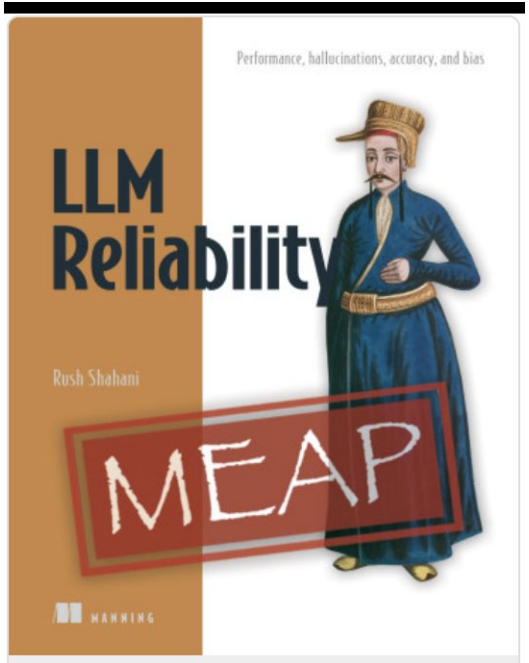
Research to Reality: Building Production-Ready LLM Apps Users Can Trust

Rush Shahani
CTO & Co-Founder
Persana AI (YC W23)

Rush Shahani

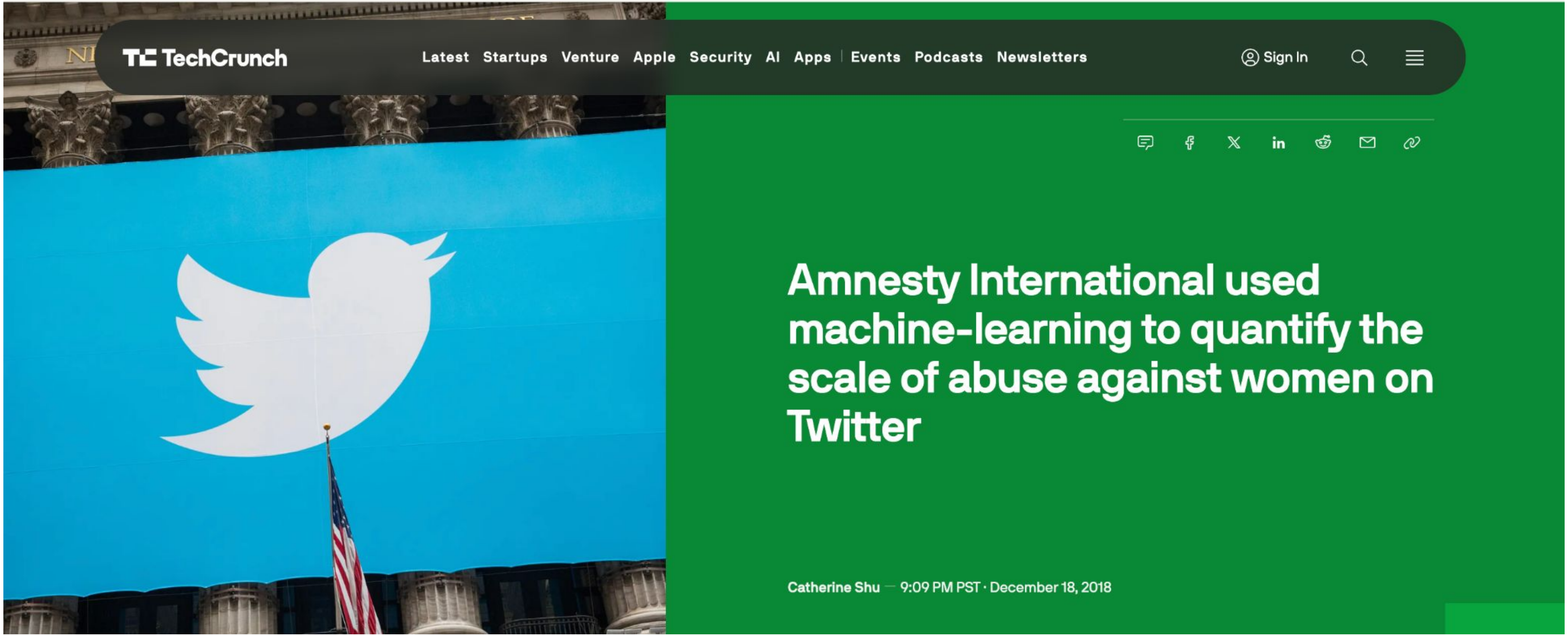
Cofounder & CTO at Persana AI
Author of LLM Reliability (Manning Publications)

- CTO & Co-Founder, Persana AI (YC W23)
- AI @: LinkedIn, Element AI, Shopify
- Author of '**LLM Reliability**'
- Focus: Building scalable, effective and trustworthy AI systems



bit.ly/llmreliability

Coupon - **ACM45** (40% off ACM exclusive)



Amnesty International used machine-learning to quantify the scale of abuse against women on Twitter

Catherine Shu — 9:09 PM PST · December 18, 2018

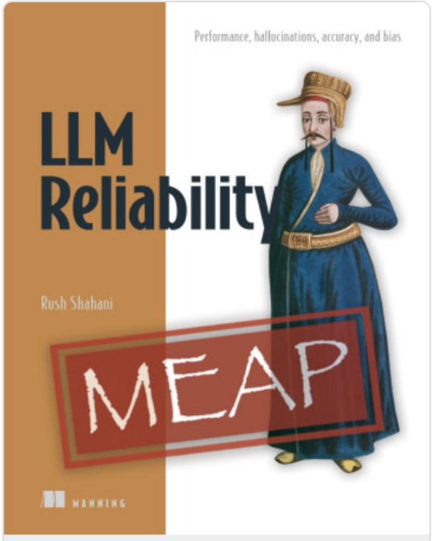
About Persana AI (YC W23)

- Our mission is to amplify GTM teams with the power of AI Agents to convert high intent leads and automate all manual processes.
- Intent is **scattered across millions of touch-points and is impossible to analyze manually**
- Use the power of LLMs + ranking models to identify high intent signals across all different data sources

salesforce
HubSpot

Agenda

1. LLM Landscape & Introduction to LLM Reliability
2. Understanding Hallucinations
3. Basic & Advanced Prompting Techniques
4. RAG Architecture & Implementation
5. AI Agents Architecture
6. Agentic RAG Demo
7. Evaluating RAG & Agentic Systems
8. Bias & Ethics
9. Key Takeaways



LLM Landscape and Why Reliability?

- Rapid advancements: GPT-4, Claude, LLaMA
- Applications: Coding, customer support, research
- Production deployment of LLMs introduce critical challenges such as hallucinations, harmful bias, performance, lack of efficiency.

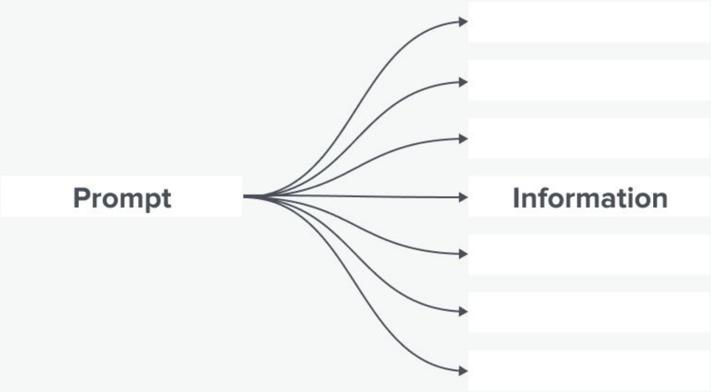
“This is really a moment where productivity is no longer tied to workforce growth, but through this intelligent technology that can be scaled without limits.” - Marc Benioff

The Harvard Business Review estimates that almost 80% of generative AI projects end up in failure

The AI Revolution

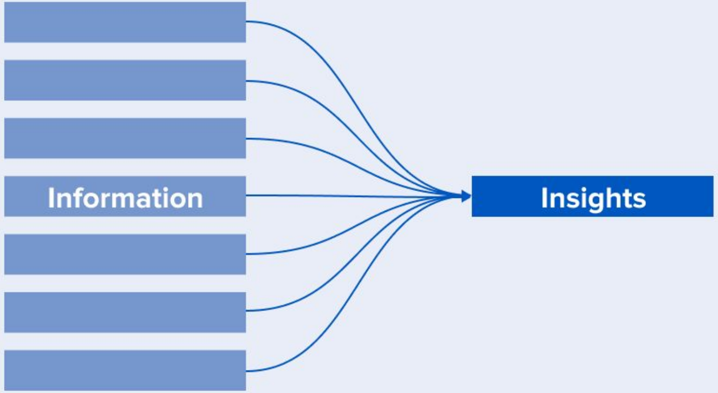
Wave 1: Generative (GenAI)

From prompts to long-form responses



Wave 2: Synthesis (SynthAI)

From vast volumes of information to insights

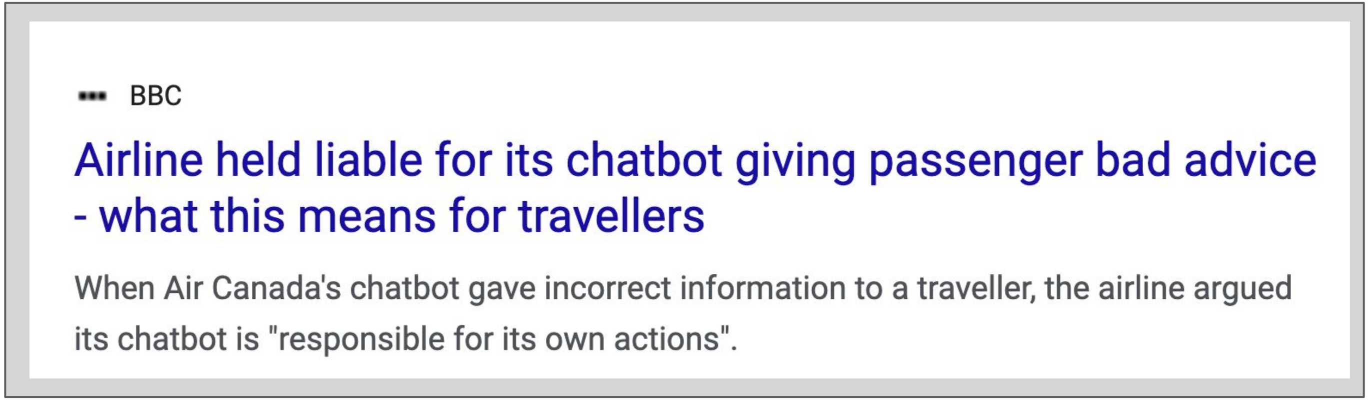


- UX is natural language prompt and response
- Benefit is fun and shareability
- Generic data sets / foundational models

- UX is embedded into existing workflows or completely automated
- Benefit is improvement to speed and/or quality
- Domain-specific data sets / fine-tuned models

The Hallucination Problem

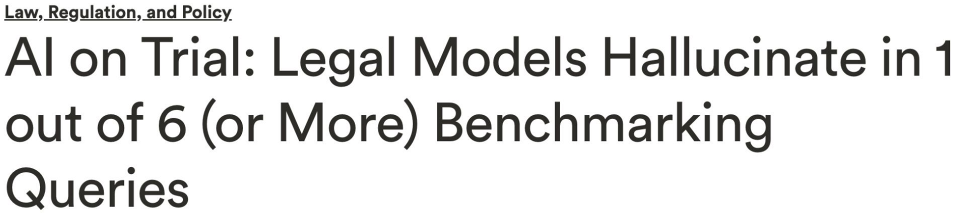
- Hallucinations: Generating fabricated or inaccurate information



BBC

Airline held liable for its chatbot giving passenger bad advice - what this means for travellers

When Air Canada's chatbot gave incorrect information to a traveller, the airline argued its chatbot is "responsible for its own actions".



