Cognitive Effects in LLMs

A Critical Analysis for Impact on Pedagogical Alignment

Source: https://arxiv.org/abs/2308.14337

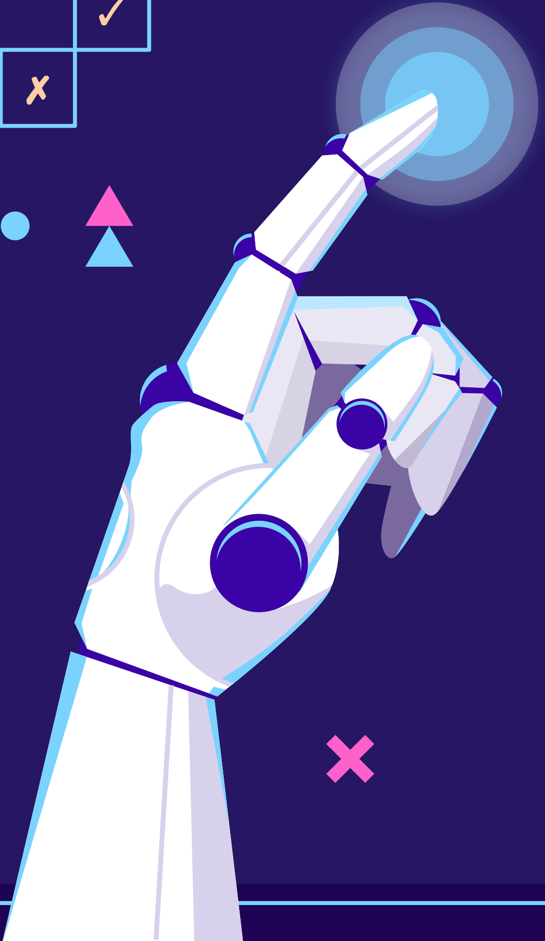
Executive Summary

GPT as LLM Agent display reasoning and decision-making abilities including biases and cognitive effects.

This paper explores 5 cognitive effects, expressions of systematic patterns, using essence of similar neuroscientific experiments.

Experiments to measure GPT's confidence at (specifically formatted) sequence processing tasks have yielded positive results for 4/5.

The hypothesis of the paper is the cognitive effects observed in humans are also present in the GPT albeit in a different descriptive theoretical essence.



Background

How LLMs work:

Compute PD over Tokens-Seq → Fed-Back for longer output
Similar Meaning Closer in Pre-Embedded Tokens Space
Transformers attention focuses on specific parts of sequence
Some generate several responses ranked by Confidence Interval of them

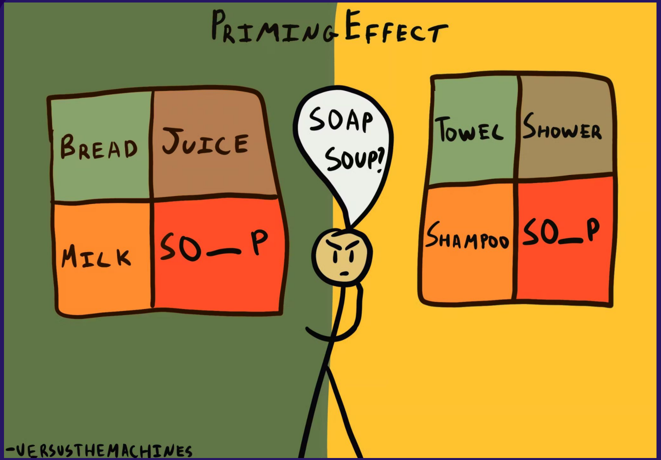
Why LLMs are Prone to Human Biases:

Possible biases and Hidden Goal Structures present or acquired
eg racism and sexism not filtered due to scale of training data

Cognitive-Effects:

Expressions of Systematic Patterns in cognitive processes

1. **Priming:** exposure to a stimulus influences the reaction to other
2. **Distance:** temporal and conceptual closeness is easier to perceive
3. **SNARC:** mental no. line (small left, big right) → spatial faster RTT
4. **Size-Congruity:** faster RTT when numerical & physical size congruent
5. **Anchoring:** perception is influenced by previous (anchoring) point



