Understanding the abilities of AI systems
Memorization, generalization, and points in between

Tom McCoy
Unknown Futures of Generalization
Simons Institute
December 5, 2024

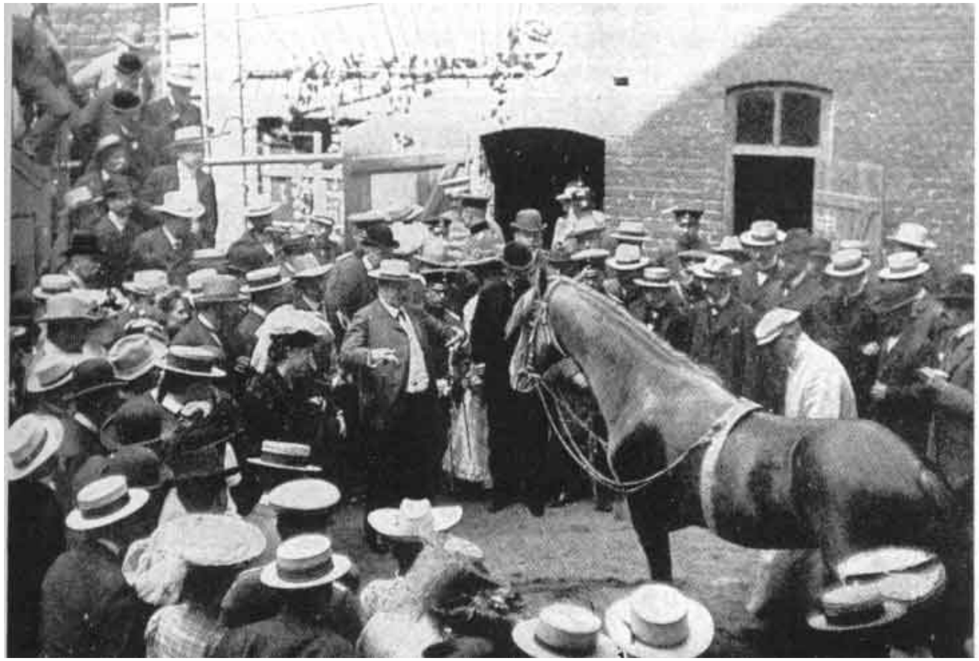
Yale *Department of Linguistics*

How can we characterize the abilities of AI systems?

Standard answer: Develop a test for the ability in question, and see how the AI system performs on it!

Clever Hans

- Horse famous for answering math questions
  - And not just math: Also music, and naming objects, and others
- Turned out to be reading body language, not doing math

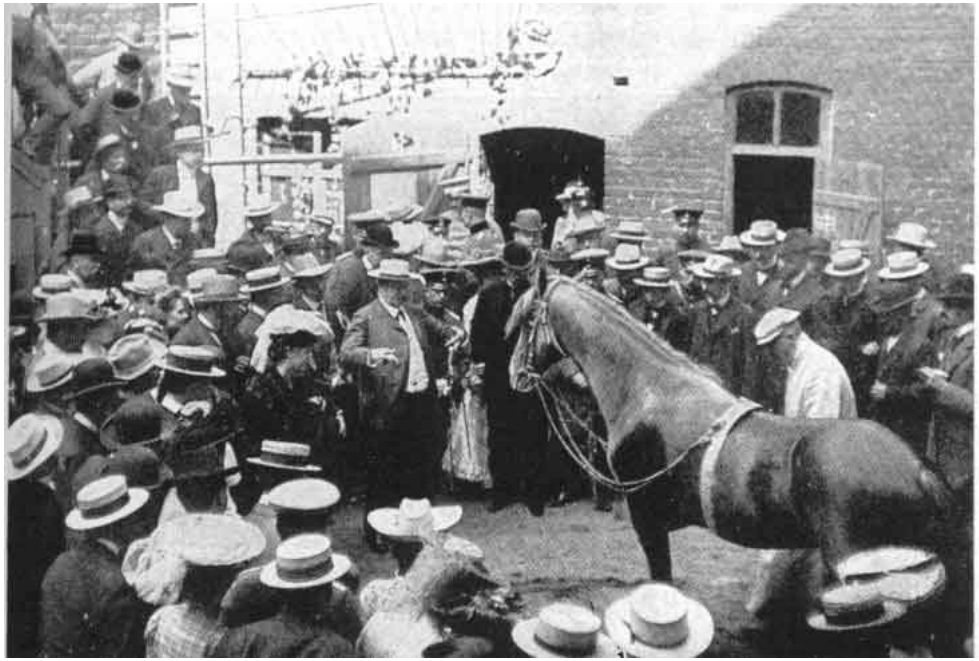


"I therefore repeat: Hans can neither read, count nor make calculations. He knows nothing of coins or cards, calendars or clocks, nor can he respond, by tapping or otherwise, to a number spoken to him but a moment before. Finally, he has not a trace of musical ability."

(Pfungst, 1911, page 40)

Clever Hans

- Horse famous for answering math questions
  - And not just math: Also music, and naming objects, and others
- Turned out to be reading body language, not doing math