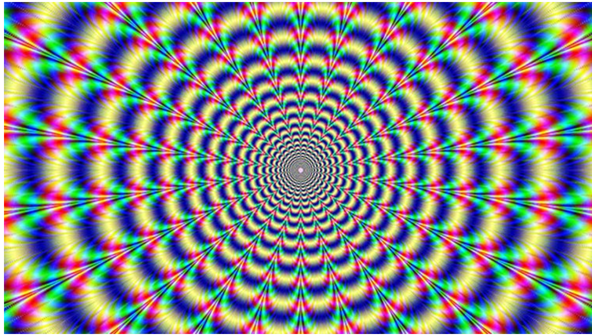
Law, Regulation, and Policy

AI on Trial: Legal Models Hallucinate in 1 out of 6 (or More) Benchmarking Queries

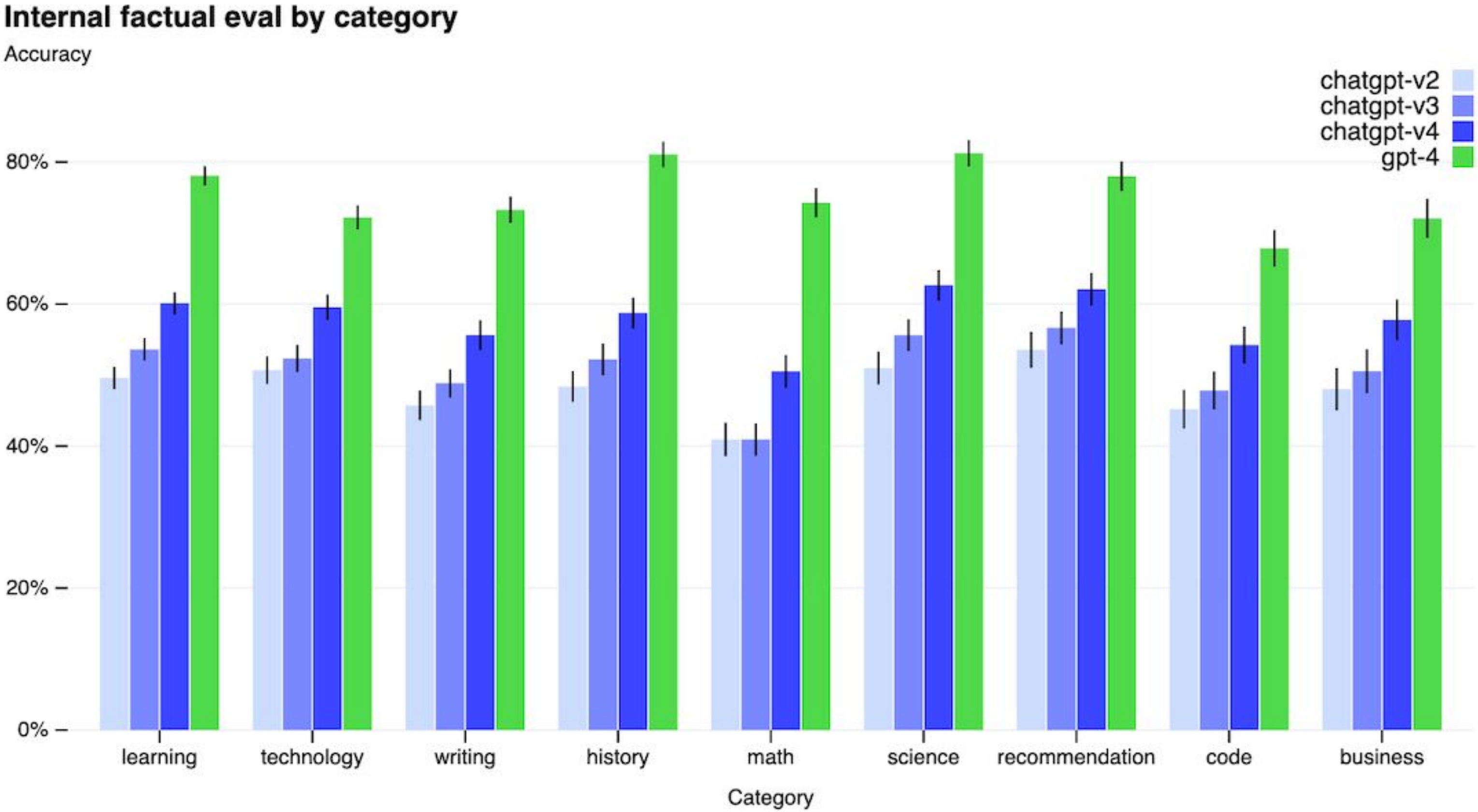
What Causes Hallucinations?

- Causes:
    - Training data limitations
    - Lack of real-time retrieval
    - Overconfidence in uncertain responses
- Impact: Erodes trust and is dangerous especially for high-stakes domains like healthcare, law

Ways to Reduce Hallucinations



- **Choice of LLM** reduces hallucinations (eg. GPT-4 reduces hallucinations compared to GPT-3.5)
- **Model Parameters**
- **Advanced Prompting techniques**
- **Fine-tuning** can make the model grounded
- **Retrieval Augmented Generation**
- **Agents and tools**



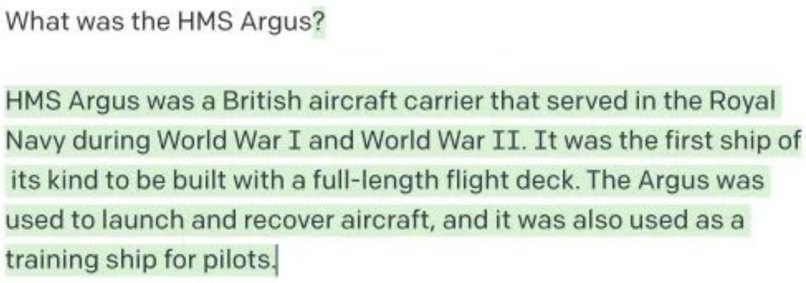
On nine categories of internal adversarially-designed factual evals, we compare GPT-4 (green) to the first three ChatGPT versions. There are significant gains across all topics. An accuracy of 1.0 means the model's answers are judged to be in agreement with human ideal responses for all questions in the eval.

Basic Prompting Tips to Reduce Hallucinations

- Use assertive voice if you want a model obey certain guidelines
- Reiterate the rules at the end of the prompt
- Be descriptive and use examples (few-shot prompting)
- Add more context in the prompt and ask the model to act as an expert for our field (eg. Act as a Expert Lawyer)
- Ask the model to substantiate its claims (Citations)

Adjusting Model Parameters

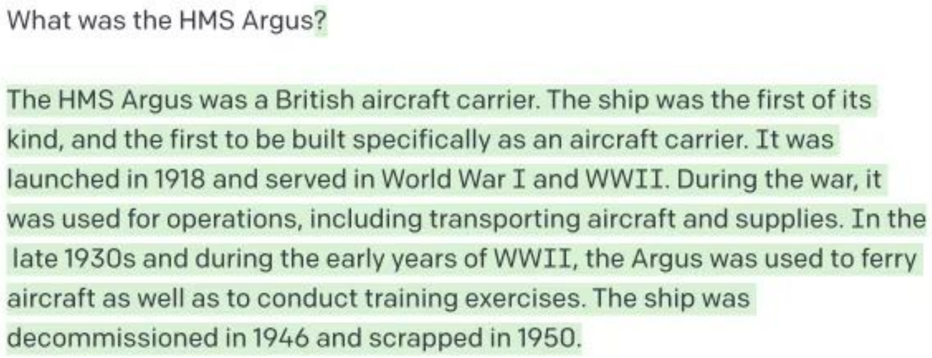
Temperature = 0



What was the HMS Argus?

HMS Argus was a British aircraft carrier that served in the Royal Navy during World War I and World War II. It was the first ship of its kind to be built with a full-length flight deck. The Argus was used to launch and recover aircraft, and it was also used as a training ship for pilots.

Temperature = 1

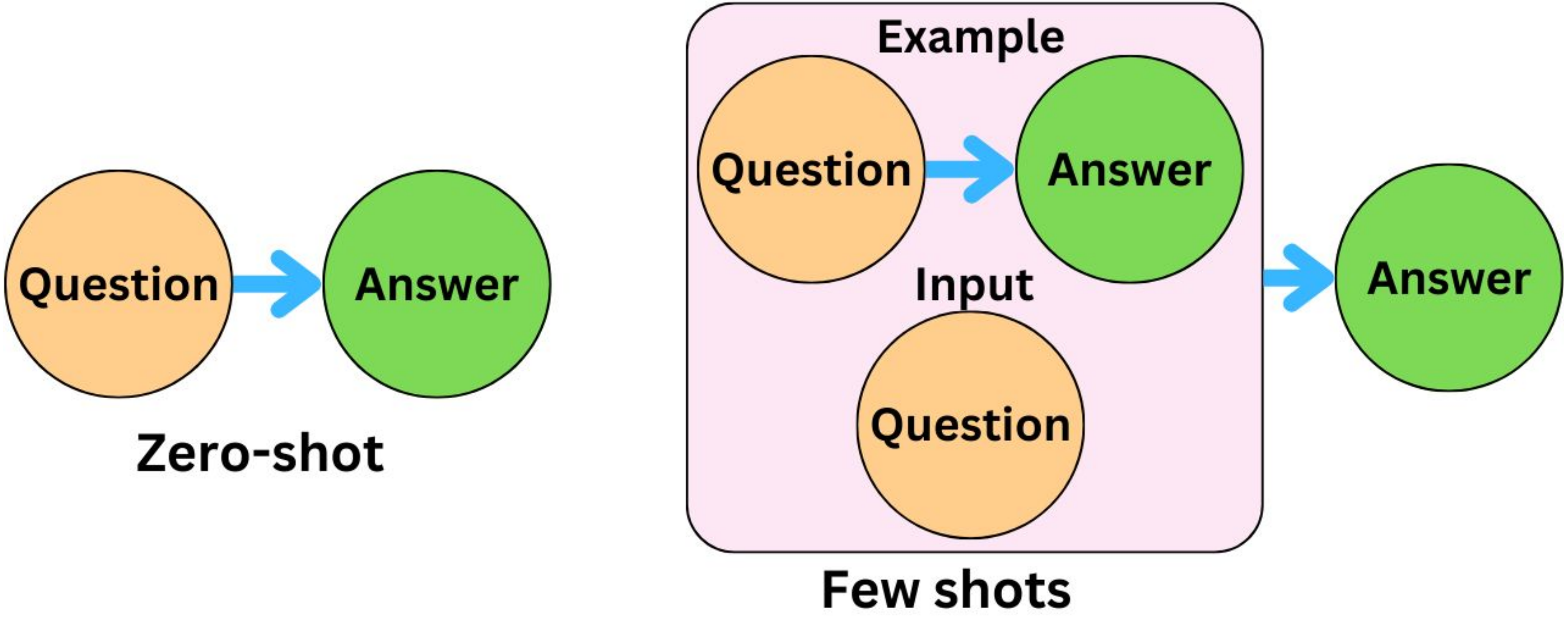


What was the HMS Argus?