Experiments & Results

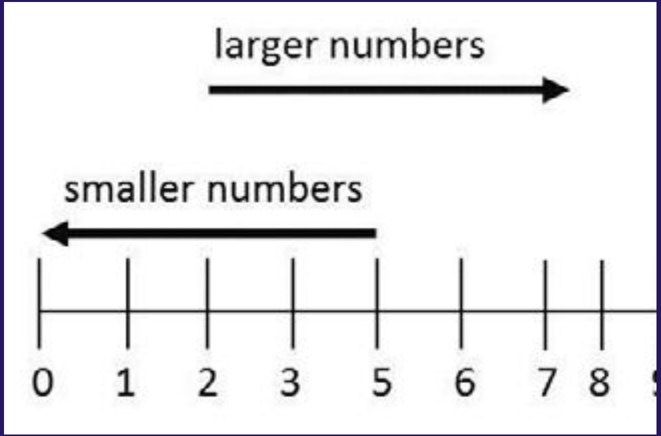
Priming Effect

GPT was given the prime stimulus (word) and asked if the letter sequence (prompt) can form (be recognized as a valid) word? (Measures CFT essence)

For example given “dog”, does “bone” form a word? Measure confidence.

Asked GPT in 3 format: Q/A, Sentence and Words only with varying spacing

Priming present in Q/A & Sentence formats but not in Words only format



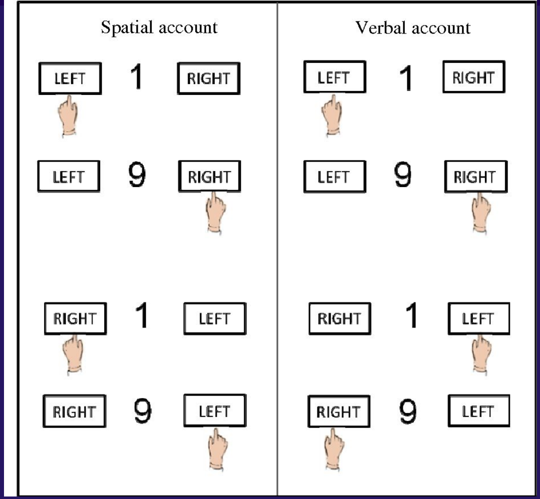
Distance Effect

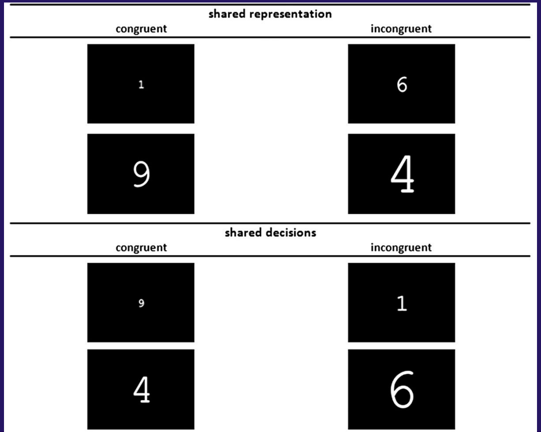
GPT was asked to compare between 2 sets of animals / numbers / months with animal names having same length for fairness and white spaces within

GPT was asked in 240 formats: is bigger / smaller than & singular & plural

Confidence was assigned to GPT’s answer of “Yes” or “No”

Distance showed more certainty in Number & Months than animals





Experiments & Results

SNARC Effect

GPT was given numbers and asked to classify them using directions.

“Magnitude” and “Parity” formats involved GPT responding, with left or right, for “compared to 5” and “even / odd”, respectively

“Vertical” format involved with GPT responding with up or down

SNARC present in Magnitude and Parity but not in Vertical format

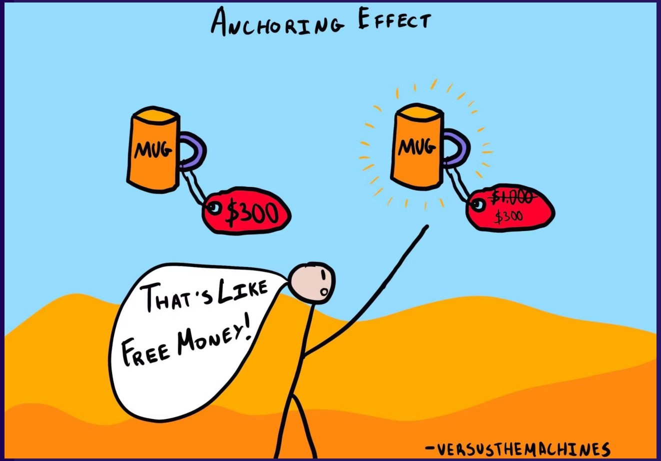
Size-Congruity Effect

GPT was given animals / numbers in upper / lower case and asked to respond with Yes / No

The question had formats as “Is first smaller / bigger than second” and vice versa

Size-Congruity was present for animals not for numbers

Experiments & Results



Anchoring

GPT was given a small / large anchor value and a sequence of tokens.