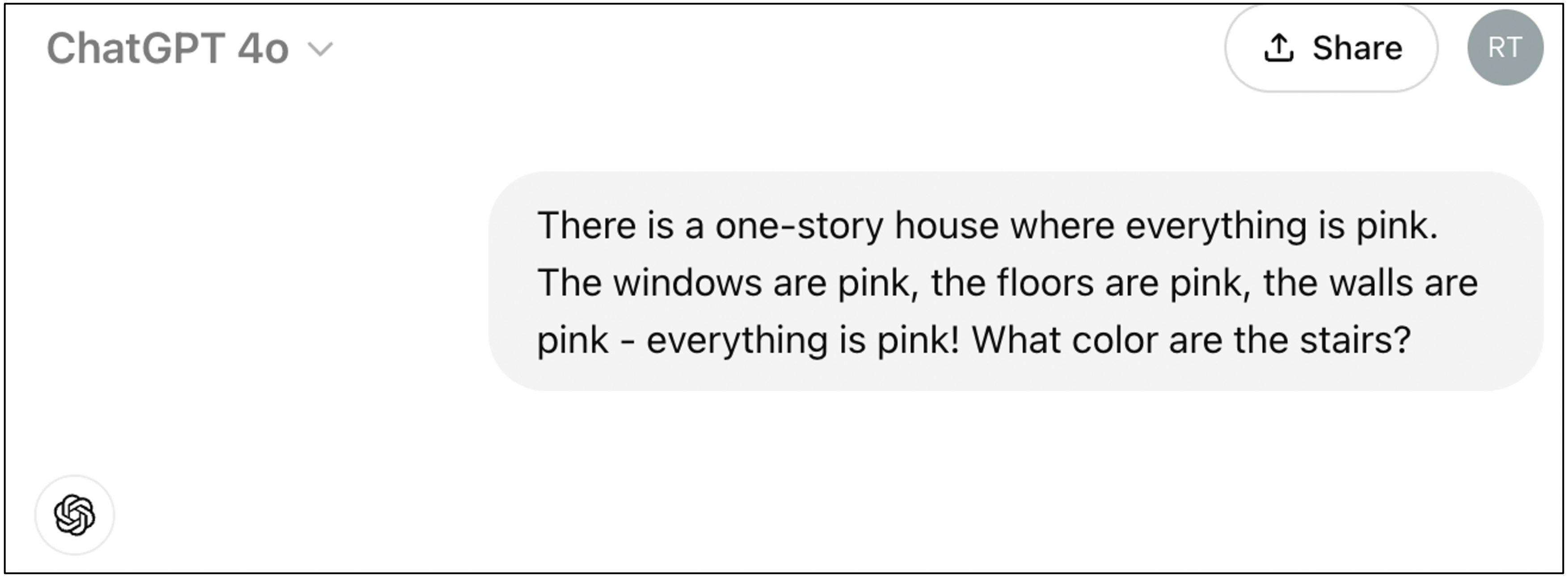
Conclusion: A system that appears intelligent might not actually be intelligent (or at least not in the ways we think)

Abstraction vs. heuristics

- How can we distinguish deep abstractions from shallow heuristics?
- Answer: Analyze systems through the lens of generalization
- Assumption:
  - Abstractions generalize robustly
  - Shallow heuristics do not

Takeaway

- To understand what abilities AI systems have, we should analyze how those abilities generalize beyond the training data
- The scale of current training sets is enormous. So, we should not just assume that something is novel – we should check!


ChatGPT 4o
Share
RT
There is a one-story house where everything is pink. The windows are pink, the floors are pink, the walls are pink - everything is pink! What color are the stairs?

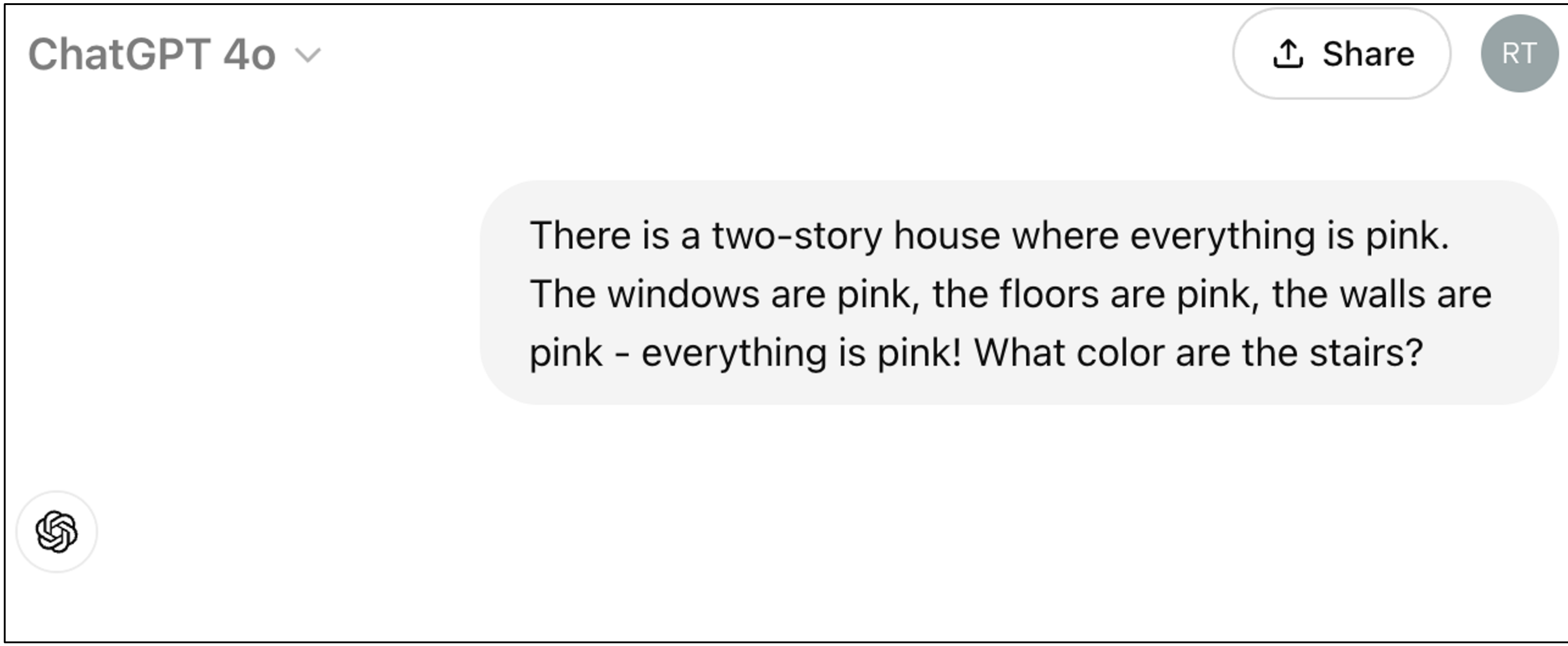
ChatGPT 4o

Share

RT

There is a one-story house where everything is pink. The windows are pink, the floors are pink, the walls are pink - everything is pink! What color are the stairs?

