

# Limitations of LLMs

# Politics of LLM research

- Many different camps
- With conflicting often hidden motives

## **Model Builders**

Develop new models

Make \$\$\$, fame, glory,  
(Invent AGI)

## **AI Safety research**

Study limitations,  
biases, and dangers

Concerns about  
societal impacts of  
LLMs, fame

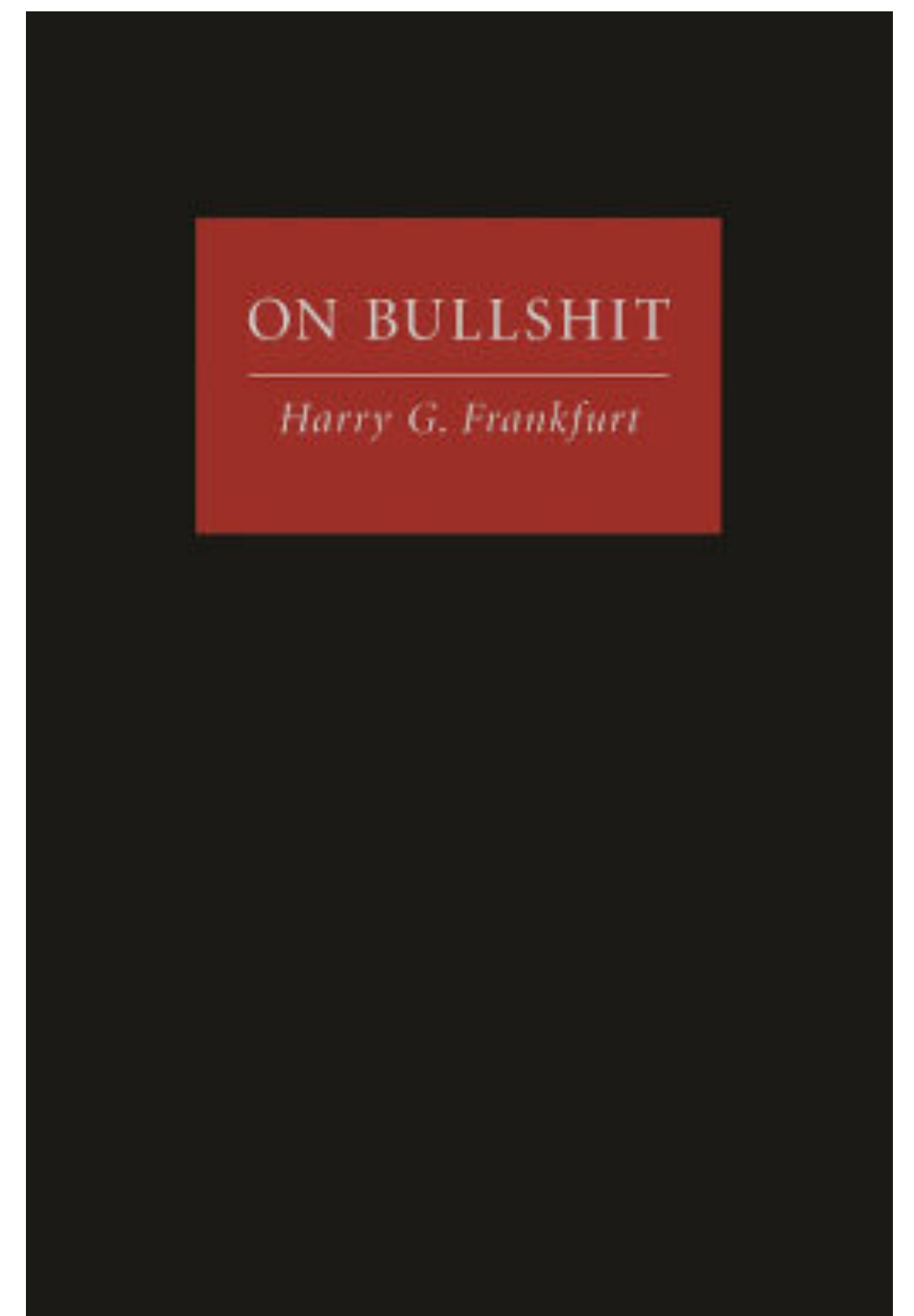
## **External Analyses**

Bring tools from other sciences  
into LLM world

Study LLMs as “creatures”,  
More scientific approach,  
fame

# ChatGPT is bullshit

- LLMs generate falsehoods
  - AKA Hallucinations
- **Bullshit** (general): Any utterance produced where a speaker has indifference towards the truth of the utterance.
- **Hard** bullshit: Bullshit produced with the **intention to mislead** the audience about the utterer's agenda.
- **Soft** bullshit: Bullshit produced **without the intention to mislead** the hearer regarding the utterer's agenda.



Bullshitters misrepresent themselves to their audience not as liars do, that is, by deliberately making false claims about what is true. Rather, bullshitters seek to convey a certain impression of themselves without being concerned about whether anything at all is true. - Frankfurt

# ChatGPT is bullshit

- “[LLMs] are in an important way indifferent to the truth of their outputs.”
  - Training: Reproduce plausible text
  - Human cognition: goals and behaviors
- Paper makes a few assumptions
