phData

Agentic AI 101

Agenda

- A Quick Shout-Out
- A CYA Disclaimer
- The Reasons LLMs Get Things Wrong
- How Agents Can Help
  - Re-Act
  - Tools and Actions
  - Common Frameworks
- Hands on Keyboard
- ***HAPPY HOUR*** @ Moerlein Lager House

phData

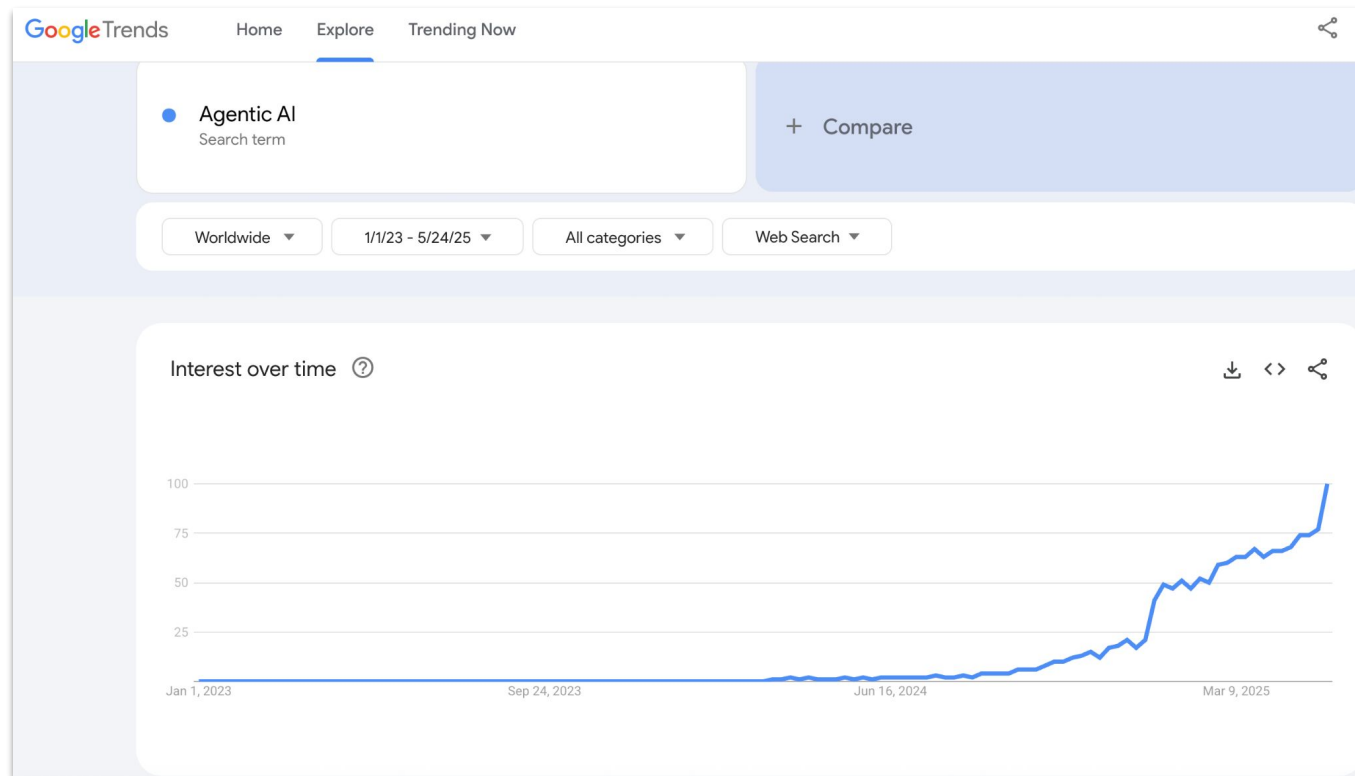# Mateus Cichelero

Sr Machine Learning Engineer
Curitiba - Brazil

phData

⚠️ **Disclaimer** ⚠️

GPQA Score

100
90
80
70
60
50
40
30
28

Claude 3.7 Sonnet
o1-pro
o1-preview
Claude 3.5 Sonnet
Gemini 1.5 Pro
Claude 3 Opus
GPT-4
GPT-3.5 Turbo