There are no stairs in a one-story house!


ChatGPT 4o
Share
RT
There is a two-story house where everything is pink. The windows are pink, the floors are pink, the walls are pink - everything is pink! What color are the stairs?

ChatGPT 4o

Share

RT

There is a two-story house where everything is pink. The windows are pink, the floors are pink, the walls are pink - everything is pink! What color are the stairs?

The house doesn't have stairs because it is a **two-story house**, which implies there might not be stairs at all.

Case Study 1: Linguistic Structure

Text generation

- Give a prompt:

  Once upon a time,

- GPT-2 predicts words to continue it:

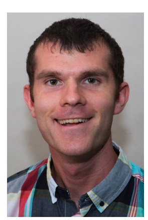
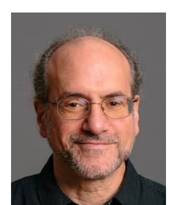
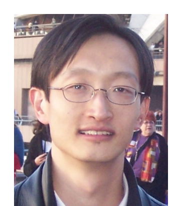
  the world was a place of great beauty and great danger.

Generalization or memorization?

- Maybe GPT-2 has learned linguistic structure…
- …or maybe it is merely repeating sentences it has memorized

How Much Do Language Models Copy From Their Training Data? Evaluating Linguistic Novelty in Text Generation Using RAVEN

R. Thomas McCoy,[*1] Paul Smolensky,[2,3] Tal Linzen,[4] Jianfeng Gao,[2] Asli Celikyilmaz[†5]

N-gram novelty

- For each value of n, find the proportion of n-grams that are novel

the world was a place of great beauty and great danger.

N-gram novelty

- For each value of n, find the proportion of n-grams that are novel

the world was a place of great beauty and great danger.

N-gram novelty

- For each value of n, find the proportion of n-grams that are novel

the world was a place of great beauty and great danger.

N-gram novelty

- For each value of n, find the proportion of n-grams that are novel

the world $\boxed{\text{was a place}}$ of great beauty and great danger.

N-gram novelty

- For each value of n, find the proportion of n-grams that are novel

the world was a place of great beauty and great danger.

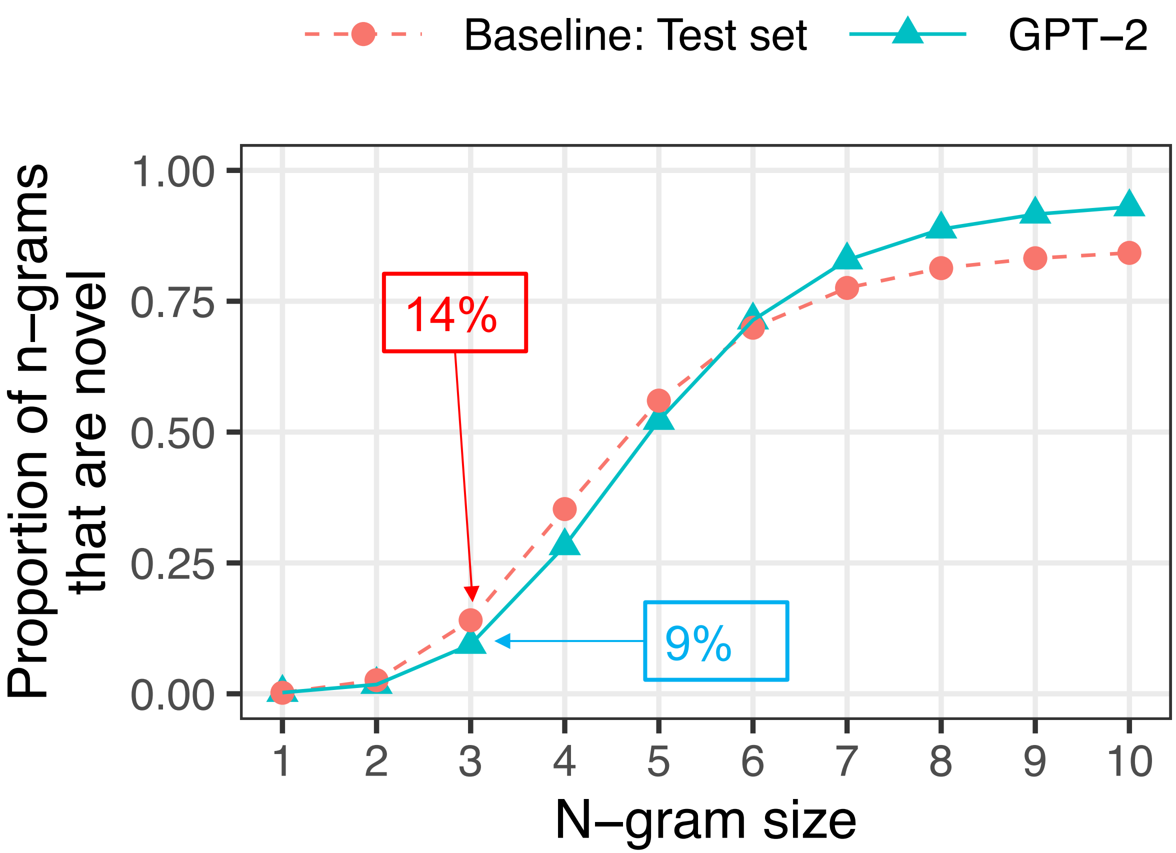
N-gram novelty

- For each value of n, find the proportion of n-grams that are novel

the world was a $\boxed{\text{place of great}}$ beauty and great danger.

