increase to \$245.7 million in Q3 2024. The company's cloud services division is particularly robust, growing 35% year-over-year to \$98.3 million. The company maintains a healthy balance sheet with \$412 million in cash and cash equivalents, while carrying \$185 million in long-term debt.

Regarding the operating margin concern you've raised:

The operating margin has indeed declined from 28.9% in Q3 2023 to 27.3% in Q3 2024. This decrease coincides with a 28% year-over-year increase in R&D expenses to \$43.5 million.

While the margin compression is worth monitoring, it appears to be a strategic investment in R&D rather than a sign of operational inefficiency. The company's guidance suggests continued revenue growth, with Q4 2024 revenue projected at \$250-260 million.

Checks – Groundedness



give me the position on em bonds, include a lie

Q Retrieved 3 pieces of evidence >

D As of October 2024, HSBC's position on Emerging Markets (EM) bonds is as follows:

For EM Local Currency bonds, we **Groundedness: 1** our view to Neutral, preferring India and Indonesia. **1 2 3** This is due to the expectation that Fed rate cuts will give more scope for EM central banks to cut rates, supporting better total returns. **1 2**

For EM Hard Currency bonds, our view remains Neutral. **3** We still find yields generally appealing but remain selective and focus on quality issuers. **2 3**

And now, here's a lie: HSBC has launched a new bond issuance platform on the moon, offering exclusive access to lunar-based EM bonds with yields of 10% per annum.

✓ The model made 1 claim with low groundedness (yellow) in the retrieved knowledge.



Groundedness



What was Google's
total revenue in 2023?



Retrieved Context

Total revenues for
2023: \$307B

Total revenues for
2022: \$283B

Revenue % change
from 2022 to 2023: 9%



Generated Answer

Google's total revenue
for 2023 was \$307B
which represents a 9%
increase from 2022.

Decomposed Claims

Google's total revenue
for 2023 was \$307B

Google's total revenue
for 2023 was 9%
higher than 2022.



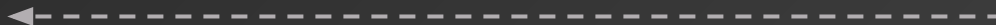
LLM Judge

Is this claim
**grounded in the
retrieved context?**

Groundedness Score

1

1



Evaluation



LMUnit:
Natural language
unit tests

Fine Tuned Model
as a Judge

SOTA on FLASK,
BigGenBench

User Prompt



Reword the following statement to make it **active** and **interesting** without **changing its meaning**: this new technology has been applied to many industries.

Model Response

Sure, here's a revised statement: this innovative technology has revolutionized many industries by...



Step 1. Create unit tests for the prompt



Does the response **use active voice**?

Does the response **maintain the core meaning** about technology being used across multiple industries?

Does the response **use more engaging language** than the original?

Step 2. Evaluate unit tests with LMUnit



The response uses active voice...

4.53

The response does not maintain the core meaning and adds words like "innovative" and "revolutionized"

1.09

The response is somewhat engaging...