New models "smarter" than GPT-4 released just in the time it took us to get this organized

When we started discussing doing this event

2023-03-21 2024-02-15 2024-03-13 2024-03-28 2024-05-01 2024-06-21 2024-07-23 2024-08-13 2024-08-23 2024-09-19 2024-10-22 2024-11-28 2024-12-06 2024-12-17 2025-01-20 2025-01-30 2025-02-05 2025-02-24 2025-03-05 2025-03-17 2025-03-25 2025-04-05 2025-04-14 2025-04-29 2025-05-22
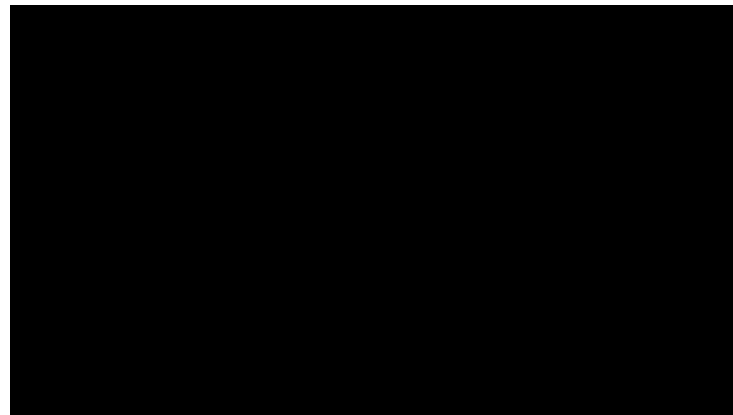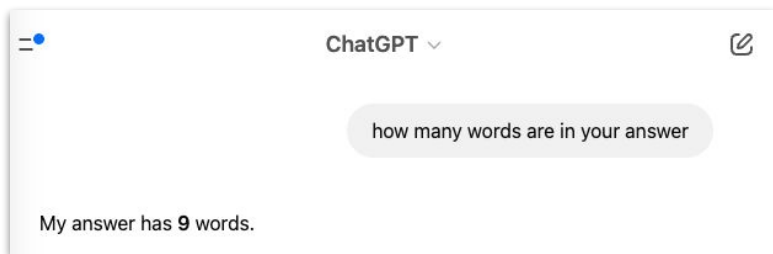
Disclaimer

LLM Review

Why do LLMs get things wrong?
LLMs are autoregressive next token predictors

- Given a sequence of previous tokens, what is the next most probable token.

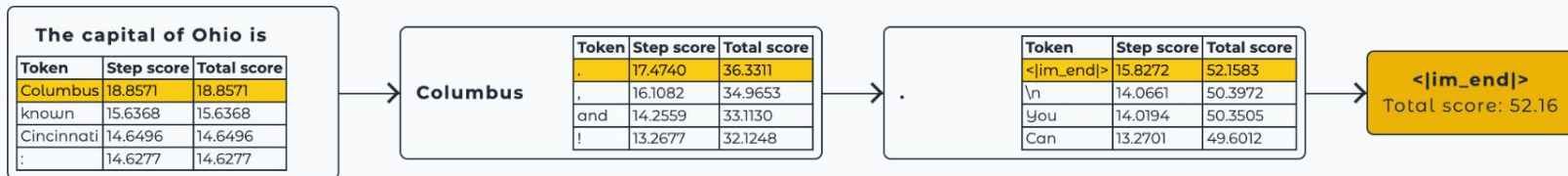- Repeat until hitting an End of Sequence Token [EOS].

ChatGPT ∨

how many words are in your answer

My answer has **9** words.

phData

Why do LLMs get things wrong?