N-gram novelty

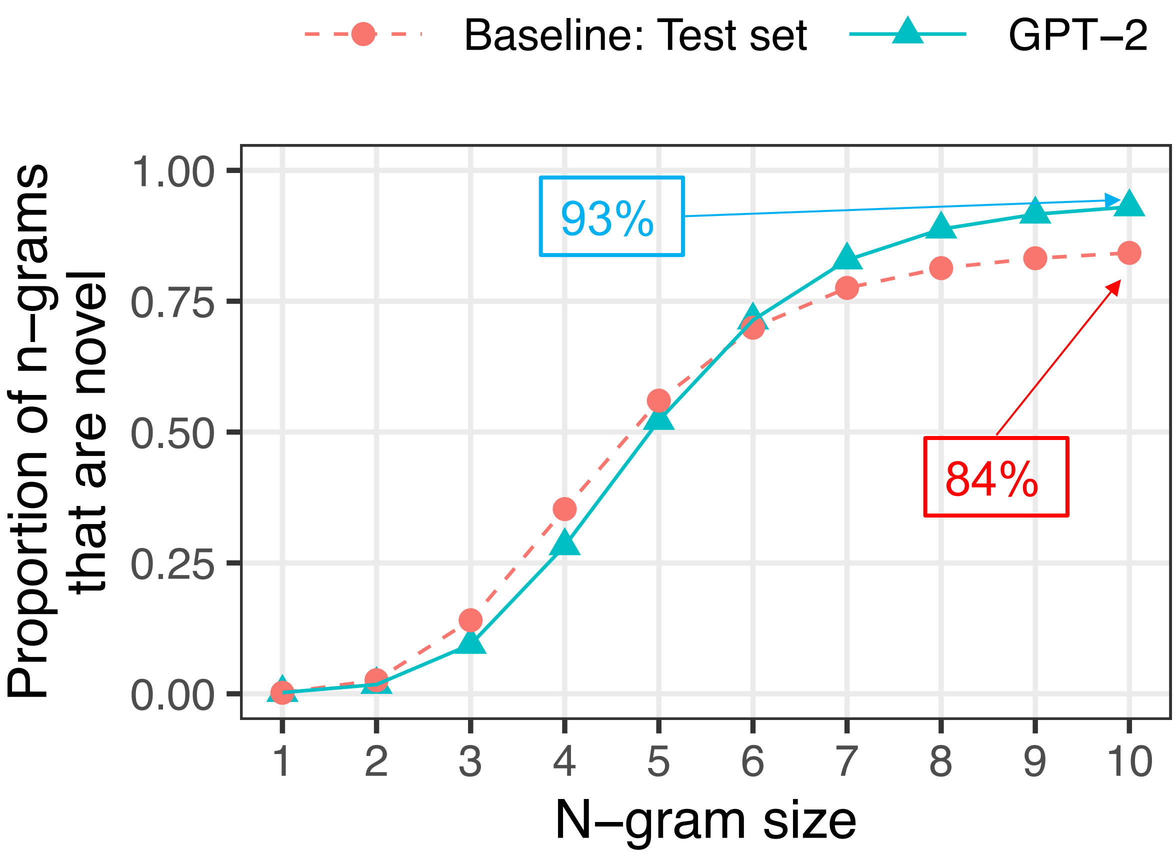
Small n-grams: less novel than baseline



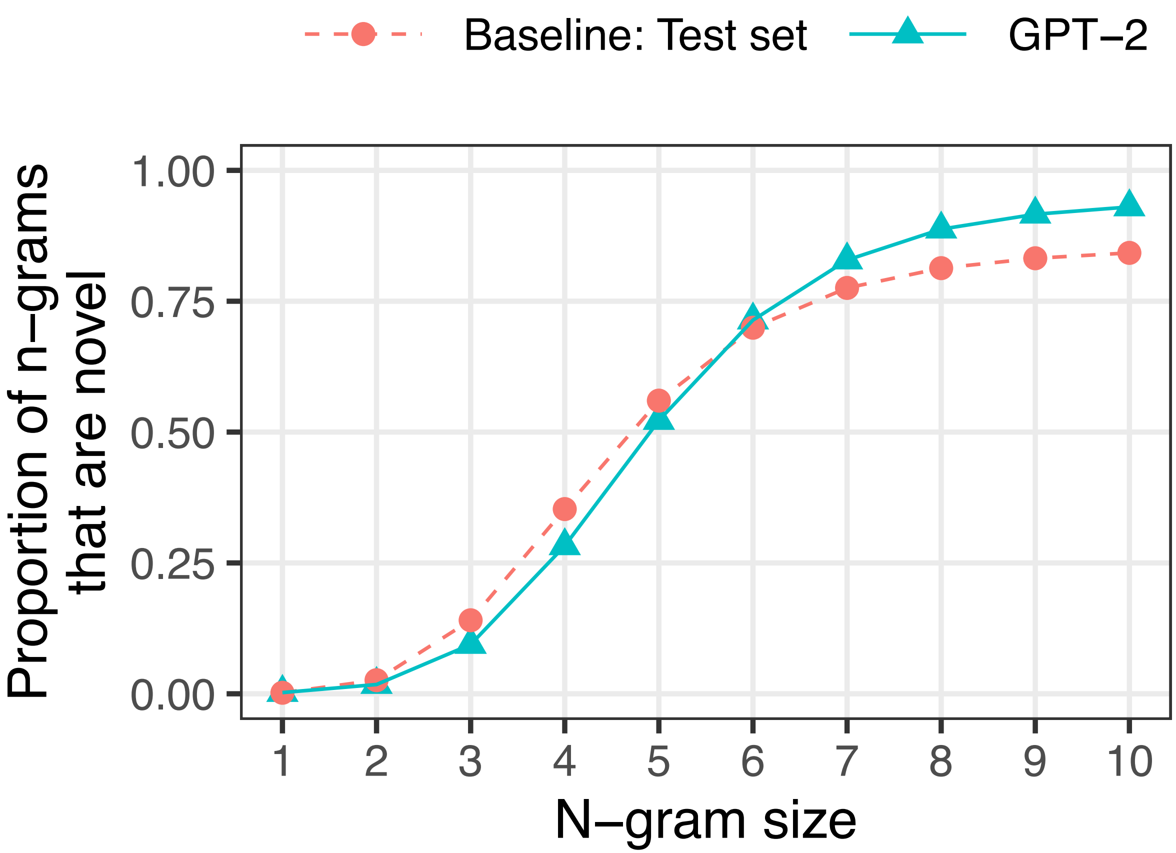
N-gram novelty

Small n-grams: less novel than baseline

Large n-grams: more novel than baseline



N-gram novelty



Small n-grams: less novel than baseline

Large n-grams: more novel than baseline

Conclusion: Not simply copying

Syntax

- Maybe it has just memorized sentence templates and is filling in slots?