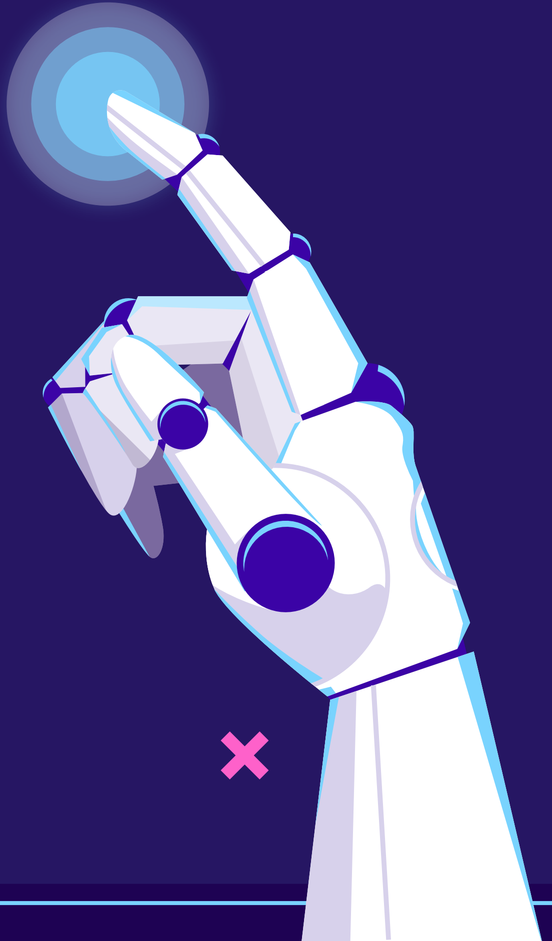
In the first experiment there was one anchor & in second there were two

The third & fourth experiments replaced earlier with anchored sequences

Anchoring was present in 3rd & 4th experiments but not first two

Thus they concluded the effect to not be present or even opposite to exist

Methodology



Why different methodology than that used for humans?

LLMs are:

1. Opaque
2. Sequence-based NOT Logic-based
3. Dependant on Temperature for Randomness

How different methodology than that used for humans?

1. Comparing Confidence of responses analog to Error instead of RTT
2. Confidence = Prob of Correct proportion to Prob assigned to all
3. Added mental load (whitespaces) when probability approaches 1
4. Simulated hands by limiting the response to ‘Left’ or ‘Right’
5. Reformatted prompts to compensate for only using a single LLM

Key Findings

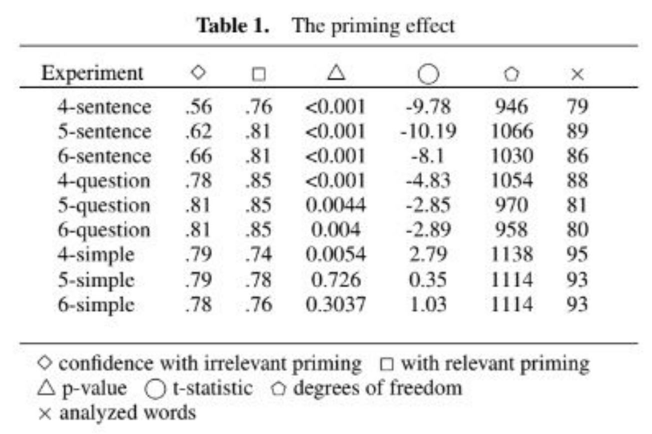


Table 1. The priming effect

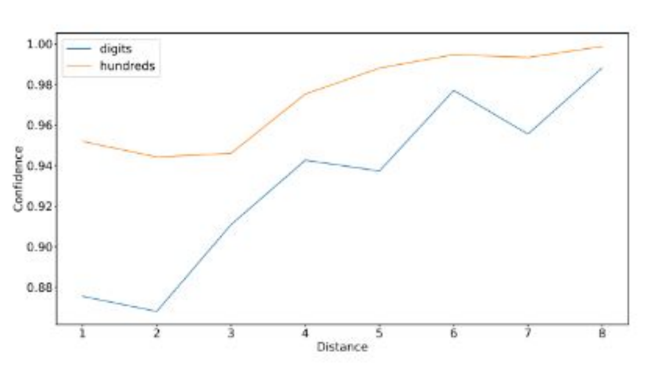
Experiment	◇	□	△	◯	⬠	×
4-sentence	.56	.76	<0.001	-9.78	946	79
5-sentence	.62	.81	<0.001	-10.19	1066	89
6-sentence	.66	.81	<0.001	-8.1	1030	86
4-question	.78	.85	<0.001	-4.83	1054	88
5-question	.81	.85	0.0044	-2.85	970	81
6-question	.81	.85	0.004	-2.89	958	80
4-simple	.79	.74	0.0054	2.79	1138	95
5-simple	.79	.78	0.726	0.35	1114	93
6-simple	.78	.76	0.3037	1.03	1114	93

◇ confidence with irrelevant priming □ with relevant priming
△ p-value ◯ t-statistic ⬠ degrees of freedom
× analyzed words

Priming Effect

The simple word format not having a priming effect could be attributed to GPT-3 making the cognitive operations used to comprehend content more accessible when associated words follow priming words.