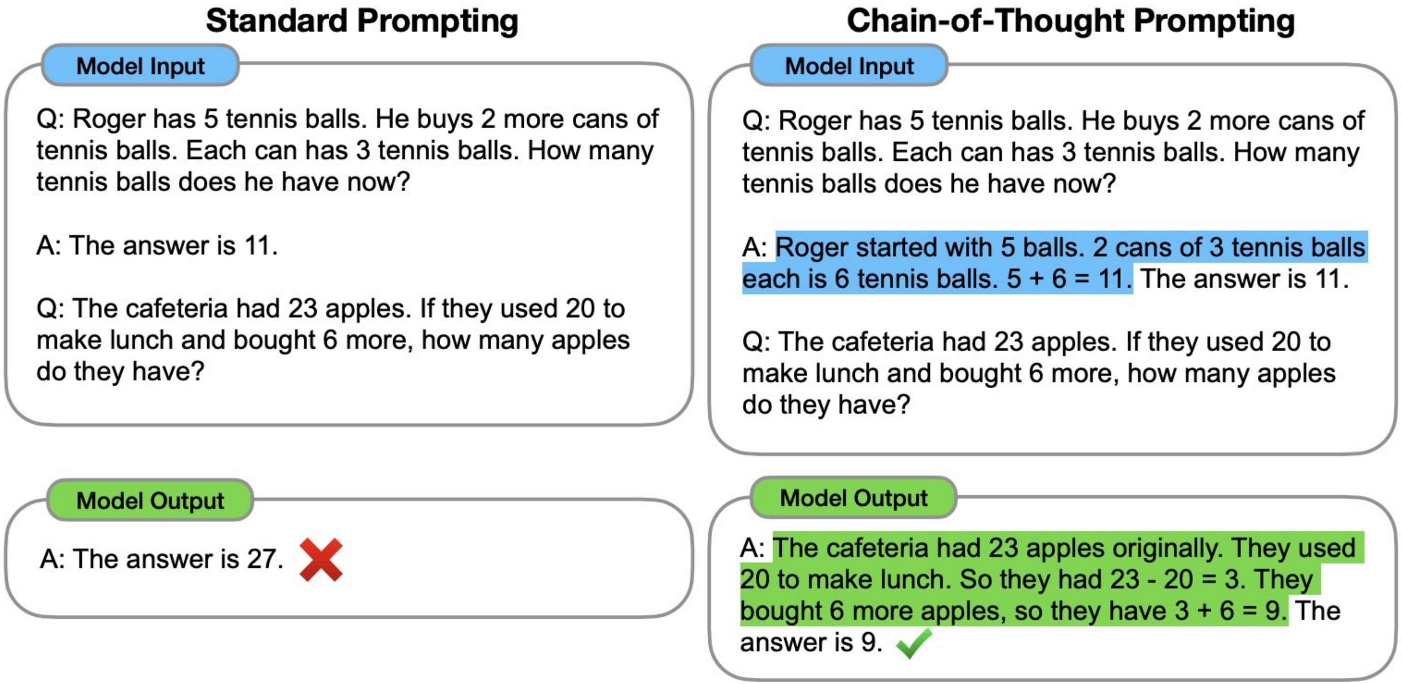
The HMS Argus was a British aircraft carrier. The ship was the first of its kind, and the first to be built specifically as an aircraft carrier. It was launched in 1918 and served in World War I and WWII. During the war, it was used for operations, including transporting aircraft and supplies. In the late 1930s and during the early years of WWII, the Argus was used to ferry aircraft as well as to conduct training exercises. The ship was decommissioned in 1946 and scrapped in 1950.

temperature=1 returns incorrect dates

Zero Shot vs. Few Shot Prompting



Chain-of-Thought Prompting



Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

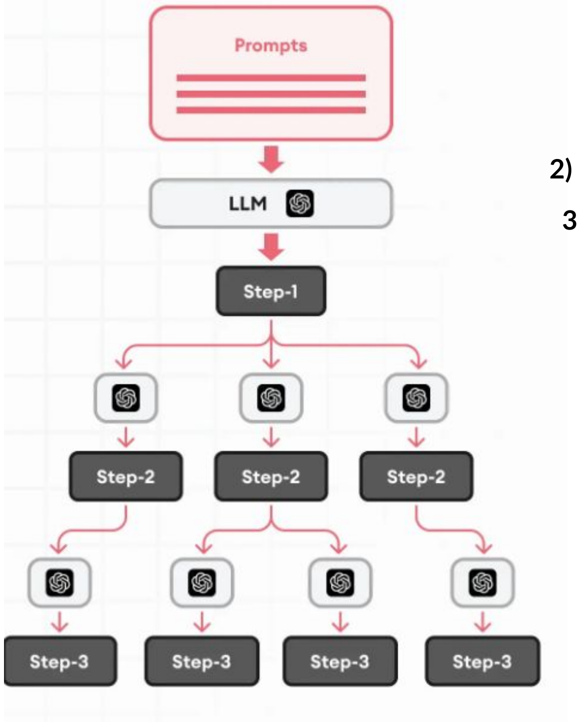
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Tree of Thought Prompting

Sudoku is a perfect example of this

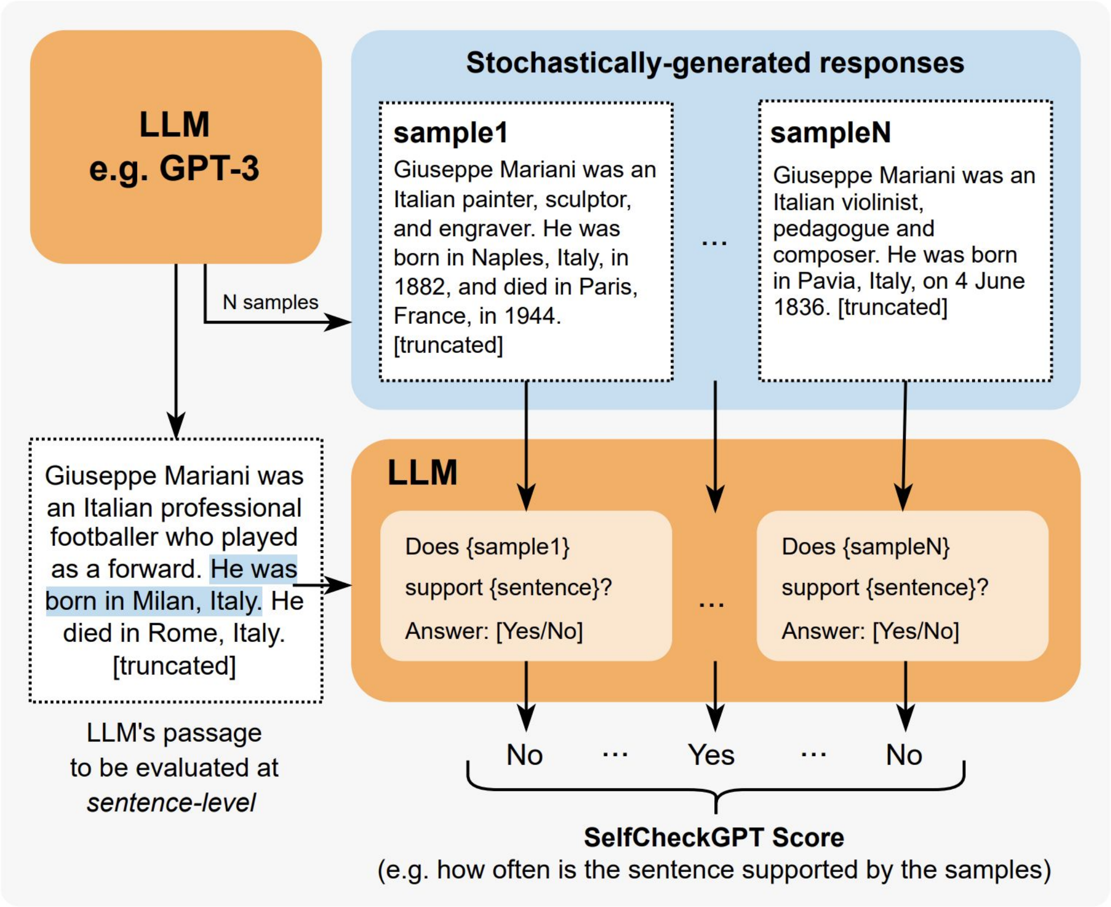


Another example is a customer telling a chatbot that his TV is malfunctioned

The LLM can evaluate 3 options:

1) Return it if its in the 30 day window
2) Manufacturer 1 year warranty if its outside of 30 day
3) Replace it for a new one if it's within 90 days of purchase

SelfCheckGPT



LLM
e.g. GPT-3

N samples

Stochastically-generated responses

sample1
Giuseppe Mariani was an Italian painter, sculptor, and engraver. He was born in Naples, Italy, in 1882, and died in Paris, France, in 1944. [truncated]

...

sampleN
Giuseppe Mariani was an Italian violinist, pedagogue and composer. He was born in Pavia, Italy, on 4 June 1836. [truncated]

Giuseppe Mariani was an Italian professional footballer who played as a forward. He was born in Milan, Italy. He died in Rome, Italy. [truncated]

LLM's passage to be evaluated at *sentence-level*

LLM

Does {sample1} support {sentence}? Answer: [Yes/No]

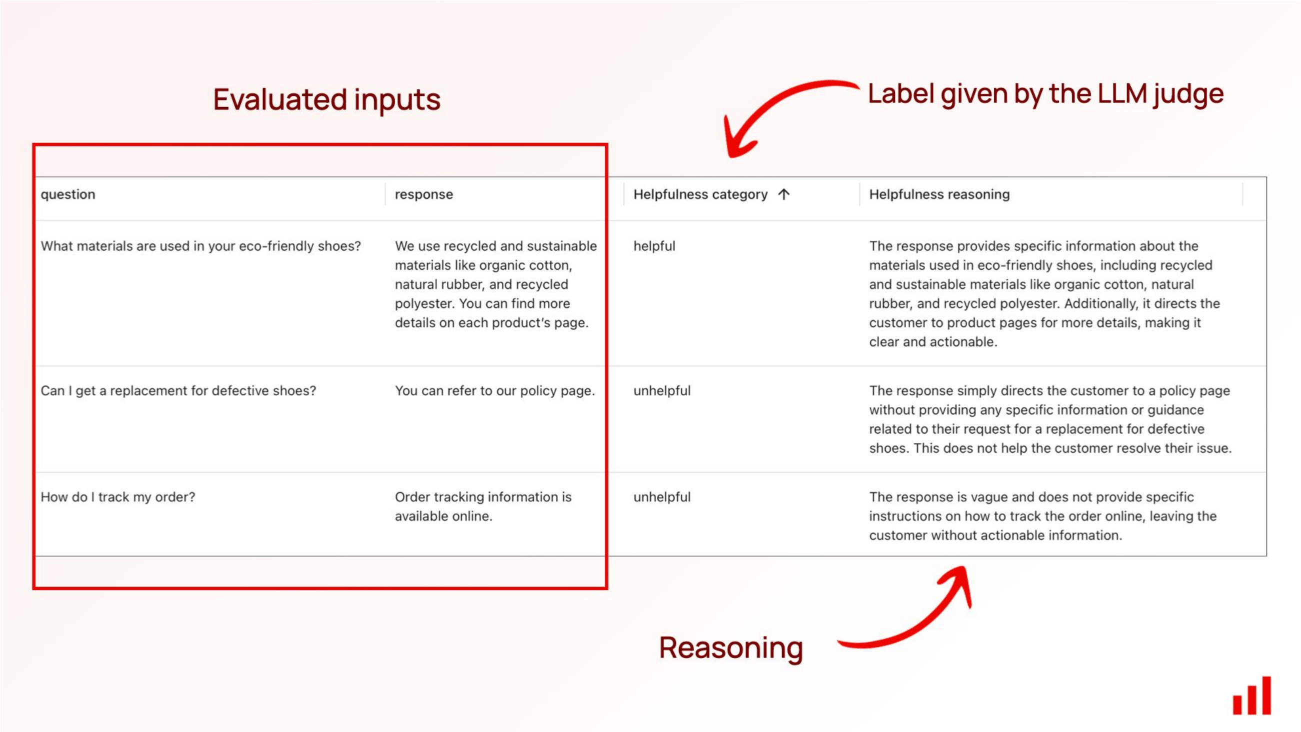
...