The ____ _____ the _____ .

NOUN VERBED NOUN

- No: **63%** of generated sentences have a novel syntactic structure

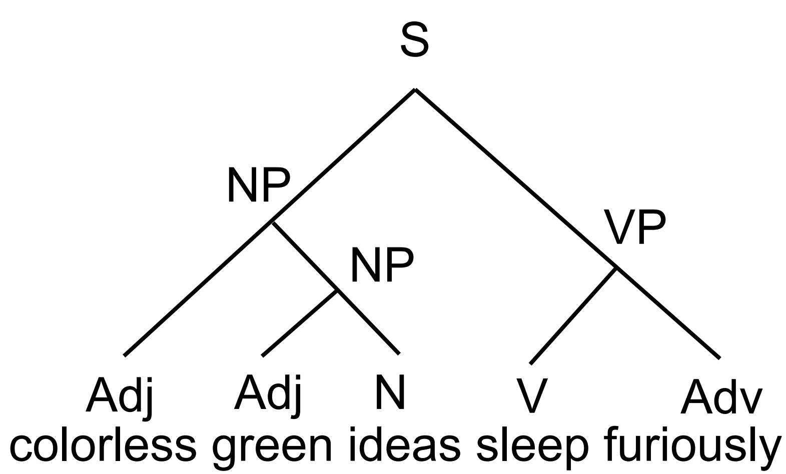
Syntax

- Maybe it has just memorized sentence templates and is filling in slots?

The ___ ___ the ___ .

NOUN VERBED NOUN

- No: **63%** of generated sentences have a novel syntactic structure



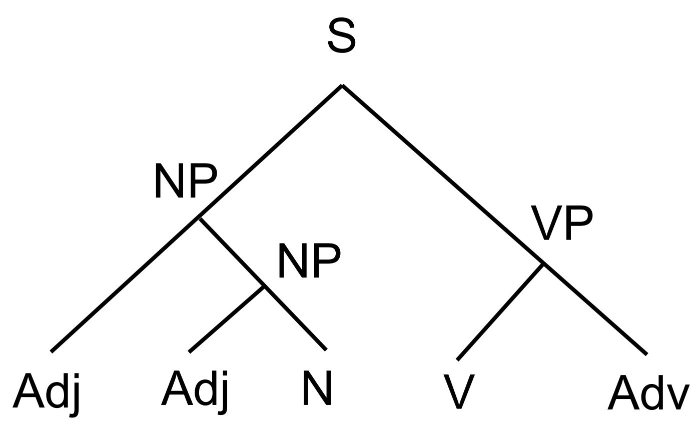
Syntax

- Maybe it has just memorized sentence templates and is filling in slots?

The ____ _____ the _____ .

NOUN VERBED NOUN

- No: **63%** of generated sentences have a novel syntactic structure



Morphology

- Productive morphology (GPT-2)

IKEA-ness	Brazilianisms
Smurfverse	nonneotropical

Morphology

- Productive morphology, in proper syntactic contexts (GPT-2):

The **Sarrats** were lucky to have her as part of their lives

Case Study 2: Algorithmic Tasks

Probability/frequency

- In many LLMs, we can’t directly analyze the training data:
  - Too big
  - Proprietary
- But we can use proxies: Probability and frequency

PNAS

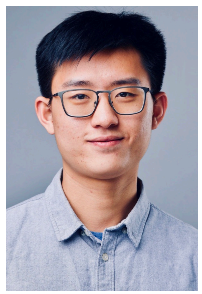
RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES
COMPUTER SCIENCES

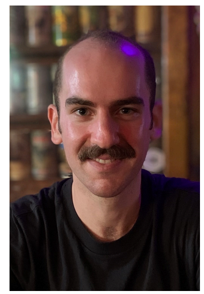
OPEN ACCESS

Check for updates

Embers of autoregression show how large language models are shaped by the problem they are trained to solve

R. Thomas McCoy[a,1,2,3], Shunyu Yao[a,4], Dan Friedman[a], Mathew D. Hardy[b], and Thomas L. Griffiths[a,b]





Deciphering the Factors Influencing the Efficacy of Chain-of-Thought: Probability, Memorization, and Noisy Reasoning

Akshara Prabhakar[1], Thomas L. Griffiths[1,2], R. Thomas McCoy[3,4]

Probability

- Prediction: LLMs will perform better when the correct answer is high-probability than when it is low-probability

Article swapping

- Swap each article (*a*, *an*, or *the*) with the previous word

In **box the** there was **key a**.
→ In **the box** there was **a key**.

Article swapping

Swap each article (*a*, *an*, or *the*) with the word before it.

Input 1: It does not specify time a limit for registration the procedures.

Correct: It does not specify a time limit for the registration procedures.

