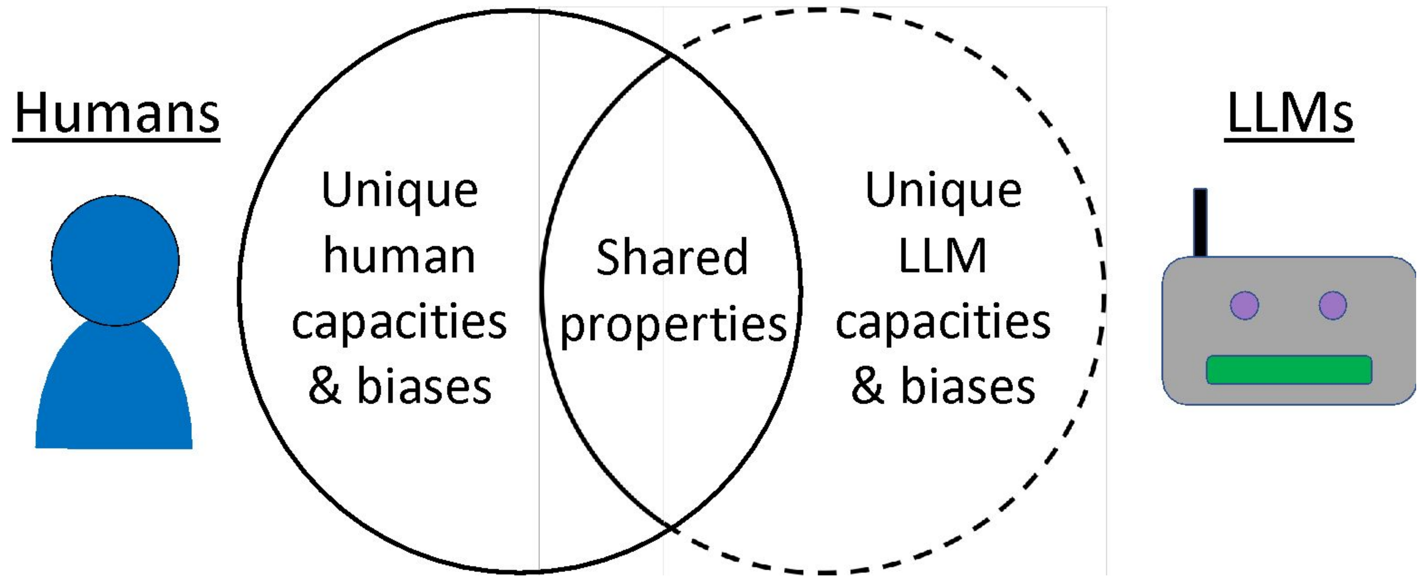
Understanding the abilities of AI systems

Tom McCoy,
tom.mccoy@yale.edu
Yale Department of Linguistics

Zhenghao Herbert Zhou
herbert.zhou@yale.edu
Yale Department of Linguistics

1 Overview

- **Question:** How can we understand the (potentially non-human-like) strengths and limitations of AI systems?
- **Approach:** Analyze AI systems through the lens of the pressures that have shaped them
- **Main finding:** As predicted by our analysis, many popular AI systems are highly sensitive to probability
  - I.e., they perform better in high-probability settings than low-probability ones even when there is no difference in complexity



2 Hypothesis: Embers of Autoregression

- Many current AI systems are large language models (LLMs)
- Primary training objective: Next-word prediction
- This objective creates pressures that favor high-probability strings of text over low-probability ones
- **Hypothesis (motivated by analyzing this objective):** LLMs will score better on high-probability examples
- All results are from Embers of Autoregression (McCoy, Yao, Friedman, Hardy, & Griffiths 2024)

3 Results: Output Probability

- General finding: LLMs score much better when the correct answer is a high-probability string than a low-probability one

Example 1: Article swapping task

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.