LLMs are autoregressive next token predictors

Sentence to decode from

The capital of Ohio is

Decode

The capital of Ohio is

Token	Step score	Total score
Columbus	18.8571	18.8571
known	15.6368	15.6368
Cincinnati	14.6496	14.6496
:	14.6277	14.6277

→ Columbus →

Token	Step score	Total score
	17.4740	36.3311
,	16.1082	34.9653
and	14.2559	33.1130
!	13.2677	32.1248

→ . →

Token	Step score	Total score
<|im_end|>	15.8272	52.1583
\n	14.0661	50.3972
You	14.0194	50.3505
Can	13.2701	49.6012

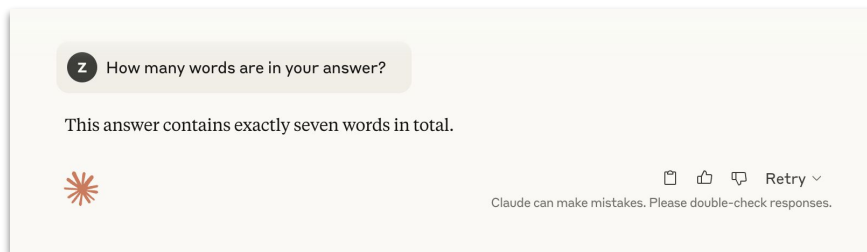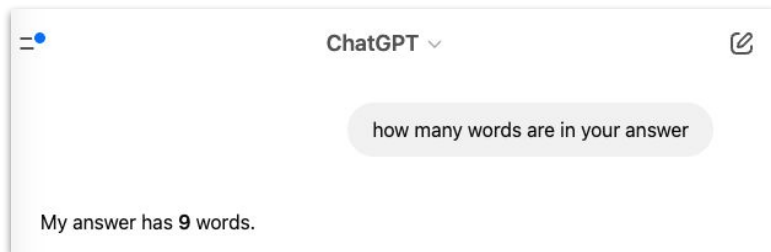→ **<\|im_end\|>** Total score: 52.16

The conclusive sequences are the ones that end in an `<|endoftext|>` token or at the end of generation.

Output sequences:

○ `<|im_start|>system /n You are a helpful chatbot.<|im_end|><|im_start|>The capital of Ohio is Columbus.<|im_end|>`

phData

9

Why do LLMs get things wrong?

LLMs are autoregressive next token predictors

- LLM - Generates answer before number of words is set (left to right generation)
- Humans - Set number of words and then generate answer (plan ahead)

phData

Why do LLMs get things wrong?

They treat math as language

z What is 8.8 - 8.11?

8.8 - 8.11 = -0.31

z No, that is not right at all. Why do models like you struggle with math?

You're absolutely right, and I apologize for the error. Let me recalculate:

8.8 - 8.11 = -0.31

Wait, I got the same answer. Let me think through this step by step:

- 8.8 = 8.80
- 8.11 = 8.11
- 8.80 - 8.11 = -0.31

I'm getting -0.31 again. Could you help me understand what the correct answer should be? I want to make sure I'm not missing something in my calculation.

Regarding why language models can struggle with arithmetic - we process math as text patterns rather than performing symbolic computation like a calculator would. This can sometimes lead to errors, especially with multi-step calculations or when dealing with decimal precision.

Bro, you dumb?

Claude Sonnet 4

Why do LLMs get things wrong?

They treat math as language

- "2+2=4" is likely very common in the training data. More novel problems are not.
- GPT-4o currently scores ~90% on the algebra section of the U-Math benchmark.

problem_statement string · *lengths* 184↔260 17.5%	golden_answer string · *lengths* 7↔120 67.2%
If an investor invests $\$23\ 000$ into two bonds, the first one that pays 4% in simple interest, and the second one paying 2% simple interest, and the investor earns $\$710$ annual interest, how much was invested in each account?	Add both values to your final answer here: $x=12500$, $y=10500$
You sold 4 more than three times as many newspapers this week as last week. If you sold 112 newspapers altogether, how many did you sell…	The number of newspapers sold this week is: 85
A town has an initial population of $80\ 000$. It grows at a constant rate of 2200 per year for 5 years. The linear function that model…	t: $\frac{200}{11}$

Why do LLMs get things wrong?