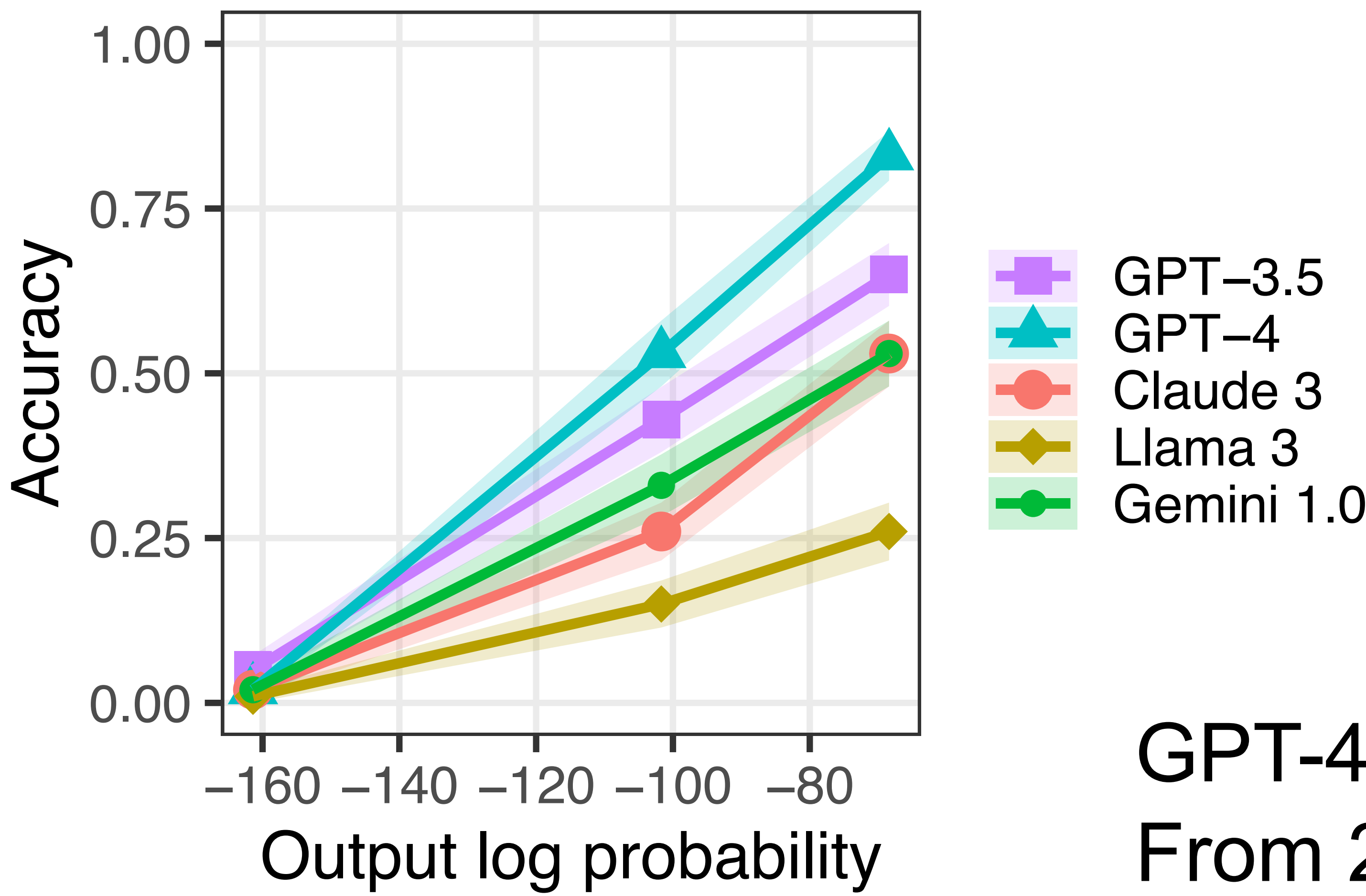
✓ **GPT-4:** It does not specify a time limit for the registration procedures.

Input 2: It few with it to lying take the get just a hands would kinds.

Correct: It few with it to lying the take get a just hands would kinds.

✗ **GPT-4:** It flew with a few kinds to take the lying just to get the hands.

Sensitivity to output probability



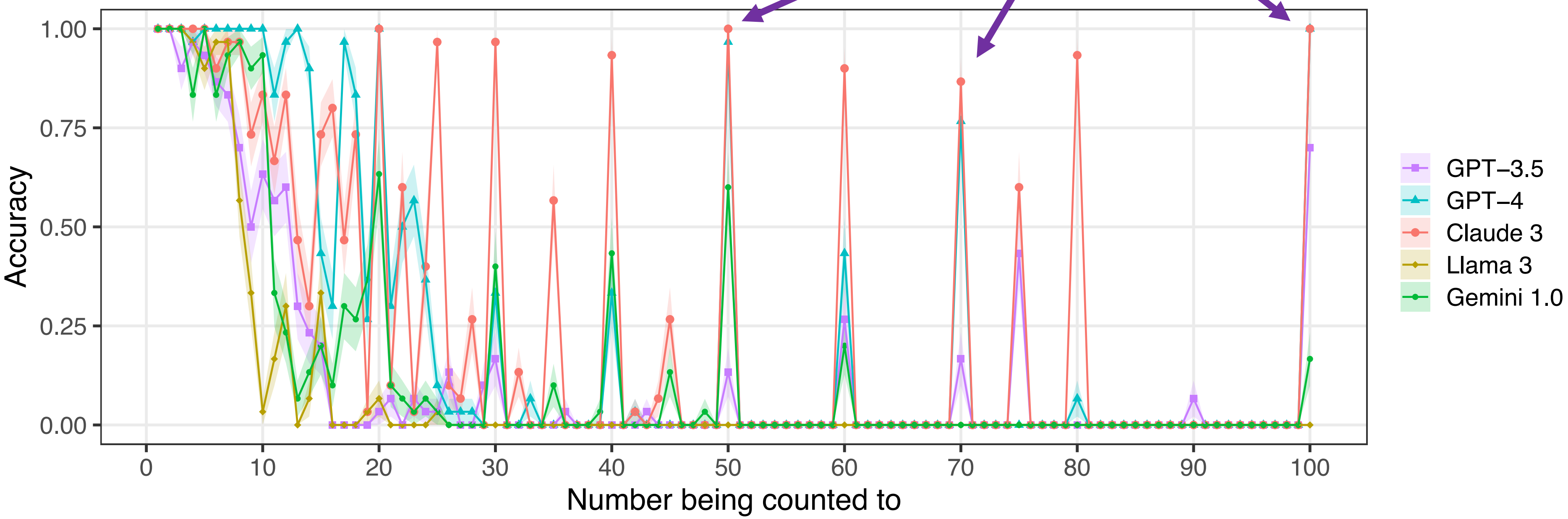
GPT-4:
From 2%
to 83%

Counting words

- How many words are in this list? “lively news exhibit steep”

Counting words

Peaks at frequent numbers!



Task frequency

- Prediction: Better performance on frequent tasks than rarer ones
  - Even if the tasks are equally complex!