Alternatively, they suggest that GPT-3 may simply handle associated words that follow previous words more frequently than unrelated words.



Distance Effect

The increase in distance increased the certainty but the greater pronunciation of the effect in the months and letters may be attributed to GPT processing various types of information differently on mental no. line

Authors rule out the possibility that the effect may be due to the distance between tokens in the embedding space after experimentation

Key Findings

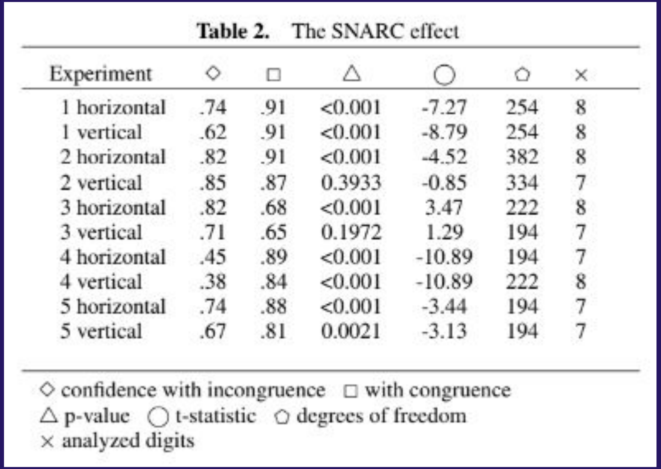


Table 2. The SNARC effect

Experiment	◇	□	△	○	⬠	×
1 horizontal	.74	.91	<0.001	-7.27	254	8
1 vertical	.62	.91	<0.001	-8.79	254	8
2 horizontal	.82	.91	<0.001	-4.52	382	8
2 vertical	.85	.87	0.3933	-0.85	334	7
3 horizontal	.82	.68	<0.001	3.47	222	8
3 vertical	.71	.65	0.1972	1.29	194	7
4 horizontal	.45	.89	<0.001	-10.89	194	7
4 vertical	.38	.84	<0.001	-10.89	222	8
5 horizontal	.74	.88	<0.001	-3.44	194	7
5 vertical	.67	.81	0.0021	-3.13	194	7

◇ confidence with incongruence □ with congruence
△ p-value ○ t-statistic ⬠ degrees of freedom
× analyzed digits

SNARC Effect

Authors suggest that the effect may be due to an association between left → bad & bad → small numbers AND right → good & good → big

The absence of the effect in Vertical experimentation and pronunciation of the effect with the inclusion of white-spaces may be attributed to the attention mechanism being more sensitive with natural presentation.

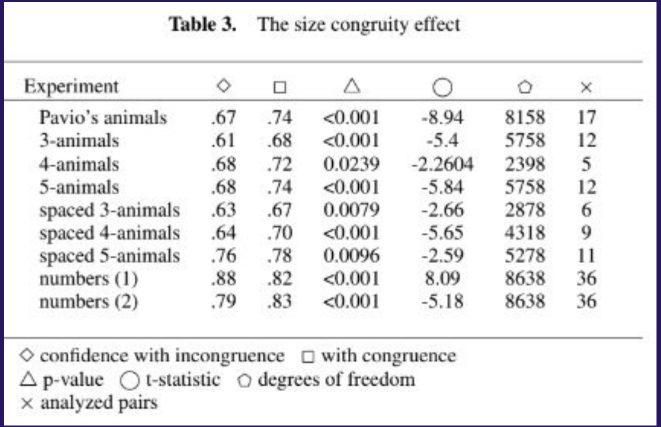


Table 3. The size congruity effect

Experiment	◇	□	△	○	⬠	×
Pavio's animals	.67	.74	<0.001	-8.94	8158	17
3-animals	.61	.68	<0.001	-5.4	5758	12
4-animals	.68	.72	0.0239	-2.2604	2398	5
5-animals	.68	.74	<0.001	-5.84	5758	12
spaced 3-animals	.63	.67	0.0079	-2.66	2878	6
spaced 4-animals	.64	.70	<0.001	-5.65	4318	9
spaced 5-animals	.76	.78	0.0096	-2.59	5278	11
numbers (1)	.88	.82	<0.001	8.09	8638	36
numbers (2)	.79	.83	<0.001	-5.18	8638	36