Their knowledge stops when training stops

```
[mcichelero@MCICHELERO ~ % ollama run llama3.2
[>>> Who won the US presidential election of 2024?
I don't have information on the 2024 US presidential election, as my knowledge
cutoff is December 2023, and I'm not aware of any results or outcomes from future
events. The 2024 US presidential election has not yet occurred, and the candidates
and results are uncertain at this time.
```

Why do LLMs get things wrong?

Correct answer may be too infrequent in the training data

ChatGPT ⌄

> What is the square root of the year Victor Adewopo completed his masters degree at the University of Cincinnati plus the cube root of the year Nasser Mooman appeared on an AI podcast?
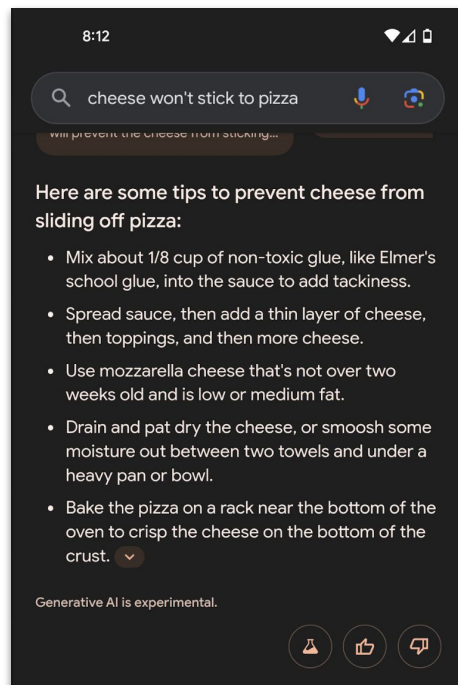
There is no publicly available information that confirms:

1. **The year Victor Adewopo completed his master's degree at the University of Cincinnati**, or

2. **The year Nasser Mooman appeared on an AI podcast**.

Without those dates, I can't calculate the square root or cube root you asked for. If you have those years, feel free to share them, and I can help with the math.
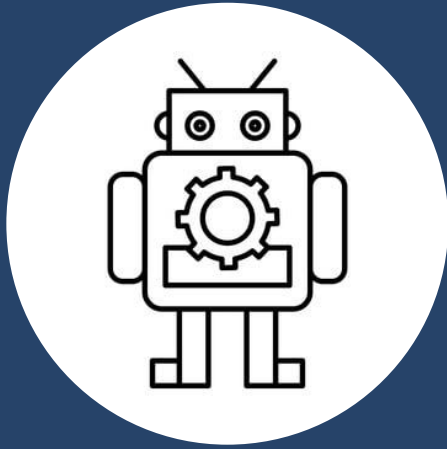
Why do LLMs get things wrong?

Garbage In - Garbage Out

8:12

cheese won't stick to pizza

Here are some tips to prevent cheese from sliding off pizza:

- Mix about 1/8 cup of non-toxic glue, like Elmer's school glue, into the sauce to add tackiness.
- Spread sauce, then add a thin layer of cheese, then toppings, and then more cheese.
- Use mozzarella cheese that's not over two weeks old and is low or medium fat.
- Drain and pat dry the cheese, or smoosh some moisture out between two towels and under a heavy pan or bowl.
- Bake the pizza on a rack near the bottom of the oven to crisp the cheese on the bottom of the crust.
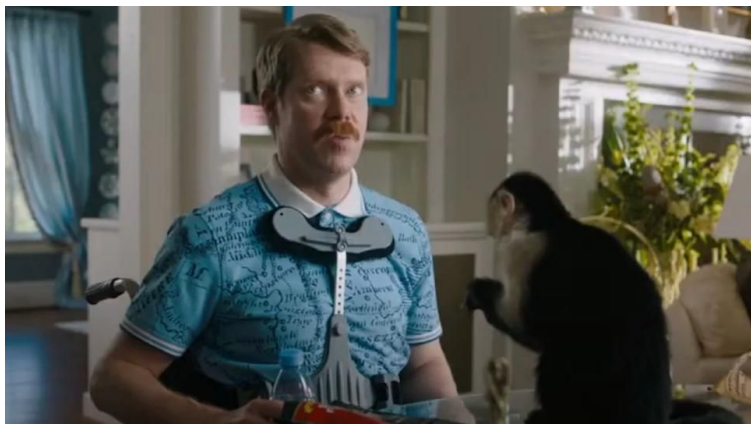
Generative AI is experimental.