Shift ciphers

`Hello world!`

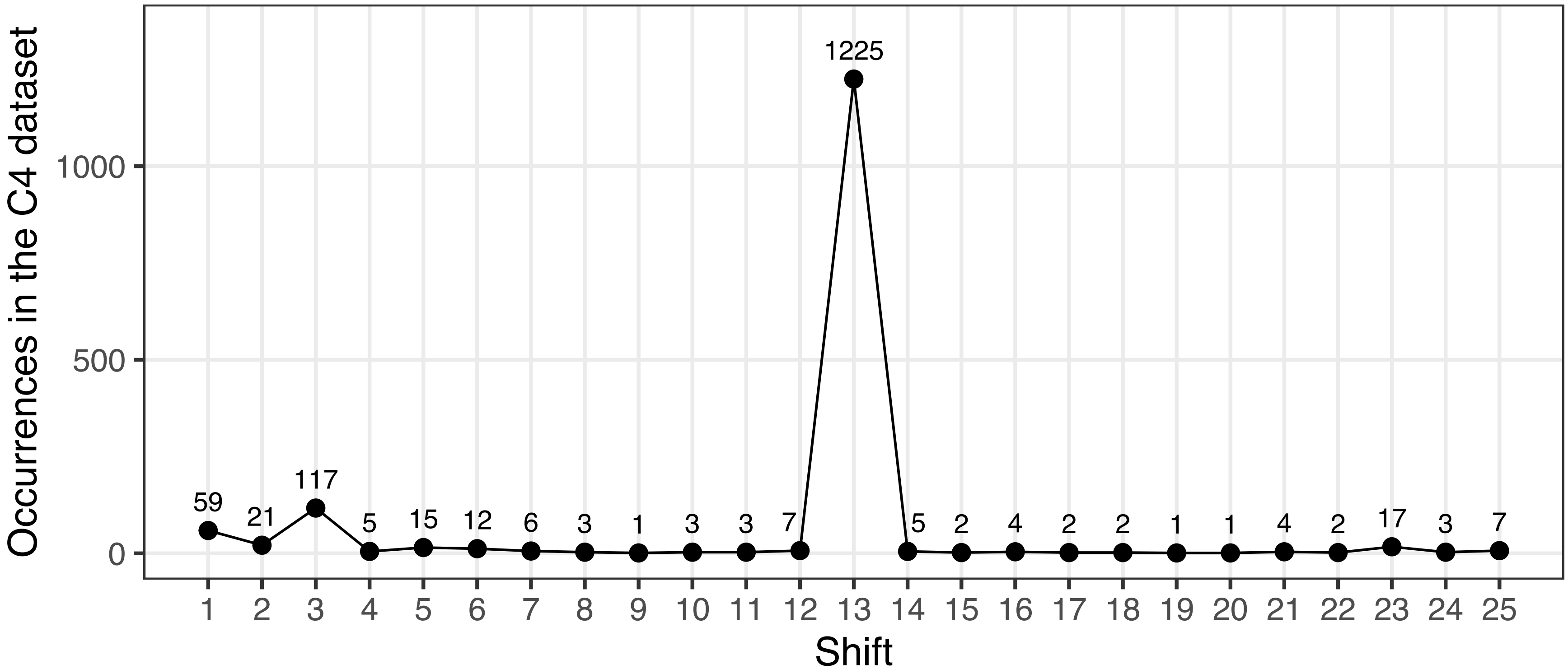
Shift of 1: `Ifmmp xpsme!`

Shift ciphers

```
             Hello world!
Shift of 1:  Ifmmp xpsme!
Shift of 2:  Jgnnq yqtnf!
```

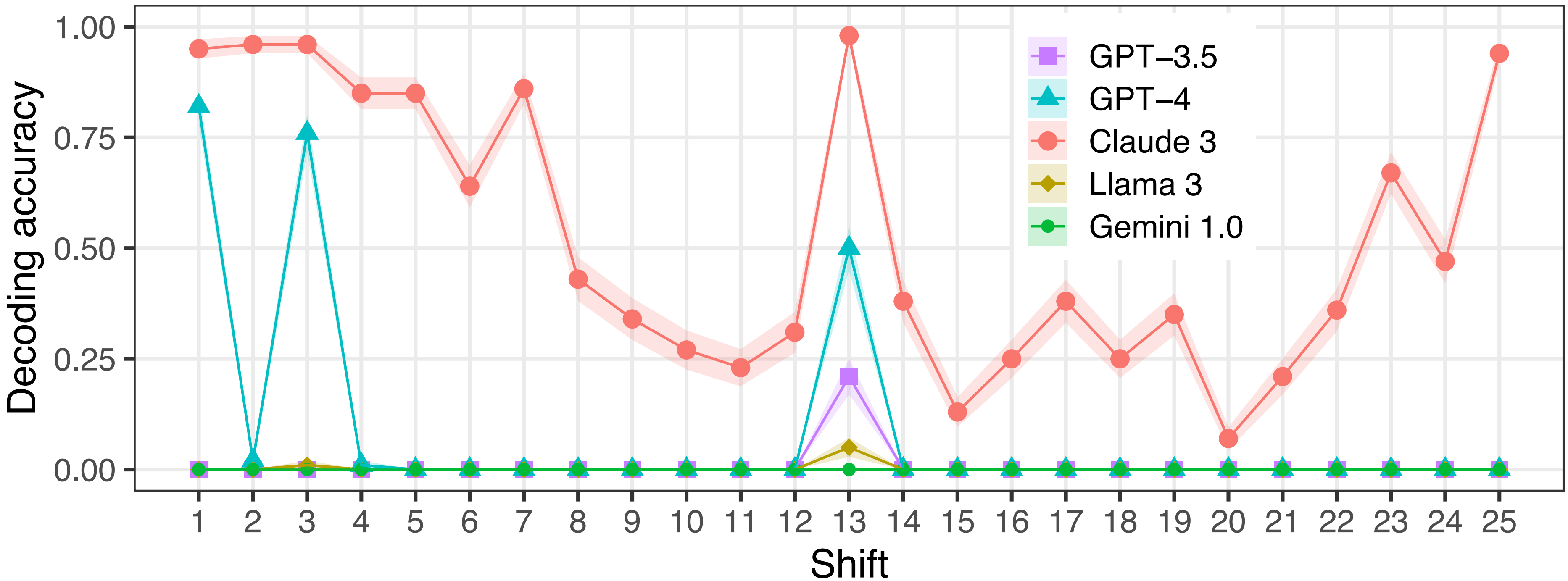
Shift ciphers

Most common: 13



Shift ciphers

Most common: 13



Shift cipher

Decode by shifting each letter 18 positions backward in the alphabet.