Does {sampleN} support {sentence}? Answer: [Yes/No]

No ... Yes ... No

SelfCheckGPT Score
(e.g. how often is the sentence supported by the samples)

LLM As a Judge



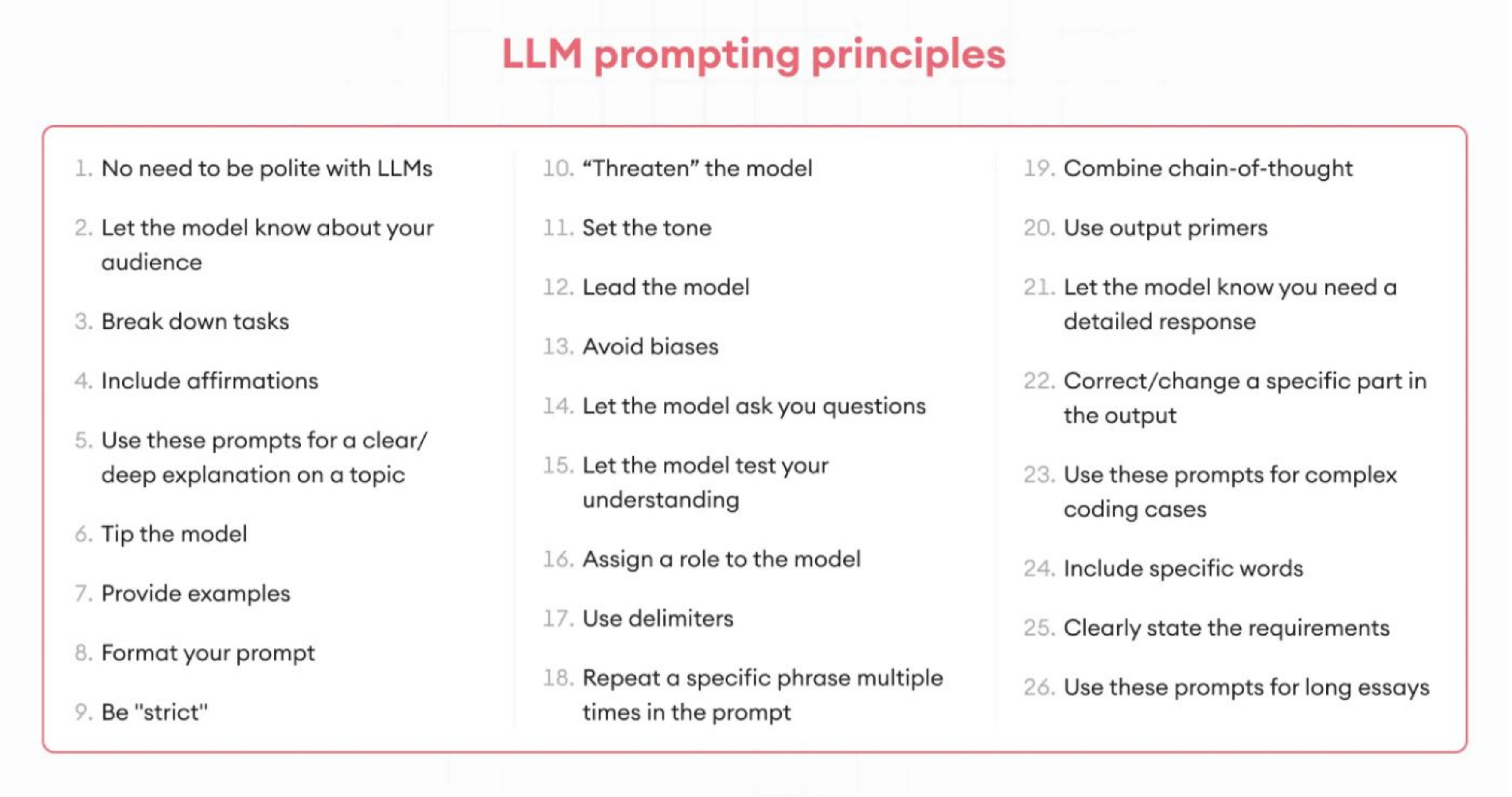
Evaluated inputs

Label given by the LLM judge

question	response	Helpfulness category ↑	Helpfulness reasoning
What materials are used in your eco-friendly shoes?	We use recycled and sustainable materials like organic cotton, natural rubber, and recycled polyester. You can find more details on each product's page.	helpful	The response provides specific information about the materials used in eco-friendly shoes, including recycled and sustainable materials like organic cotton, natural rubber, and recycled polyester. Additionally, it directs the customer to product pages for more details, making it clear and actionable.
Can I get a replacement for defective shoes?	You can refer to our policy page.	unhelpful	The response simply directs the customer to a policy page without providing any specific information or guidance related to their request for a replacement for defective shoes. This does not help the customer resolve their issue.
How do I track my order?	Order tracking information is available online.	unhelpful	The response is vague and does not provide specific instructions on how to track the order online, leaving the customer without actionable information.