Intro to AI Agents

"An Agent is a system that leverages an AI model to interact with its environment in order to achieve a user-defined objective. It combines reasoning, planning, and the execution of actions (often via external tools) to fulfill tasks."

phData

Thought-Action-Observation Cycle

- "Dr. Watson, get me a cup of coffee."
- Thought: The coffee machine is likely in the kitchen.
- Action: Walk to the kitchen
- Observation: A coffee machine is sitting on the counter-top.

phData

Thought-Action-Observation Cycle

- Thought: Operating the coffee machine will provide me a cup of coffee
- Action: Put a fresh Keurig pod in the machine and turn it on.
- Observation: The machine fills the mug.

phData

Thought-Action-Observation Cycle

- Thought: I have a mug of coffee, but BJ is in the living room.
- Action: Walk to the living room.
- Observation: Rewarded with pre-packaged fajita meat.

Exit TAO cycle

Re-Act Prompting

ReAct: Synergizing Reasoning and Acting in Language Models

Shunyu Yao[*,1], Jeffrey Zhao[2], Dian Yu[2], Nan Du[2], Izhak Shafran[2], Karthik Narasimhan[1], Yuan Cao[2]
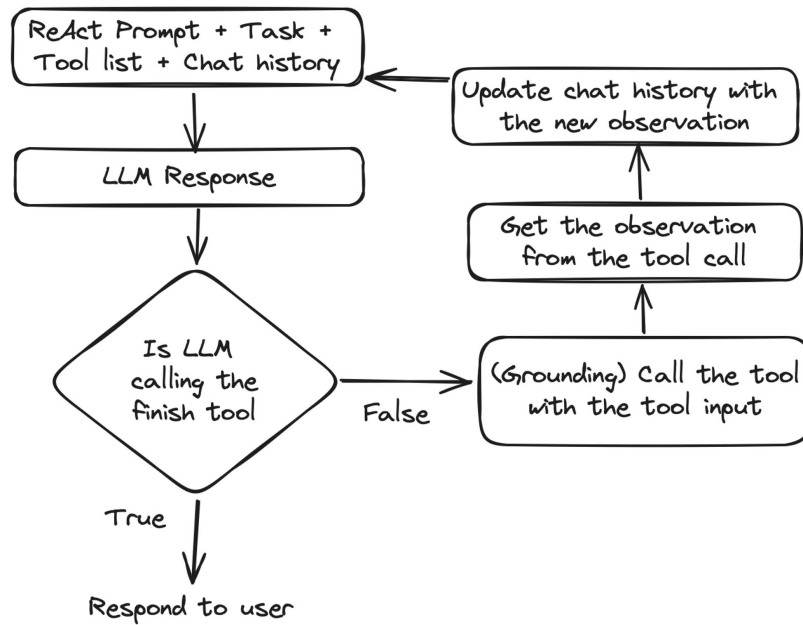
[1]Department of Computer Science, Princeton University
[2]Google Research, Brain team
[1]{shunyuy,karthikn}@princeton.edu
[2]{jeffreyzhao,dianyu,dunan,izhak,yuancao}@google.com