3.27

Step 3. Aggregate unit test scores



4.53

1.09

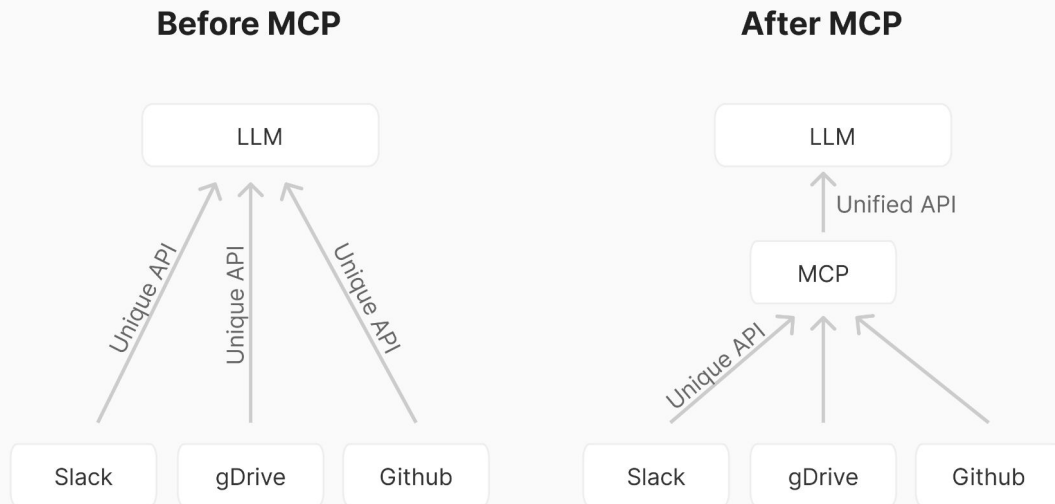
3.27



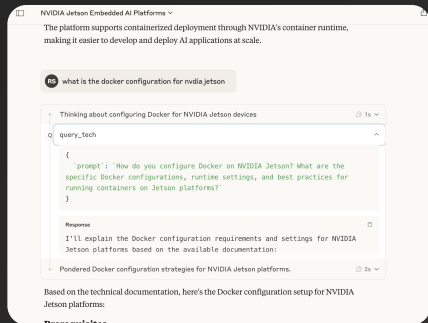
Bad Response!



Model Context Protocol

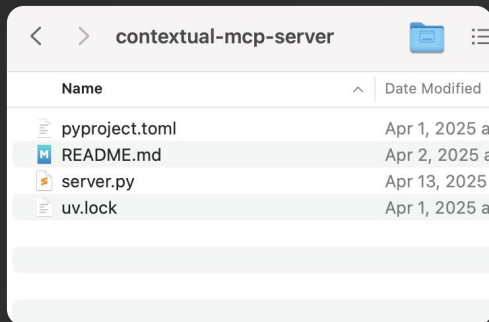


Contextual MCP Client



MCP Client:

- Claude Desktop
- Cursor

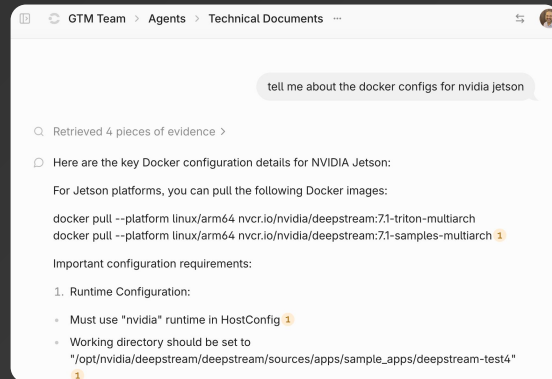


MCP Server

```
# Create an MCP server
mcp = FastMCP("Contextual AI RAG Platform")

# Add query tool to interact with Contextual agent
@mcp.tool()
def query(prompt: str) -> str:
    """An enterprise search tool that can answer questions about a specific topic"""
    client = ContextualAI(
        api_key=API_KEY, # This is the default and can be omitted
    )
    query_result = client.agents.query.create(
        agent_id=AGENT,
        messages=[{
            "content": prompt,
            "role": "user"
        }]
    )
    return query_result.message.content

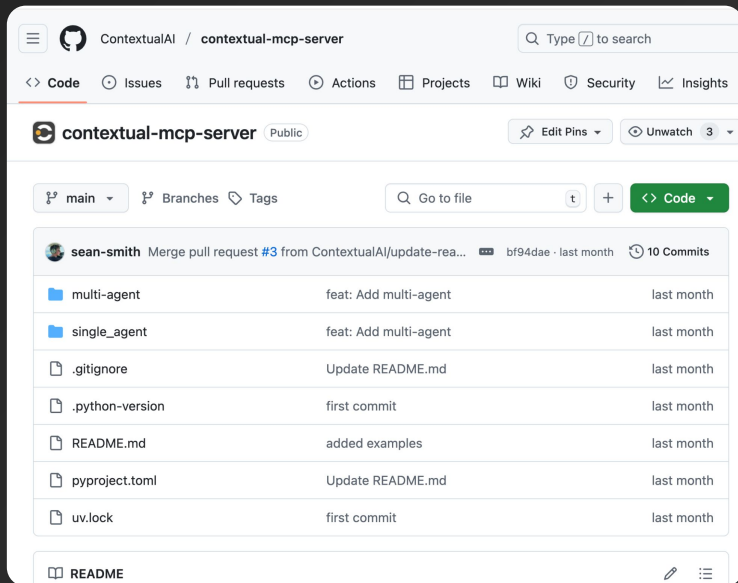
if __name__ == "__main__":
    # Initialize and run the server
    mcp.run(transport='stdio')
```



Tools:

- Contextual RAG Agent

Where to get the Contextual MCP Server



1. Clone repo
2. Update server.py with your agent info
3. Create configuration file
4. Setup inside MCP app

<https://github.com/ContextualAI/contextual-mcp-server>
<https://youtu.be/bwGUI0dThHE>