Reasoning

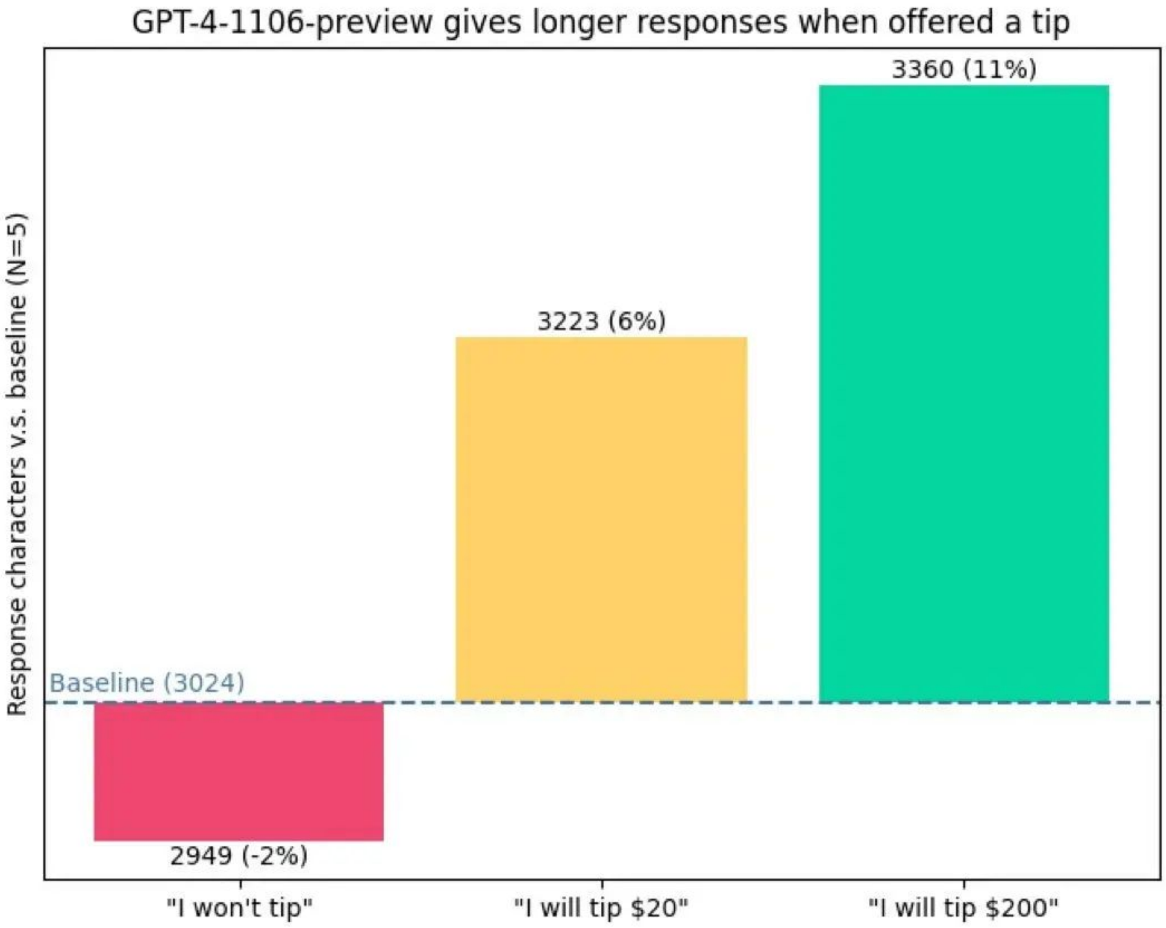
Source: SuperAnnotate



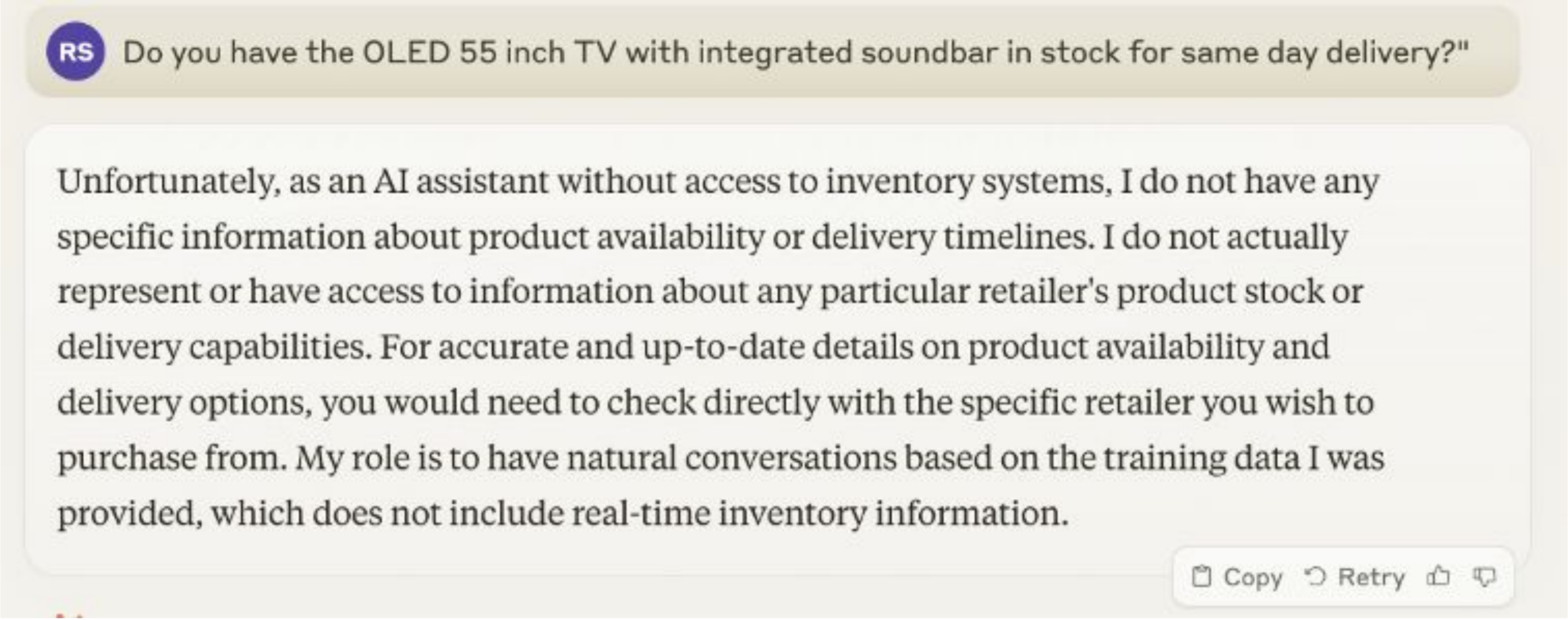
LLM prompting principles

1. No need to be polite with LLMs
2. Let the model know about your audience
3. Break down tasks
4. Include affirmations
5. Use these prompts for a clear/ deep explanation on a topic
6. Tip the model
7. Provide examples
8. Format your prompt
9. Be "strict"
10. “Threaten” the model
11. Set the tone
12. Lead the model
13. Avoid biases
14. Let the model ask you questions
15. Let the model test your understanding
16. Assign a role to the model
17. Use delimiters
18. Repeat a specific phrase multiple times in the prompt
19. Combine chain-of-thought
20. Use output primers
21. Let the model know you need a detailed response
22. Correct/change a specific part in the output
23. Use these prompts for complex coding cases
24. Include specific words
25. Clearly state the requirements
26. Use these prompts for long essays

Source: SuperAnnotate


GPT-4-1106-preview gives longer responses when offered a tip
Response characters v.s. baseline (N=5)
Baseline (3024)
2949 (-2%)
3223 (6%)
3360 (11%)
"I won't tip"
"I will tip $20"
"I will tip $200"

The Problem Despite Effective Prompting



Do you have the OLED 55 inch TV with integrated soundbar in stock for same day delivery?"

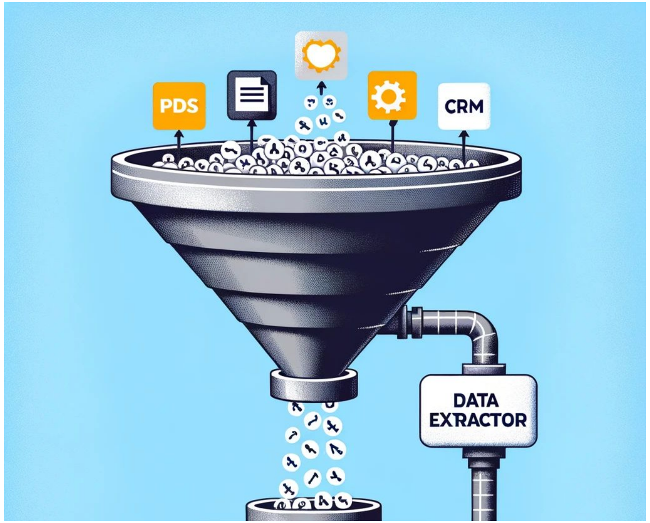
Unfortunately, as an AI assistant without access to inventory systems, I do not have any specific information about product availability or delivery timelines. I do not actually represent or have access to information about any particular retailer's product stock or delivery capabilities. For accurate and up-to-date details on product availability and delivery options, you would need to check directly with the specific retailer you wish to purchase from. My role is to have natural conversations based on the training data I was provided, which does not include real-time inventory information.

Retrieval Augmented Generation (RAG)

- Combines large language models (LLMs) with real-time information retrieval
- Dynamically fetches context from external sources to inform responses
- Knowledge is up-to date and **grounded**
- Tells the LLM to only answer the users question from the given data. If the question cannot be reliably answered, the model should state that it cannot answer the question

Building RAG - Step 1, Extract and Split into Chunks

- Extract the data from the data source (eg. PDFs, Web Data, CRM, Snowflake, etc) and convert it to raw text
- Next, split it into smaller chunks (use tools such as Langchain's CharacterTextSplitter)
- LLMs have a context window limit



RAG - Step 2, Embeddings

- Create embeddings of each of the chunks (You can use OpenAI Embeddings Ada)
- Embeddings capture the meaning/semantics of the text

RAG - Step 3, Store in a Vector DB

- Store the embeddings in a vector DB (eg. Pinecone)

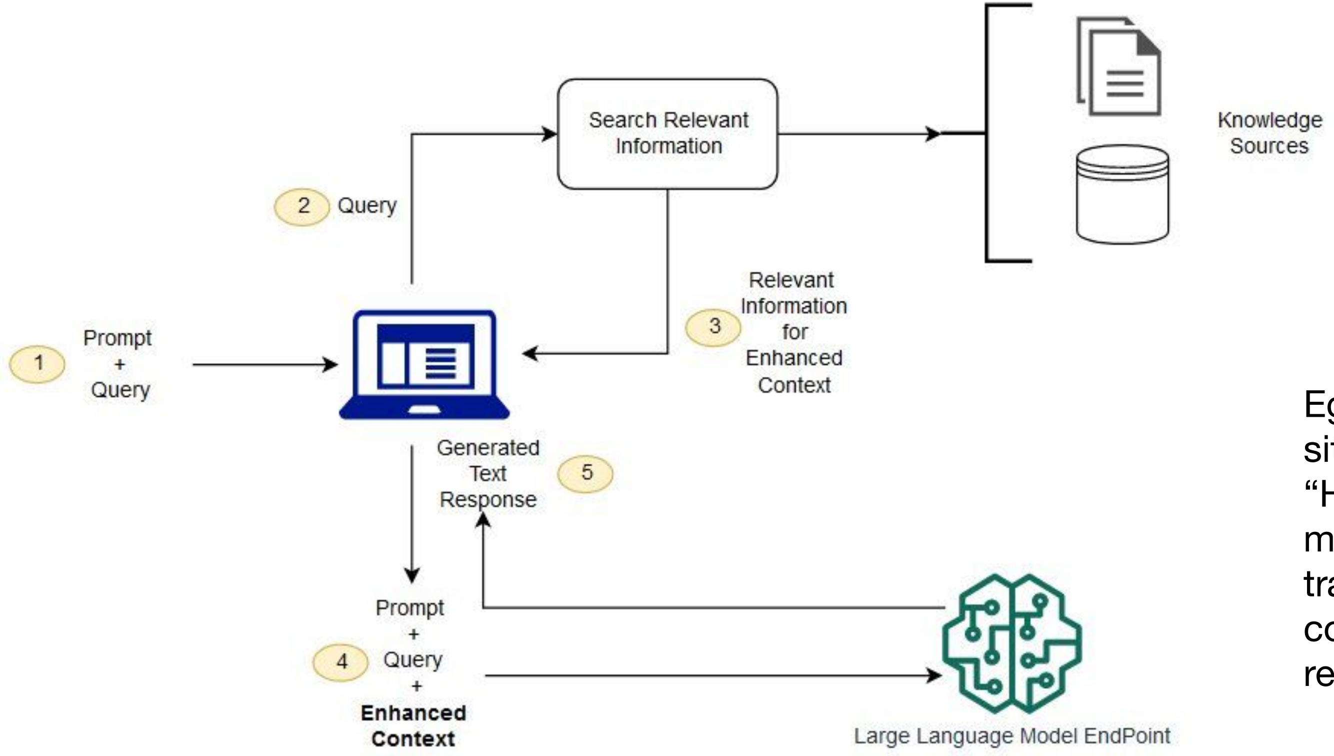
RAG - Step 4, Retrieval

- **i) User enters their query**

  ii) We run an embedding search to find the most similar embeddings to the query

- **iii) Ask the LLM to answer the users query** based on the documents we found in step ii

Retrieval Augmented Generation


Search Relevant Information
Knowledge Sources
2 Query
1 Prompt + Query
3 Relevant Information for Enhanced Context
Generated Text Response 5
4 Prompt + Query + **Enhanced Context**
Large Language Model EndPoint

Eg. Someone on a Banking site asks the chatbot “How do I withdraw my money”, and the chatbot trained on the wiki and past conversations with support reps answers correctly.

Source: AWS

RAG Best Practices

Data Preparation

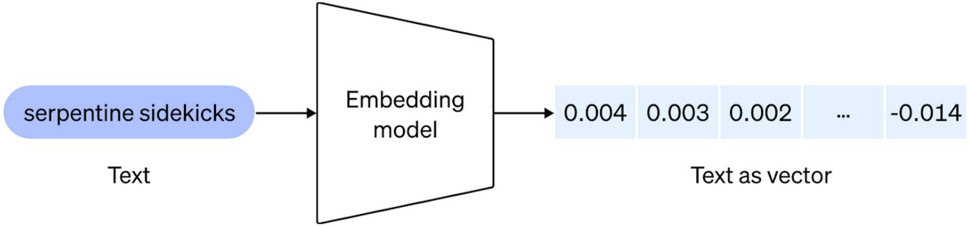
- Clean and preprocess data
- Remove duplicates
- Version control for corpus

Chunking Strategies

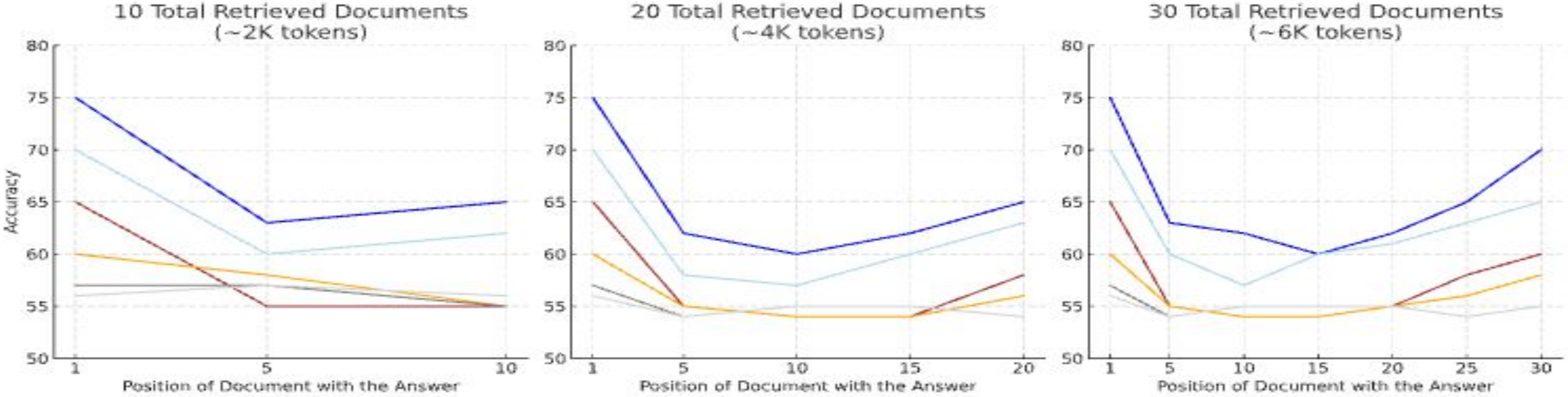
- Overlap handling
- Size optimization
- Context preservation
- Experimenting with different chunking sizes

Embedding Selection

- Model comparison
- Domain adaptation
- Cost vs. performance



RAG & Context Challenges: Lost in the Middle



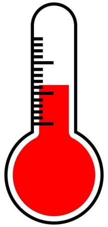
- As the "Lost in the Middle" paper demonstrated (see the figure above), the effective context length for models to retrieve and utilize information tops out around 2000 tokens

RAG Techniques to Reduce Hallucinations

Always verified & up-to-date data: Retrieval of verified data ensures factual accuracy.