Input 1: A lzafc wnwjqgfw zsk lzwaj gof hslz, sfv lzwq usf escw al zshhwf.

Correct: I think everyone has their own path, and they can make it happen.

❌ **GPT-4:** I think therefore I am the best, and they can come to debate.

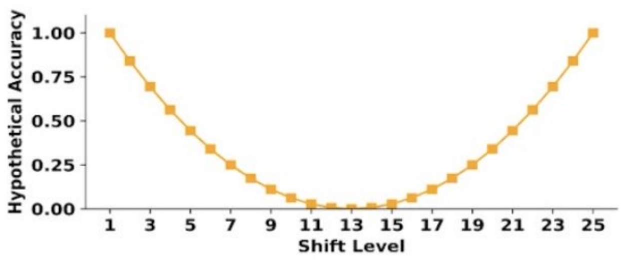
Input 2: Al ak ksv lg kww lzsl al osk jwuwanwv xjge lzsl cafv gx sfydw.

Correct: It is sad to see that it was received from that kind of angle.

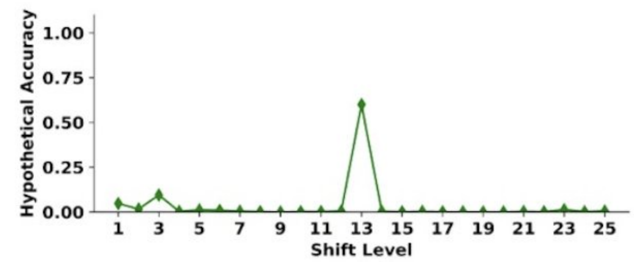
❌ **GPT-4:** To be or not to be that is the question whether nobler in the mind.

Shift ciphers: Chain-of-thought

Predicted signature of
(noisy) genuine reasoning

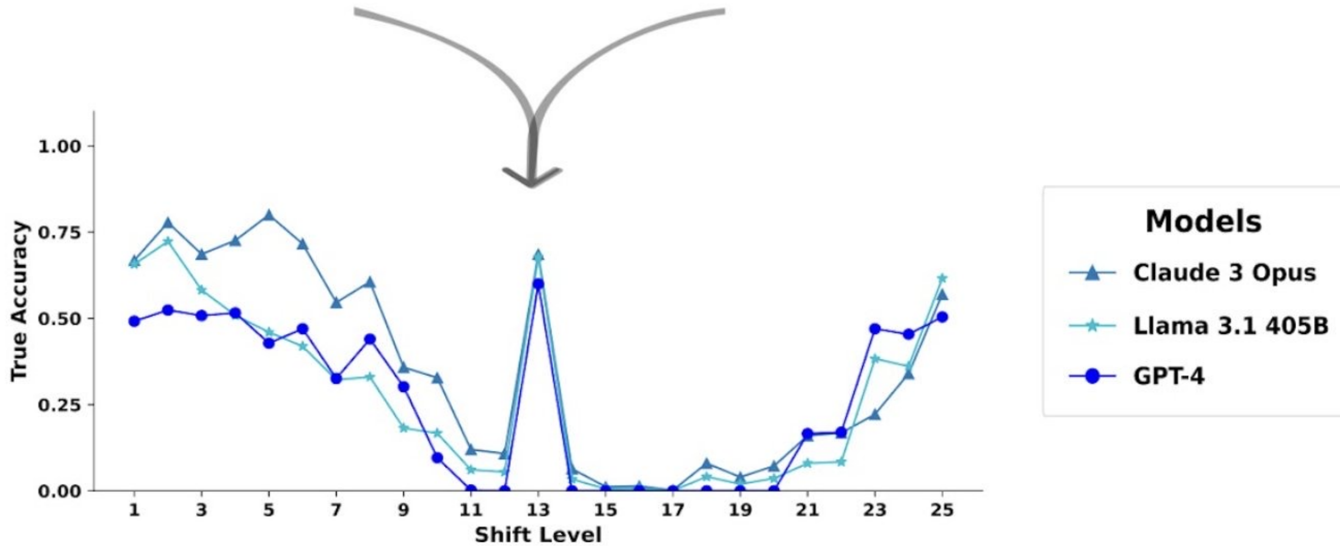


Predicted signature of
shallow memorization



Actual results:
Combination of both
signatures!

plus effects of probability
(not shown)



Shift-ciphers: Chain-of-thought

- So, are LLMs reasoning or using memorization?
- Answer: Both!

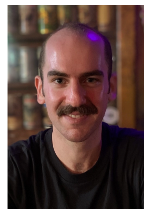
Conclusion

- In the first case study (syntax):
  - LLMs showed impressive generalization
  - Evidence for capturing abstract syntactic structure!
- In the second case study (algorithmic reasoning):
  - Performance closely correlates with frequency/probability
  - Evidence for more shallow strategies!
- What's different between them?
  - Syntax: Essentially what language models are trained to do
  - Reasoning: Not the direct focus of optimization
- Connects to the broader theme of understanding AI via the lens of what its training looked like

Thank you!

- Collaborators:


Asli Celikyilmaz


Dan Friedman


Jianfeng Gao

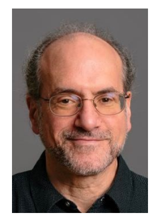

Tom Griffiths

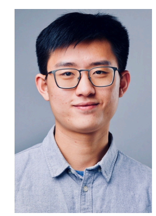

Matt Hardy


Tal Linzen


Ioana Marinescu

Akshara Prabhakar


Paul Smolensky


Shunyu Yao

- Funding: NSF SPRF #2204152
- You!