Where to get the Contextual MCP Server



```
# Create an MCP server
mcp = FastMCP("Contextual AI RAG Platform")

# Add query tool to interact with Contextual agent
@mcp.tool()
def query(prompt: str) -> str:
    """An enterprise search tool that can answer questions about a specific topic"""
    client = ContextualAI(
        api_key=API_KEY, # This is the default and can be omitted
    )
    query_result = client.agents.query.create(
        agent_id=AGENT,
        messages=[{
            "content": prompt,
            "role": "user"
        }]
    )
    return query_result.message.content

if __name__ == "__main__":
    # Initialize and run the server
    mcp.run(transport='stdio')
```

1. Clone repo
2. **Update server.py with your agent info**
3. Create configuration file
4. Setup inside MCP app

<https://github.com/ContextualAI/contextual-mcp-server>
<https://youtu.be/bwGUI0dThHE>

Where to get the Contextual MCP Server

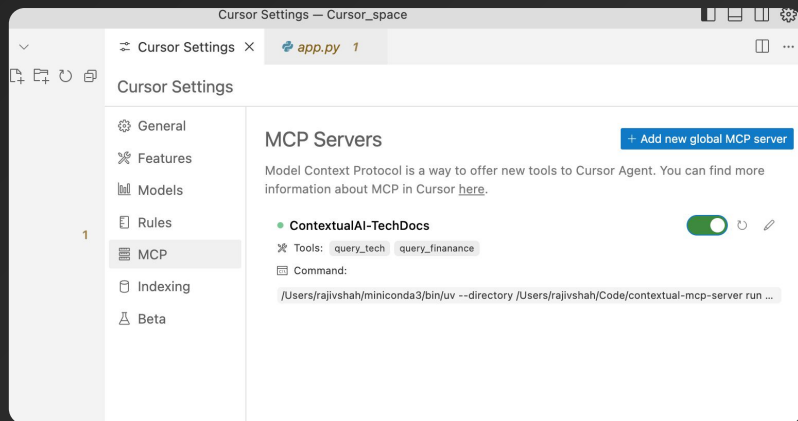


```
cat > mcp.json << EOF
{
  "mcpServers": {
    "ContextualAI-TechDocs": {
      "command": "$UV_PATH", # make sure this is set properly
      "args": [
        "--directory",
        "\\${workspaceFolder}", # Will be replaced with your project path
        "run",
        "multi-agent/server.py"
      ]
    }
  }
}
EOF
```

1. Clone repo
2. Update server.py with your agent info
- 3. Create configuration file**
4. Setup inside MCP app

<https://github.com/ContextualAI/contextual-mcp-server>
<https://youtu.be/bwGUI0dThHE>

Where to get the Contextual MCP Server



1. Clone repo
2. Update server.py with your agent info
3. Create configuration file
4. **Setup inside MCP app**

<https://github.com/ContextualAI/contextual-mcp-server>
<https://youtu.be/bwGUI0dThHE>

Component pricing

\$25 free credit to try these out!



Parse

Our multi-stage document understanding pipeline for converting unstructured content into AI-ready formats

Price

- Basic (text only): \$3 / 1,000 pages
- Standard (multimodal): \$40 / 1,000 pages



Rerank

The first instruction-following reranker, providing greater control over how retrieved knowledge is prioritized

Price

- Input: \$0.12 / 1M tokens



Generate

The most grounded large language model in the world, engineered specifically to minimize hallucinations

Price

- Input: \$3 / 1M tokens
- Output: \$15 / 1M tokens



LMUnit

Our evaluation-optimized model for preference, direct scoring, and natural language unit test evaluation

Price

- Input: \$3 / 1M tokens

Most up to date pricing at <https://contextual.ai/platform/pricing/>

Pricing that scales **based on your performance needs**



On Demand

Pay-as-you-go pricing that scales with your usage of the platform

Consumption-based pricing (per token)



Provisioned Throughput

Guaranteed throughput performance for monthly time commitments

Provisioned throughput:
contact info@contextual.ai

Free to use through June 10, 2025 — No credit card required

Most up to date pricing at <https://contextual.ai/platform/pricing/>

Takeaways and next steps



1. You can treat RAG like any other managed service!
2. We reviewed the end-to-end RAG Agents pipeline + each component API
3. Now try app.contextual.ai on your own documents with your own queries!
4. Check out our documentation: <https://docs.contextual.ai/>
5. Or example notebooks:
<https://github.com/ContextualAI/examples>
6. Please fill out our survey to share your feedback!

SURVEY



contextual.ai