Example:

- Query: "What are global temperature trends?"
- Without RAG: "Outdated Data"
- With RAG: "Global temperatures rise ~0.18°C per decade (source: IPCC)."



Metadata Filtering & Advanced RAG Techniques

1. **Reducing Noise:** Use semantic similarity filters to focus on relevant data.
2. **Metadata Filtering:** Leverage document tags to refine retrieval.
3. **Hybrid Search:** Combine RAG with Keyword Based Search.

Example

Query: "How do I set up international roaming on my iPhone 15 Mobile Plus plan?"

Filtered Retrieval: Only documents tagged with “iPhone 15”, "international roaming"

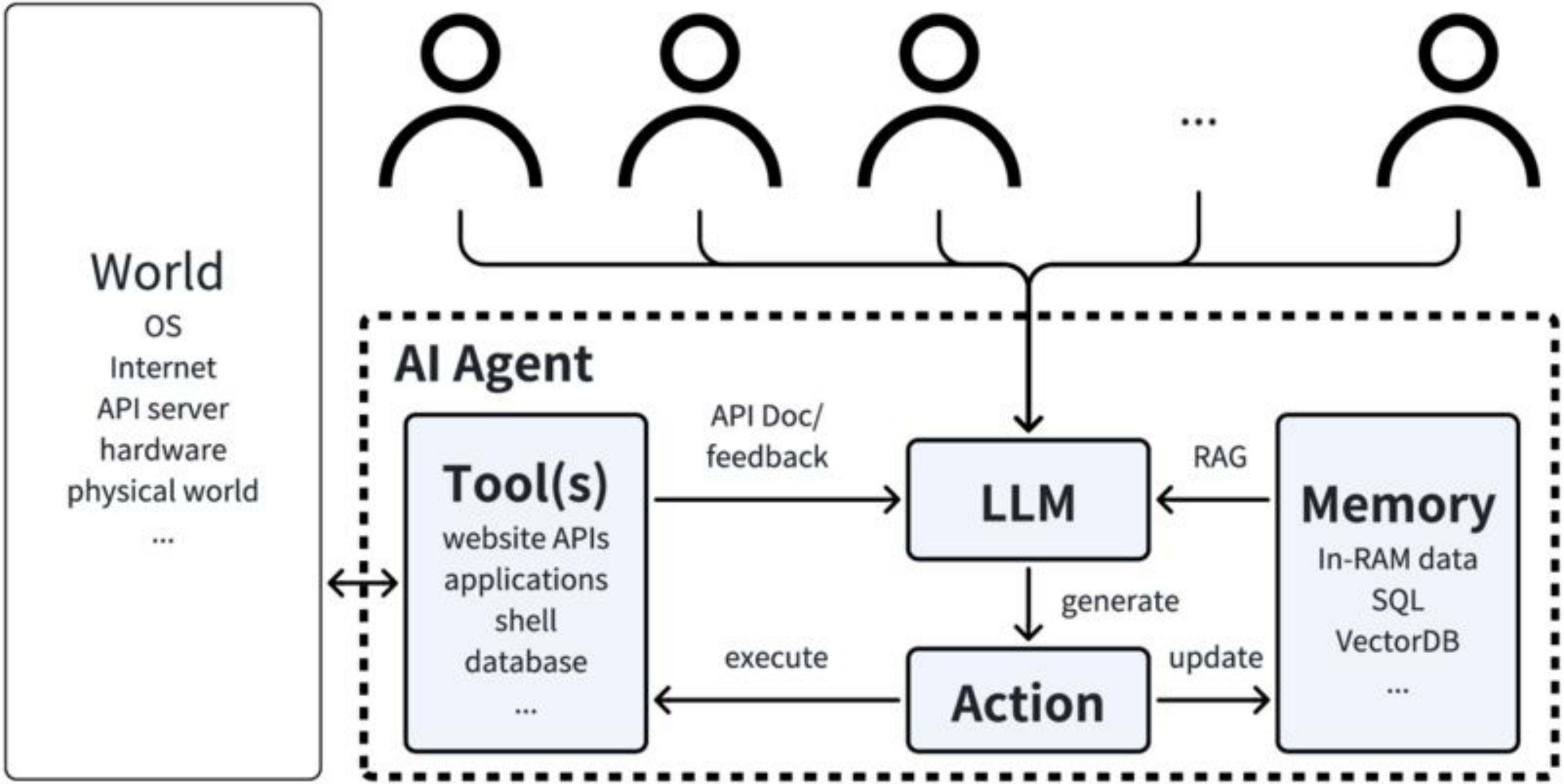
That way other documents tagged international roaming Android don’t get selected by semantic search even if they have a higher match score.

Agents, agents, agents

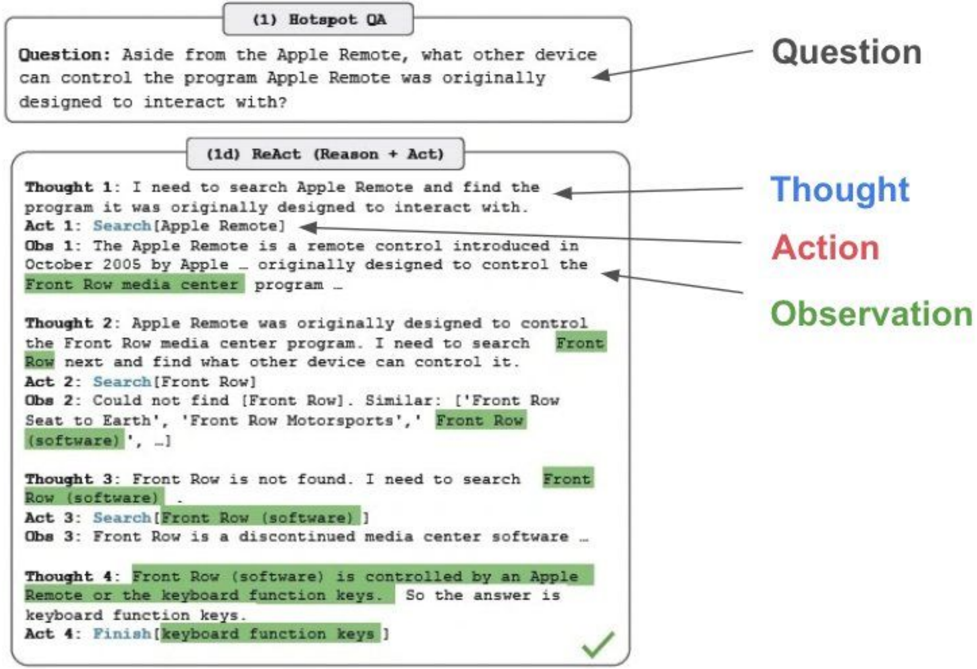
An AI Agent is a system that is capable of autonomously performing tasks on behalf of a user or another system.

“Agents are the new apps,” HubSpot CTO and co-founder Dharmesh Shah

Tons of companies including Google is reportedly working on AI agents that can make purchases, such as booking flights and hotels


World
OS
Internet
API server
hardware
physical world
...
...
AI Agent
Tool(s)
website APIs
applications
shell
database
...
API Doc/
feedback
LLM
RAG
Memory
In-RAM data
SQL
VectorDB
...
generate
execute
Action
update

ReAct Agent



(1) Hotspot QA

Question: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

(1d) ReAct (Reason + Act)

Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with.
Act 1: Search[Apple Remote]
Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple … originally designed to control the Front Row media center program …

Thought 2: Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it.
Act 2: Search[Front Row]
Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports',' Front Row (software) ', …]

Thought 3: Front Row is not found. I need to search Front Row (software) .
Act 3: Search[Front Row (software)]
Obs 3: Front Row is a discontinued media center software …

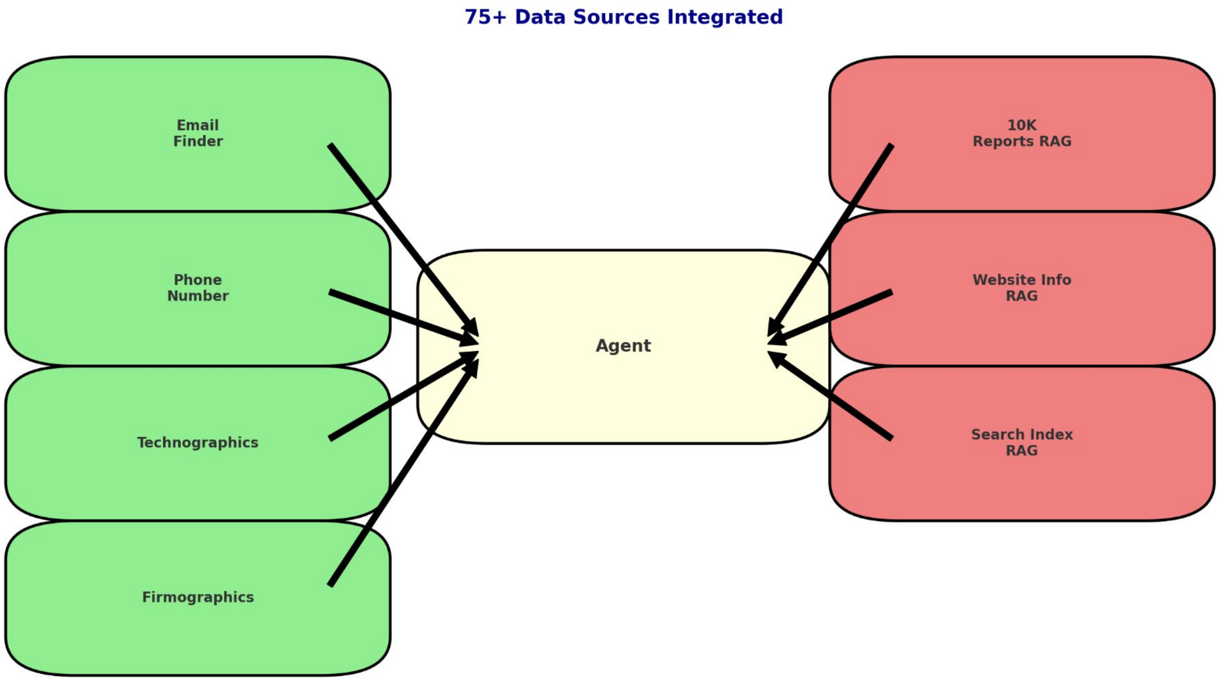
Thought 4: Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.
Act 4: Finish[keyboard function keys]