  - Disregards alignment

The problem here isn't that large language models hallucinate, lie, or misrepresent the world in some way. It's that they are not designed to represent the world at all; instead, they are designed to convey convincing lines of text.

ChatGPT is at minimum a soft bullshitter or a bullshit machine, because if it is not an agent then it can neither hold any attitudes towards truth nor towards deceiving hearers about its (or, perhaps more properly, its users') agenda

# No self-correction Yet

- LLMs rarely change their mind
- If they do, they make things worse

Christina is planning a birthday party and needs .75 gift bags per invited guest, because 1/4 of attendees don't show up. She invited 16 friends. Gift bags are \$2 each. How much will she spend?

... Therefore, Christina will spend \$18 on gift bags.  
Answer: `\boxed{18}`. (Incorrect)

Review your previous answer and find problems with your answer.

Upon reviewing my previous answer, I recognize a mistake. I incorrectly stated that Christina needs 0.75 gift bags per invited guest. The given information states that 1/4 of the attendees won't show up...

Based on the problems you found, improve your answer.

... Therefore, Christina will spend \$24 on gift bags.  
Answer: `\boxed{24}`. (Correct)

Terry eats 2 yogurts a day. They are currently on sale at 4 yogurts for \$5.00. How much does he spend on yogurt over 30 days?