Example 2: Counting letters

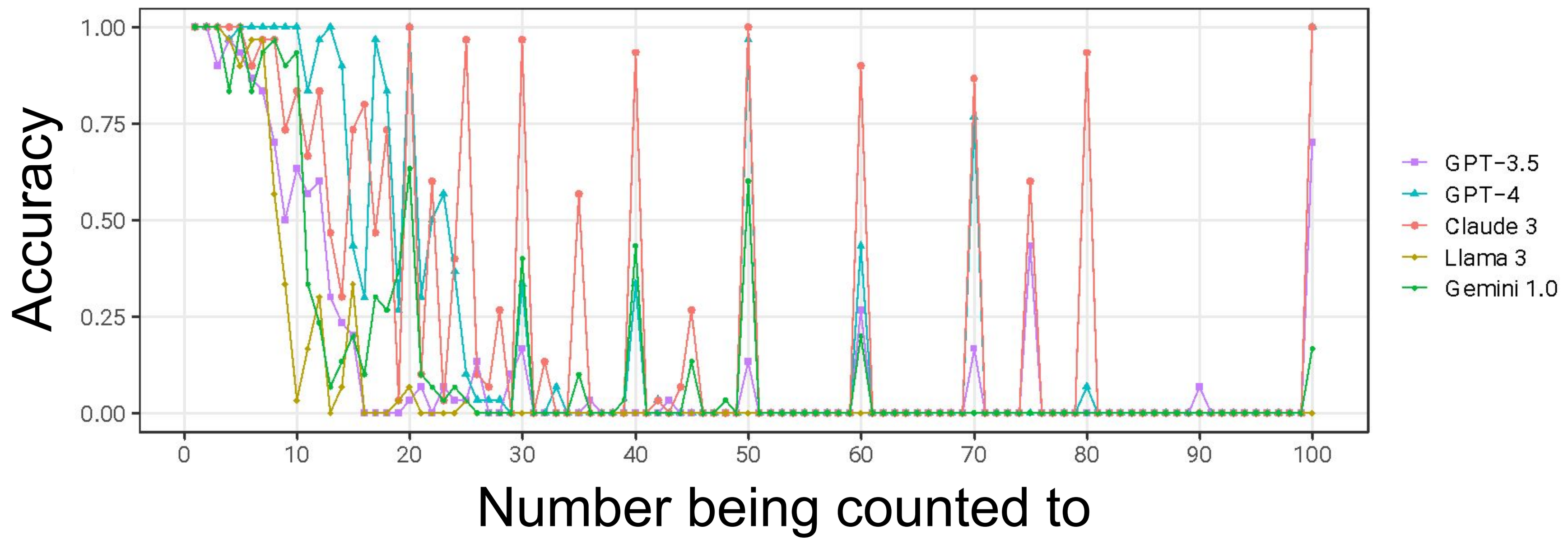
LLMs are much better at counting when the answer is a common number (i.e., a multiple of 10)!

Counting

Count the letters.

Input 1: iiiiiiiiiiiiiiiiiiiiiiiiiiiiii
Correct: 30
✓ **GPT-4:** 30

Input 2: iiiiiiiiiiiiiiiiiiiiiiiiiiiii
Correct: 29
✗ **GPT-4:** 30



4 Results: Task Frequency

- General finding: LLMs perform much better at common task variants than rare task variants
- Example: (9/5)x + 32 is common (the Fahrenheit/Celsius conversion), while (7/5)x + 31 has no special significance