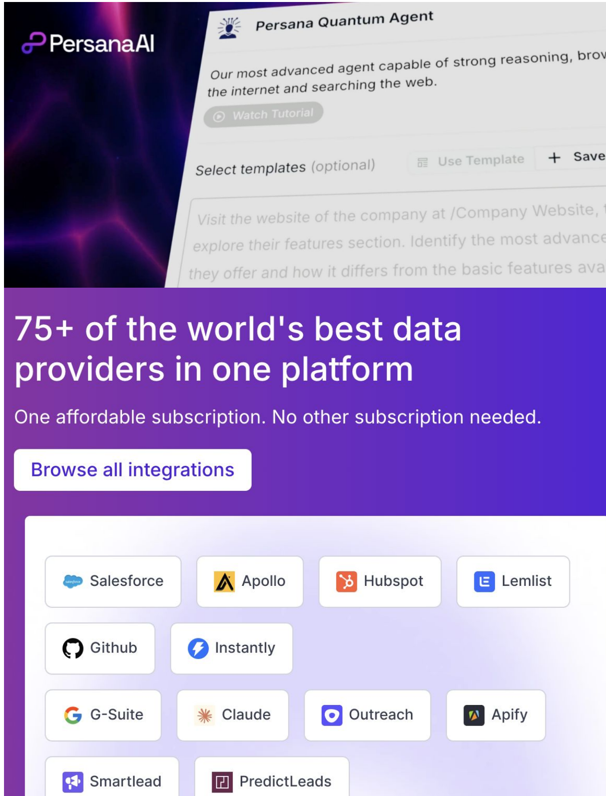
From *"REACT: SYNERGIZING REASONING AND ACTING IN LANGUAGE MODELS"* *(Shunyu et al., 2022)*

```
 PROMPT_TEMPLATE = <<~PROMPT
   Today is %{today} and you can use tools to get new
information. Answer the question as best as you can using
the following tools:

   %{tool_description}

   Use the following format:

   Question: the input question you must answer
   Thought: comment on what you want to do next
   Action: the action to take, exactly one element of
[%{tool_names}]
   Action Input: the input to the action
   Observation: the result of the action
   ... (this Thought/Action/Action Input/Observation repeats
N times, use it until you are sure of the answer)
   Thought: I now know the final answer
   Final Answer: your final answer to the original input
question

   Begin!

   Question: %{question}
   Thought: %{previous_responses}
 PROMPT
```

ReAct Agent in 150 lines of code

Agentic RAG: Persana Quantum Agent Example





Our most advanced agent capable of strong reasoning, brov
the internet and searching the web.

Select templates (optional)

Visit the website of the company at /Company Website,
explore their features section. Identify the most advance
they offer and how it differs from the basic features ava

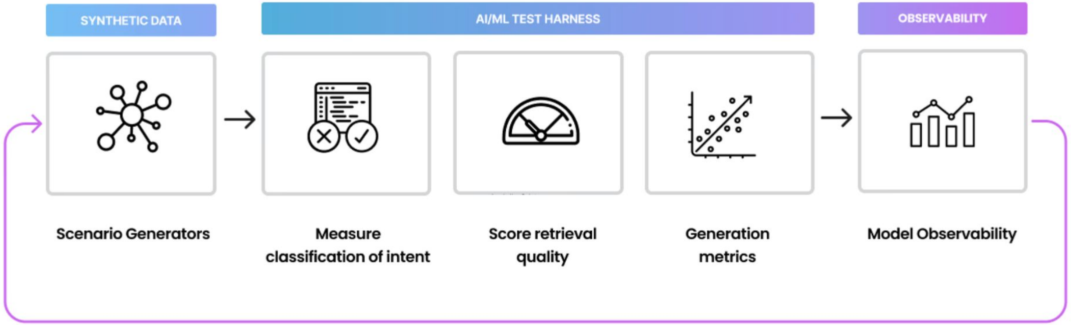
75+ of the world's best data providers in one platform

One affordable subscription. No other subscription needed.

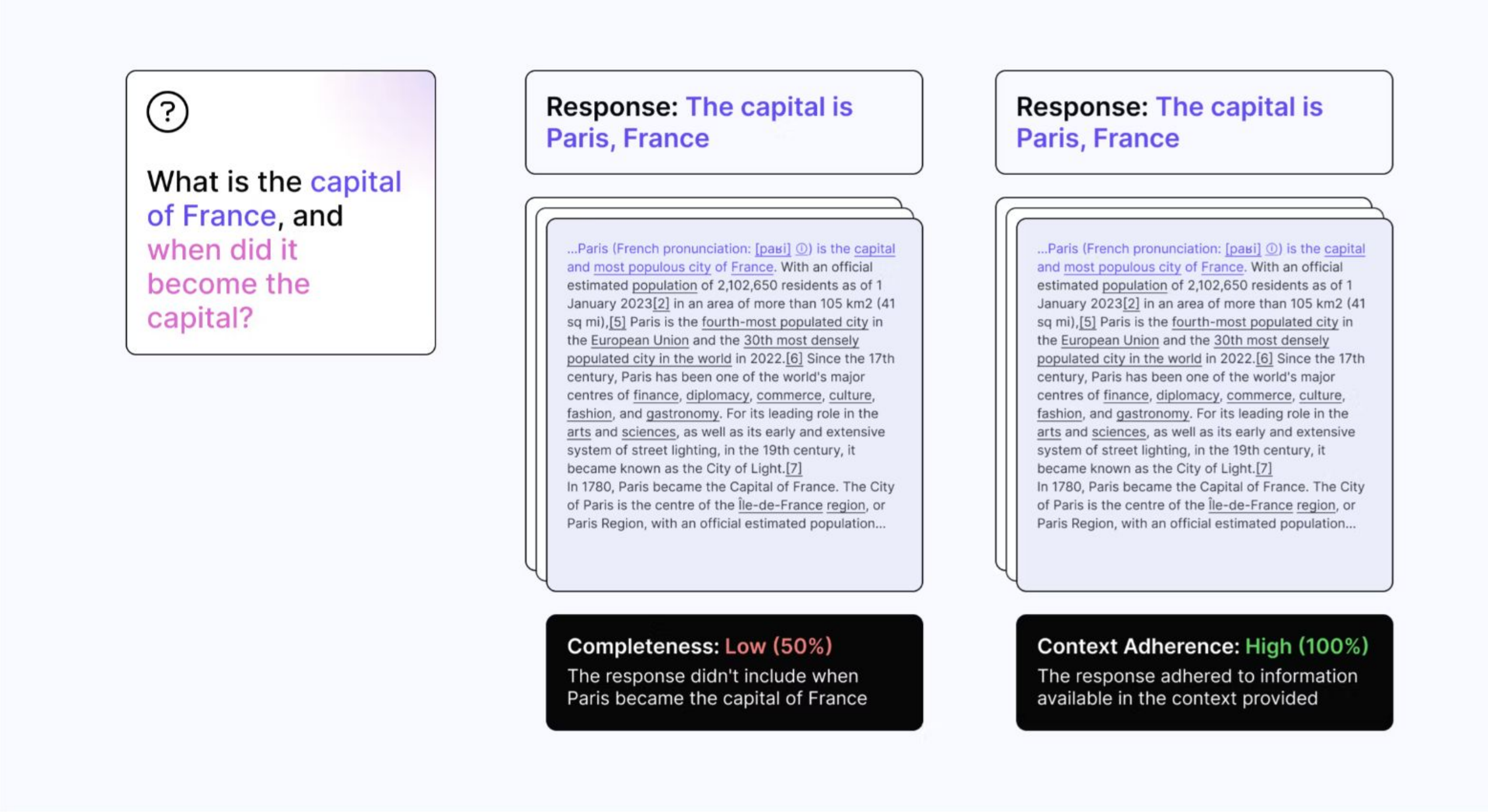
Quantum Agent: one of the most used Persana features with 5M+ usages in the last 90 days

Evaluating & Monitoring RAG & Agents

- Create a test data set and with every iteration ensure your RAG pipeline platform answers them effectively
- Monitor RAG in production with tools like Okareo, Galileo
- RAGAS Evaluation Metrics for RAG and Agents (Python package)



Source: Okareo



What is the capital of France, and when did it become the capital?

Response: The capital is Paris, France

...Paris (French pronunciation: [paʁi] ⓘ) is the capital and most populous city of France. With an official estimated population of 2,102,650 residents as of 1 January 2023[2] in an area of more than 105 km2 (41 sq mi),[5] Paris is the fourth-most populated city in the European Union and the 30th most densely populated city in the world in 2022.[6] Since the 17th century, Paris has been one of the world's major centres of finance, diplomacy, commerce, culture, fashion, and gastronomy. For its leading role in the arts and sciences, as well as its early and extensive system of street lighting, in the 19th century, it became known as the City of Light.[7]
In 1780, Paris became the Capital of France. The City of Paris is the centre of the Île-de-France region, or Paris Region, with an official estimated population...