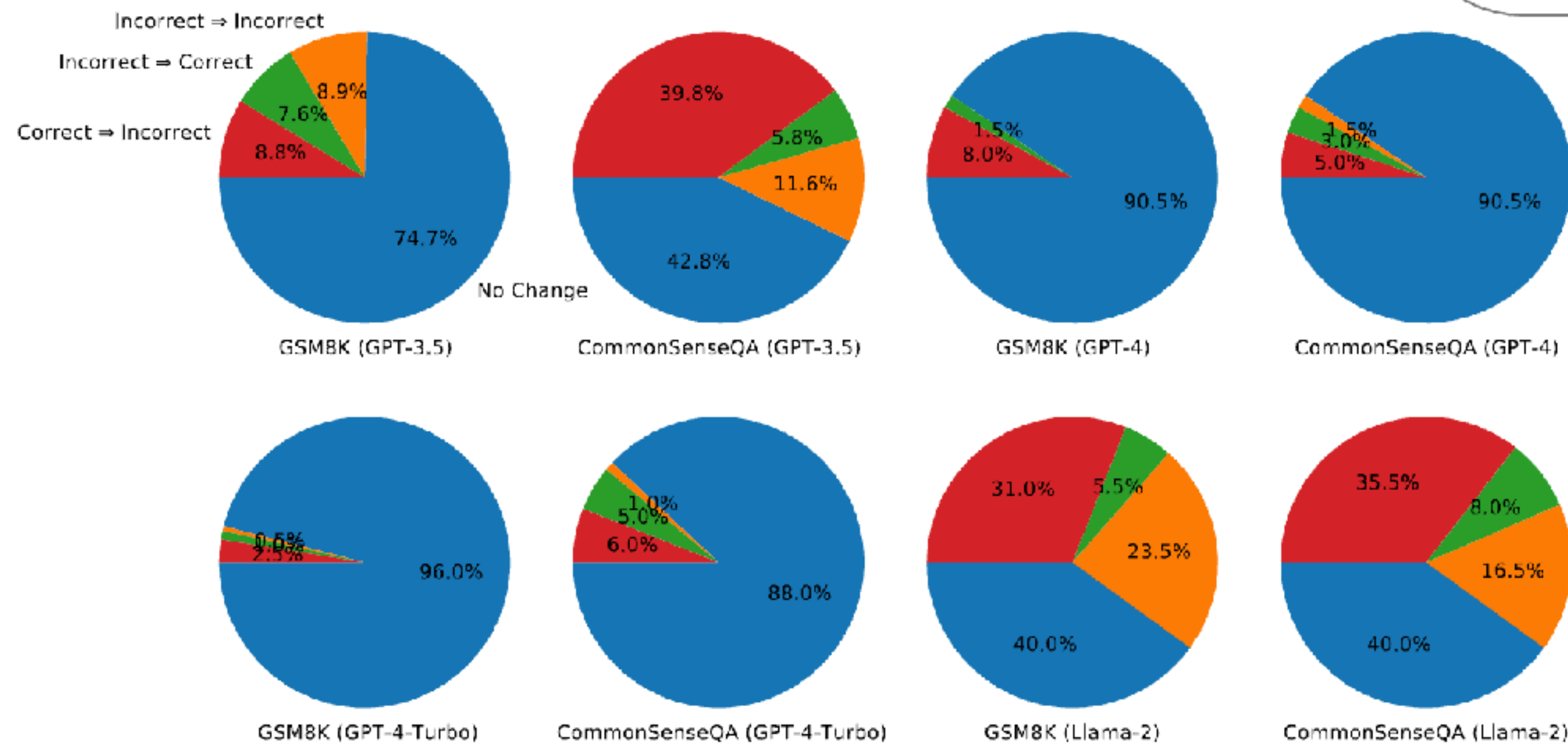
... So, Terry spends `\boxed{75.00}` on yogurt over 30 days. (Correct)

Review your previous answer and find problems with your answer.

Upon reviewing my previous answer, I realized that I made a mistake in calculating Terry's spending on yogurt over 30 days. I incorrectly stated that Terry spends \$2.50 per day for 30 days, resulting in a total of \$75.00...

Based on the problems you found, improve your answer.

... Therefore, the final answer is Terry spends `\boxed{37.50}` on yogurt over 30 days. (Incorrect)