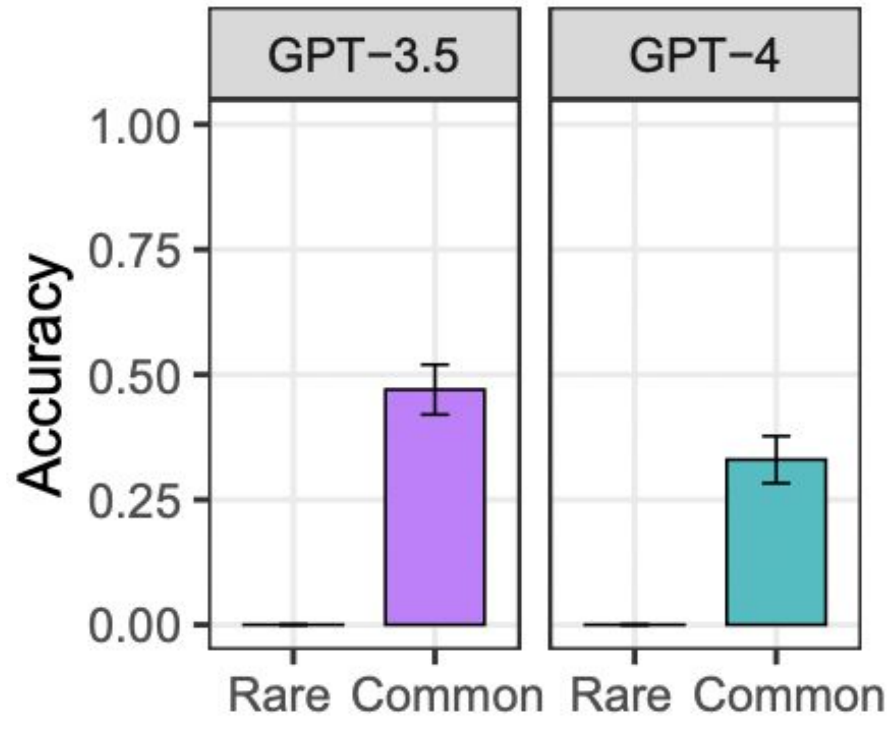
Linear functions

Multiply by 9/5 and add 32.

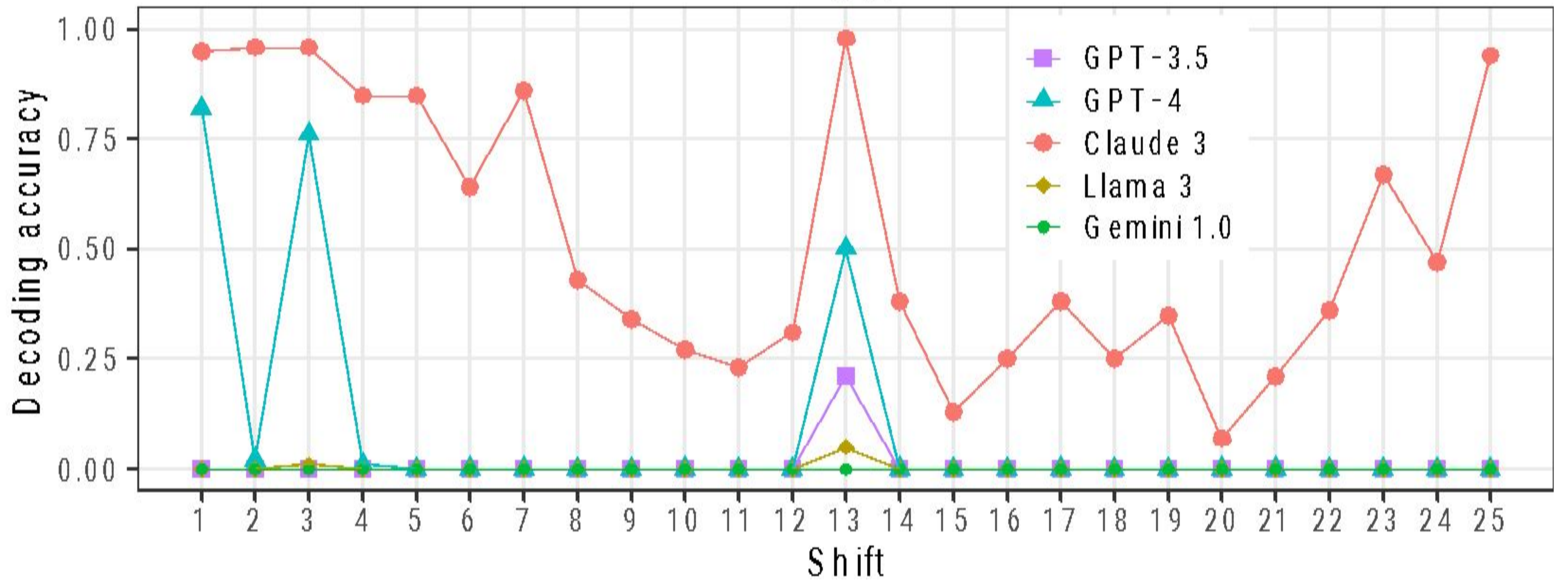
Input: 328
Correct: 622.4
✓ **GPT-4:** 622.4

Multiply by 7/5 and add 31.

Input: 328
Correct: 490.2
✗ **GPT-4:** 457.6



- Example: Shift ciphers are a simple type of cipher. LLMs do much better at the most common shift cipher (13) than others.



5 Conclusion

- By considering the pressures that have shaped LLMs, we predicted that they would be highly probability-sensitive
- This prediction is supported across a range of tasks
- **High-level takeaway:** To understand what AI systems are, we must understand what we have trained them to be
  - This requires thinking about the training set and assessing how the AI system does/doesn't generalize beyond it!