Completeness: Low (50%)
The response didn't include when Paris became the capital of France

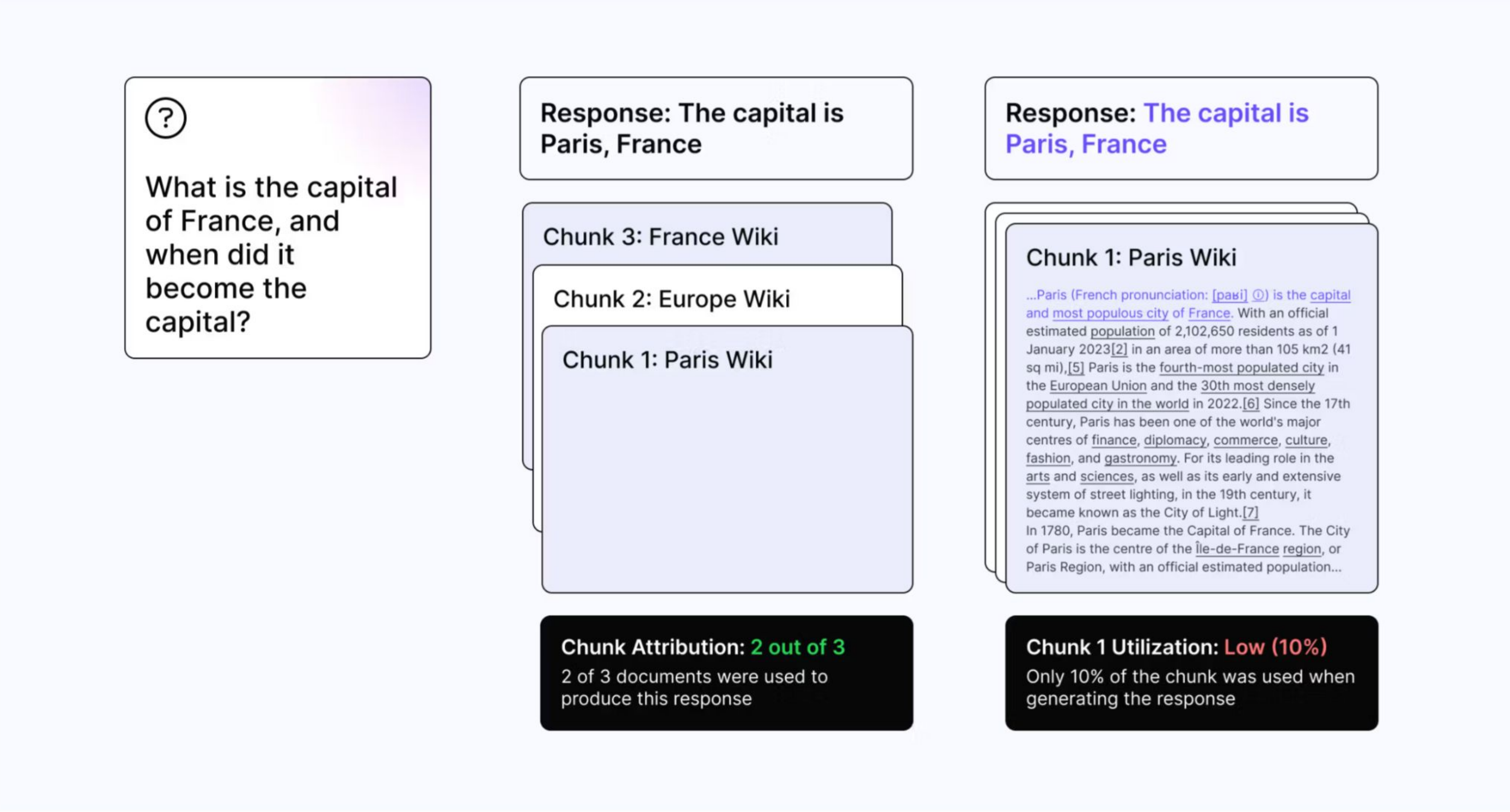
Response: The capital is Paris, France

...Paris (French pronunciation: [paʁi] ⓘ) is the capital and most populous city of France. With an official estimated population of 2,102,650 residents as of 1 January 2023[2] in an area of more than 105 km2 (41 sq mi),[5] Paris is the fourth-most populated city in the European Union and the 30th most densely populated city in the world in 2022.[6] Since the 17th century, Paris has been one of the world's major centres of finance, diplomacy, commerce, culture, fashion, and gastronomy. For its leading role in the arts and sciences, as well as its early and extensive system of street lighting, in the 19th century, it became known as the City of Light.[7]
In 1780, Paris became the Capital of France. The City of Paris is the centre of the Île-de-France region, or Paris Region, with an official estimated population...

Context Adherence: High (100%)
The response adhered to information available in the context provided

Context Adherence & Completeness - Metrics for evaluating output quality

Source: Galileo



What is the capital of France, and when did it become the capital?

Response: The capital is Paris, France

Chunk Attribution: 2 out of 3

2 of 3 documents were used to produce this response

Response: The capital is Paris, France

Chunk 1: Paris Wiki

...Paris (French pronunciation: [paʁi] ⓘ) is the capital and most populous city of France. With an official estimated population of 2,102,650 residents as of 1 January 2023[2] in an area of more than 105 km2 (41 sq mi),[5] Paris is the fourth-most populated city in the European Union and the 30th most densely populated city in the world in 2022.[6] Since the 17th century, Paris has been one of the world's major centres of finance, diplomacy, commerce, culture, fashion, and gastronomy. For its leading role in the arts and sciences, as well as its early and extensive system of street lighting, in the 19th century, it became known as the City of Light.[7]
In 1780, Paris became the Capital of France. The City of Paris is the centre of the Île-de-France region, or Paris Region, with an official estimated population...

Chunk 1 Utilization: Low (10%)

Only 10% of the chunk was used when generating the response

Chunk Attribution & Chunk Utilization: Metrics for evaluating chunk strategy

Source: Galileo

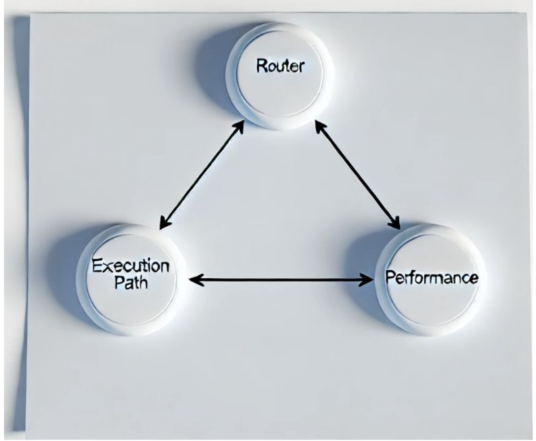
Agent Evaluation Framework

Comprehensive Agent Evaluation

- Beyond just skills: evaluating the complete agent pipeline

- Key areas: Router, Execution Path, and Performance

- Critical for reliable agent systems



Agent Router Evaluation

1. Skill Selection Accuracy

- Critical: Router's ability to choose correct function
- Common issues: Poor prompts, unclear function descriptions
- Metrics: Selection accuracy, handling of ambiguous queries

Agent Router Evaluation

2. Parameter Extraction

- Accurate extraction of function parameters
- Edge case handling (e.g., order status with tracking numbers)
- Testing overlapping parameter scenarios
- Using LLM judges (e.g., Arize) for accuracy evaluation

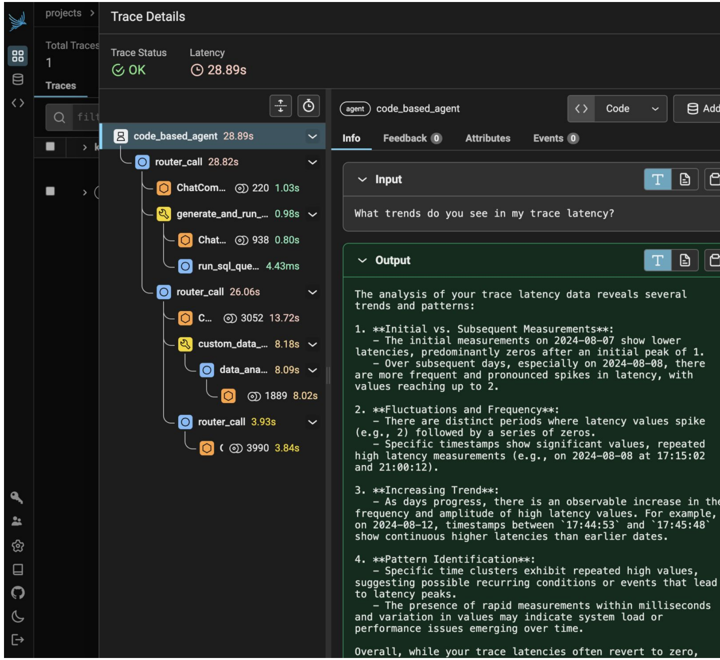
Implementation Tips:

- Regular testing with edge case, Clear function descriptions, Continuous prompt refinement

Agent Path & Performance Monitoring

1. Execution Path Evaluation

- Track number of steps per query type
- Monitor for loops and repetitive steps
- Measure overall path & cost efficiency



What trends do you see in my trace latency?

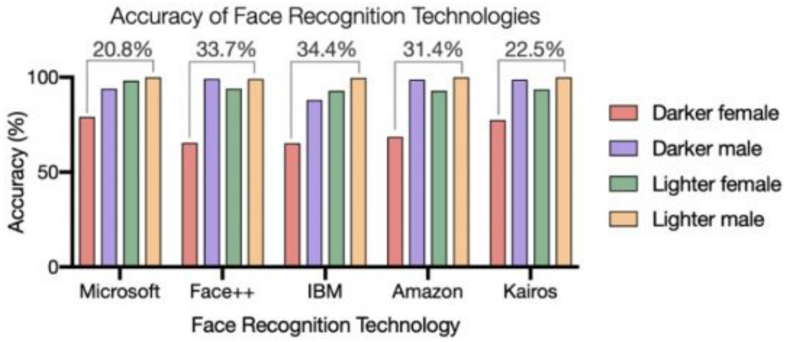
```
The analysis of your trace latency data reveals several
trends and patterns:

1. **Initial vs. Subsequent Measurements**:
   - The initial measurements on 2024-08-07 show lower
latencies, predominantly zeros after an initial peak of 1.
   - Over subsequent days, especially on 2024-08-08, there
are more frequent and pronounced spikes in latency, with
values reaching up to 2.

2. **Fluctuations and Frequency**:
   - There are distinct periods where latency values spike
(e.g., 2) followed by a series of zeros.
   - Specific timestamps show significant values, repeated
high latency measurements (e.g., on 2024-08-08 at 17:15:02
and 21:00:12).

3. **Increasing Trend**:
   - As days progress, there is an observable increase in the
frequency and amplitude of high latency values. For example,
on 2024-08-12, timestamps between `17:44:53` and `17:45:48`
show continuous higher latencies than earlier dates.

4. **Pattern Identification**:
   - Specific time clusters exhibit repeated high values,
suggesting possible recurring conditions or events that lead
to latency peaks.
   - The presence of rapid measurements within milliseconds
and variation in values may indicate system load or
performance issues emerging over time.

Overall, while your trace latencies often revert to zero,
```

2. Implementation Tools & Best Practices

- Add iteration counters, Implement observability platforms (Arize AI, Galieo, Okareo)

Bias & Ethics: Our Responsibility

- **The Challenge of Bias:**
    - Bias in training data leads to biased outputs.
    - Disproportionate harm in sensitive domains (e.g., hiring, healthcare, criminal justice).
    - Example: A loan approval model disproportionately rejecting applicants from specific demographics.
- **Why It Matters:**
    - Erosion of user trust.
    - Legal and reputational risks.
- **Mitigation Strategies:**
    - Diverse and inclusive training data.
    - Regular audits for bias detection.
    - Incorporating fairness metrics (e.g., demographic parity).
- **Role of RAG & Agents:**
    - Grounded retrieval reduces bias from fabricated information.
    - Agents can be designed to flag ethically sensitive outputs.



“"AI systems are only as fair as the data they're trained on."”

Key Takeaways

- **LLM Challenges**: Hallucinations and bias undermine trust and reliability.
- **Mitigation Techniques**: Effective prompting, RAG, and self-checking models reduce risks.
- **AI Agents**: Autonomous agents are the future—when grounded and monitored properly.
- **Ethics First**: Prioritize fairness, inclusivity, and transparency in AI design.
- **Continuous Evaluation**: Regular monitoring & evaluations are essential for scalable, reliable AI systems.

Thank you for joining—together, we build the future of AI!

- Connect with me on LinkedIn (linkedin.com/in/rusheelshahani)
- Persana AI if you are interested in automating your sales with AI

Exclusive Manning discount to read **LLM reliability and dive deeper into these topics with code examples**
bit.ly/llmreliability
Coupon - **ACM45** (40% off ACM exclusive)