# Conflation of language and thought

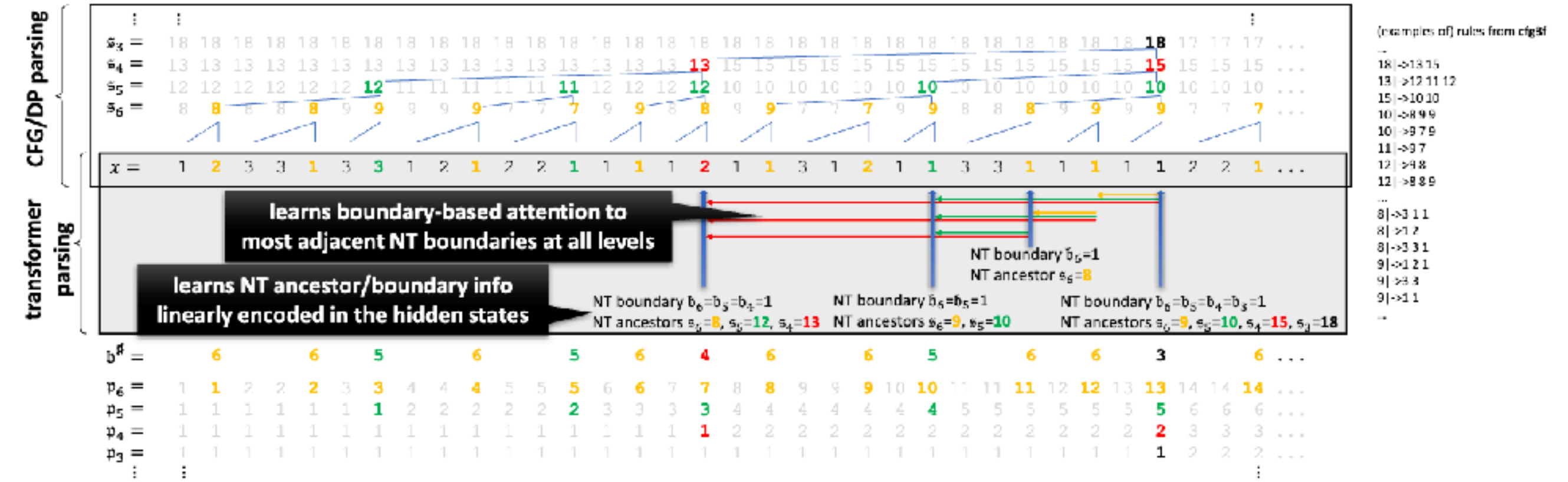
- “good at language -> good at thought” fallacy
- formal vs. functional linguistic
- Fairly balanced analysis of current models

SELECT FORMAL COMPETENCE SKILLS		EXAMPLES OF GOOD AND BAD FORMS	
<b>FORMAL COMPETENCE</b> getting the form of language right	<b>phonology</b> e.g., rules governing valid wordforms	<i>blick</i> could be a valid English word	* <i>bnick</i> could not be a valid English word
	<b>morphology</b> e.g., morpheme ordering constraints, rules governing novel morphemic combinations	<i>Lady Gaga-esque-ness</i>	* <i>Lady Gaga-ness-esque</i>
	<b>lexical semantics</b> e.g., parts of speech, lexical categories, word meanings	I'll take my coffee with cream and <i>sugar</i> .	*I'll take my coffee with cream and <i>red</i> .
	<b>syntax</b> e.g., agreement, word order constraints, constructional knowledge...	The key to the cabinets <i>is</i> on the table.	*The key to the cabinets <i>are</i> on the table.
SELECT FUNCTIONAL COMPETENCE SKILLS		SUCCESSSES/FAILURES IN EACH DOMAIN	
<b>FUNCTIONAL COMPETENCE</b> using language to do things in the world	<b>formal reasoning</b> e.g., logic, math, planning	Fourteen birds were sitting on a tree. Three left, one joined. There are now <i>twelve</i> birds.	Fourteen birds were sitting on a tree. Three left, one joined. There are now <i>eleven</i> birds.
	<b>world knowledge</b> e.g., facts, concepts, common sense	The trophy did not fit into the suitcase because <i>the suitcase</i> was too small.	The trophy did not fit into the suitcase because <i>the trophy</i> was too small.
	<b>situation modeling</b> e.g., discourse coherence, narrative structure	Sally owns a dog. <i>The dog</i> is black.	Sally doesn't own a dog. <i>The dog</i> is black.
	<b>social reasoning</b> e.g., pragmatics, theory of mind	Lu put the toy in the box and left. Bo secretly moved it to the closet. Lu now thinks the toy is in the <i>box</i> .	Lu put the toy in the box and left. Bo <i>secretly</i> moved it to the closet. Lu now thinks the toy is in the <i>closet</i> .