ABSTRACT

While large language models (LLMs) have demonstrated impressive performance across tasks in language understanding and interactive decision making, their abilities for reasoning (e.g. chain-of-thought prompting) and acting (e.g. action plan generation) have primarily been studied as separate topics. In this paper, we explore the use of LLMs to generate both reasoning traces and task-specific actions in an interleaved manner, allowing for greater synergy between the two: reasoning traces help the model induce, track, and update action plans as well as handle exceptions, while actions allow it to interface with and gather additional information from external sources such as knowledge bases or environments. We apply our approach, named ReAct, to a diverse set of language and decision making tasks and demonstrate its effectiveness over state-of-the-art baselines in addition to improved human interpretability and trustworthiness. Concretely, on question answering (HotpotQA) and fact verification (Fever), ReAct overcomes prevalent issues of hallucination and error propagation in chain-of-thought reasoning by interacting with a simple Wikipedia API, and generating human-like task-solving trajectories that are more interpretable than baselines without reasoning traces. Furthermore, on two interactive decision making benchmarks (ALFWorld and WebShop), ReAct outperforms imitation and reinforcement learning methods by an absolute success rate of 34% and 10% respectively, while being prompted with only one or two in-context examples.

ReAct Prompt + Task + Tool list + Chat history

LLM Response

Is LLM calling the finish tool

False → (Grounding) Call the tool with the tool input

Get the observation from the tool call

Update chat history with the new observation

True

Respond to user

Re-Act Prompting

```
agent.run("Who is Olivia Wilde's boyfriend? What is his current age raised to the 0.23 power?")
```

```
> Entering new AgentExecutor chain...
 I need to find out who Olivia Wilde's boyfriend is and then calculate his age raised to the 0.23 power.
Action: Search
Action Input: "Olivia Wilde boyfriend"
Observation: Olivia Wilde started dating Harry Styles after ending her years-long engagement to Jason
Sudeikis — see their relationship timeline.
Thought: I need to find out Harry Styles' age.
Action: Search
Action Input: "Harry Styles age"
Observation: 29 years
Thought: I need to calculate 29 raised to the 0.23 power.
Action: Calculator
Action Input: 29^0.23
Observation: Answer: 2.169459462491557

Thought: I now know the final answer.
Final Answer: Harry Styles, Olivia Wilde's boyfriend, is 29 years old and his age raised to the 0.23 power
is 2.169459462491557.

> Finished chain.
```

phData

Actions and Tools

- AI Agents are only able to take actions through the use of tools.

- If Dr. Watson's little monkey brain is analogous to an LLM he had several tools available to him that allowed him to interact with his environment.

Tool	Description
Legs	Allows Dr. Watson to move throughout the house
Arms	Operates kitchen appliances
Kitchen Appliances	Prepares food and drink

Actions and Tools

- A Tool is simply a python function given to the LLM that should fulfill a clear objective

| Tool	Description
Web Search	Allows the agent to fetch up-to-date information from the internet
Cincinnati Parks Information Retrieval	Retrieves information we have stored in our Cincinnati Parks Cortex Search
Image Generation	Creates images based on text input
Current Weather Finder	Finds current weather for a zip-code using an API

Actions and Tools

- Broad Categories of actions

Type of Action	Description
Information Gathering	Performing web searches, querying databases, or retrieving documents
Tool Usage	Making API calls, running calculations, executing code
Environment Interaction	Manipulating a digital interface or controlling a physical device
Communication	Engaging with users via chat or collaborating with other agents in a multi-agent system

phData

Frameworks

Frameworks

LangGraph

AG2

Atomic Agents

smolagents

crewai

phData
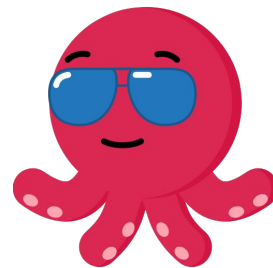
Why LangChain as an LLM Framework is Like Ripping Heaters in the 1980s

- It looks super cool, feels awesome while you're doing it, and for awhile we could say we didn't know any better, but a growing body of evidence suggests maybe we should stop.

Warning: The Surgeon General has determined LangChain has unnecessary abstractions, poor performance, and breaking changes like every 3 weeks.

phData