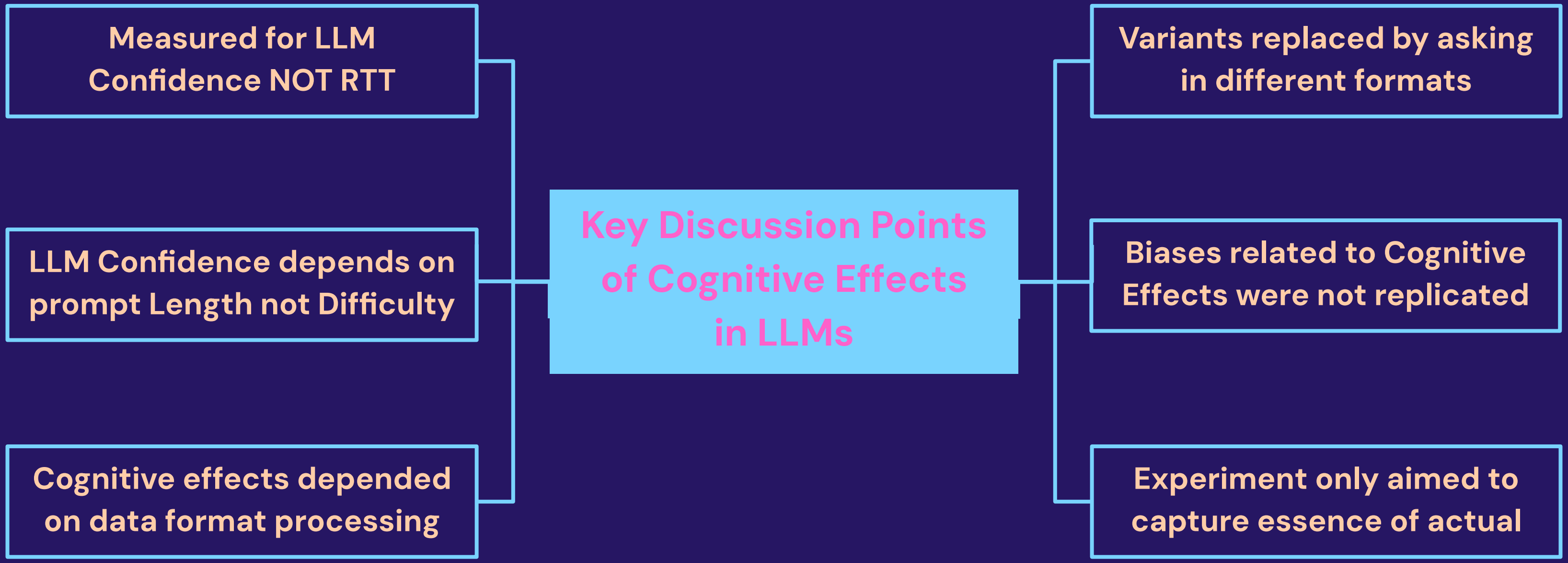
◇ confidence with incongruence □ with congruence
△ p-value ○ t-statistic ⬠ degrees of freedom
× analyzed pairs

Size-Congruity Effect

The authors suggest that the size congruity effect observed for numbers not animals could be explained by either the shared decisions / representation model

They note that it is surprising that GPT-3 treats letter capitalization so similarly to how people treat font sizes for numbers.

Key Discussion Points



Key Discussion Points of Cognitive Effects in LLMs

Measured for LLM Confidence NOT RTT

LLM Confidence depends on prompt Length not Difficulty

Cognitive effects depended on data format processing

Variants replaced by asking in different formats

Biases related to Cognitive Effects were not replicated

Experiment only aimed to capture essence of actual

Limitations & Open-Q.s

Limitations:

1. Reformatting the prompts do not capture essence of control group HENCE authors should have used variants of GPT-3

2. Lack of comparison with specific human specific data e.g.s for deeper insights

3. Lack of formal proofs eg using BMC for behavioural economics

4. Could have checked for impact of reversal curse on these effects

Open-Questions:

1. How can these cognitive effects be used when pedagogically aligning LLMs for teaching agents in EdTech?

2. What impact especially biases enforcement these cognitive effects have on GPT-3 model-based RL phase & even neurosymbolic-GNNs?

3. What guardrails can be applied prevent nudge of maliciousness?

4. How would they impact Dynamic Memory Mechanism of LDM^2?