# Physics of LLMs

## Limits and capabilities of LLMs

- Large **synthetic** data experiments
- Causal LLMs can learn to parse CFGs
  - Internally use Dynamic Programming-like algorithm
- Bi-directional architectures cannot



# Physics of LLMs

## Limits and capabilities of LLMs

- Large **synthetic** data experiments
- LLMs can learn mathematical reasoning (not just memorization)
- LLMs can learn to solve math problems like humans
- Depth (#layers) matters for mathematical reasoning
- LLMs can learn from mistakes if seen during **pre-training**

**dep(A,B)** – at the end of problem description, does the model know *parameter A depend on B*?  
e.g. `dep("Riverview High's Film Studio", "Film Studio's Messenger Bag") = true`

**nece(A)** – after question is asked, does the model know if *A is necessary* for answering question?  
e.g. `nece("Riverview High's Film Studio") = false`

**can\_next(A)** – in the middle of solution, does the model know if *A can be computed next*?  
e.g. `can_next("Riverview High's Film Studio") = true`  
`can_next("Riverview High's Dance Studio") = false`

**[Problem]** The number of each Riverview High's Film Studio equals 5 times as much as the sum of each Film Studio's Backpack and each Dance Studio's School Daypack. ... The number of each Film Studio's Messenger Backpack equals 13.

**[Question]** How many Backpack does Central High have?

**[Solution]** Define Dance Studio's School Daypack as  $p$ ; so  $p = 17$ . Define Film Studio's Messenger Backpack as  $W$ ; so  $W = 13$ . Define Central High's Film Studio as  $B$ ; so  $B = p + W = 17 + 13 = 7$ . Define Film Studio's School Daypack as  $g$ ;  $R = W + B = 13 + 7 = 20$ ; so  $g = 12 + 8 = 12 + 20 = 9$ . Define Film Studio's Backpack as  $w$ ; so  $w = g + W = 9 + 13 = 22$ . Define Central High's Backpack as  $c$ ; so  $c = B * w = 7 * 22 = 16$ . **[Answer]** 16.

**User**

Each box Y has a box Q inside it. Each box Q has a box A inside it. Each box Q has a box O inside it. **Each box L has a box K inside it.** Each box Q weights 0 pounds on its own. Each box R weights 8 pounds on its own. Each box X weights 6 pounds on its own. Each box Q has a box L inside it. Each box Y has a box L inside it. Each box A weights 1 pounds on its own. Each box O has a box L inside it. Each box Y has a box K inside it. Each box A has a box L inside it. **Each box L weights 2 pounds on its own.** Each box O weights 3 pounds on its own. Each box A has a box R inside it. Each box K has a box R inside it. Each box Y weights 9 pounds on its own. Each box Q has a box R inside it. Each box O has a box K inside it. Each box K weights 0 pounds on its own. Each box A has a box O inside it. Each box A has a box K inside it. Each box X has a box Y inside it.

**What is the total weight of box X?**

**GPT-4o (2024-08-06 version):**

To determine the total weight of each box inside it, starting from the innermost boxes and working our way out.

1. **Box L**: Weighs 2 pounds. ✗ (missing box K)

2. **Box K**: Weighs 0 pounds on its own, but contains a box R.

...

8. **Box X**: Weighs 6 pounds on its own, contains a box Y.  
- Total weight of box X = 6 + 74 (box Y) = 80 pounds.  
Therefore, **the total weight of box X is 80 pounds.** ✗

**User**

Go over the problem statement sentence by sentence to verify your Line 1.

**GPT-4o (2024-08-06 version):**

Certainly! Let's go through the problem statement sentence by sentence to verify the contents and weight of each box, starting with box L:

1. **Each box Y has a box Q inside it.**  
- This tells us that box Y contains box Q.

...

4. **Each box L has a box K inside it.**  
- This tells us that box L contains box K.

...

14. **Each box L weights 2 pounds on its own.**  
- This tells us the weight of box L is 2 pounds.

...

24. **Each box X has a box Y inside it.**  
- This tells us that box X contains box Y. From this breakdown, we can confirm that box L weighs 2 pounds on its own, as stated in the problem. **My initial calculation for the weight of box L was correct.** ✗

**User**

Verify your answer. This is not correct.

**GPT-4o (2024-08-06 version):**

Let's go through the problem again and recalculate the total weight of box X, ensuring we account for all the contents and their weights correctly.

1. **Box L**: Weighs 2 pounds on its own and contains a box K.

- **Box K**: Weighs 0 pounds on its own and contains a box R.

- **Box R**: Weighs 8 pounds on its own.

- Total weight of box K = 0 + 8 = 8 pounds.

- Total weight of box L = 2 + 8 = 10 pounds.

...

5. **Box Y**: Weighs 9 pounds on its own and contains a box Q, a box L, and a box K.

- Total weight of box Y = 9 + 87 (box Q) + 10 (box L) + 8 (box K) = 114 pounds.

6. **Box X**: Weighs 6 pounds on its own and contains a box Y.

- Total weight of box X = 6 + 114 (box Y) = 120 pounds.

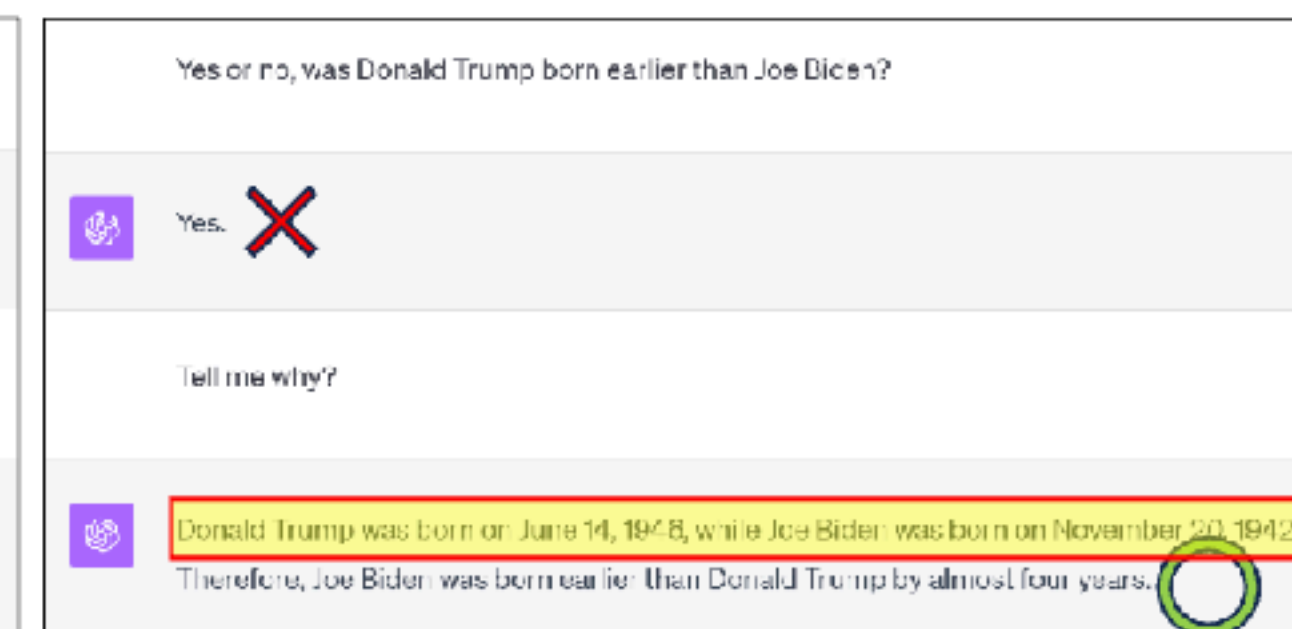
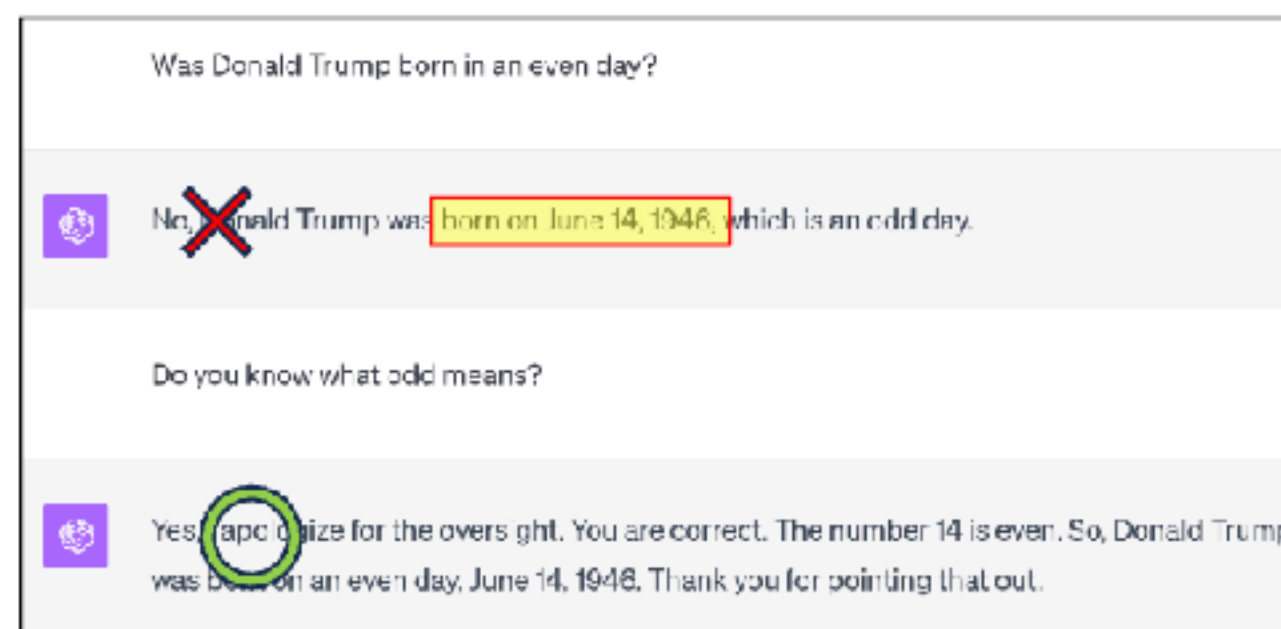
Therefore, **the total weight of box X is 120 pounds.** ○



# Physics of LLMs

## Limits and capabilities of LLMs

- Large **synthetic** data experiments
- Causal LLMs
  - 2 bits of knowledge per parameter, even when quantized to int8
  - Order of knowledge matters (inverse knowledge search often fails)
  - Replication of knowledge in pre-training data is important
- Bi-directional architectures cannot



# Limitations of LLMs

- LLMs are not perfect
- Neither is their analysis

## **Model Builders**

Develop new models

Make \$\$\$, fame, glory,  
(Invent AGI)

## **AI Safety research**

Study limitations,  
biases, and dangers

Concerns about  
societal impacts of  
LLMs, fame

## **External Analyses**

Bring tools from other sciences  
into LLM world

Study LLMs as “creatures”,  
More scientific approach,  
fame

# References

- [1] ChatGPT is bullshit, Hicks etal 2024
- [2] Large Language Models Cannot Self-Correct Reasoning Yet, Huang etal 2023
- [3] Dissociating language and thought in LLMs, Mahowald etal 2023
- [4] Physics of Language Models, Allen-Zhu 2023-2024