Conditional Reasoning with Causation in Large Language Models

ARPE Defense

Samuel Debray

August 30, 2023

École Normale Supérieure Paris Saclay

What this talk will be about

Premise: If it rains, it take my umbrella with me.
Hypothesis: It doesn't rain.
Conclusion: Therefore, I don't take my umbrella with me.

Is this inference valid?

Premise: If Carol disiliones a ostrine, then the glotysoppeti is unamaned.
Hypothesis: Carol doesn't disilione a ostrine.
Conclusion: Therefore, the glotysoppeti isn't unamaned.

What about this one?

Outline

Description of the problem

First attempts using GPT-3

Conditional reasoning in OPT

- Accuracy and the like
- Comparison with humans

Concluding remarks

Description of the problem

The old debate of reasoning with conditionals

Endorsement rates:

MP	95% – 100%	$A \to B, A \models B$
MT	66%	$A \to B, \neg B \models \neg A$
DA	50%	$A \to B, \neg A \not\models \neg B$
AC	50%	$A \to B, B \not\models A$

Many people have proposed a theory, e.g.

- Denise Cummins (1995);
- Nick Chater and Mike Oaksford (lifelong);
- Philip Fernbach et al. (2010, 2013);
- Niels Skovgaard-Olsen et al. (2016, 2019, 2021);
- Katrin Schulz and Robert van Rooij (2021);
- etc.

But there is no consensus at all...

A new hope

2017: new class of Artificial Neural Network arises, the Transformer.

How can we use them?

- → We understand the architecture of models better than the architecture of the human brain.
- → model cannot handle causality $\Longrightarrow$ there is something else in the brain

Several articles tackle the connection between Transformer model and human cognition (at different levels: from psychology to brain imaging).

2022: Dasgupta et al. show that Transformer models show human-like content effects on reasoning

What about causal reasoning?

An interesting theory (among many others)

From Cummins et al. (1991) and Cummins (1995):

$$\text{Premise}, \text{Hypothesis} \models \text{Conclusion}?$$

Two important cues: number of alternative causes and number of disablers.

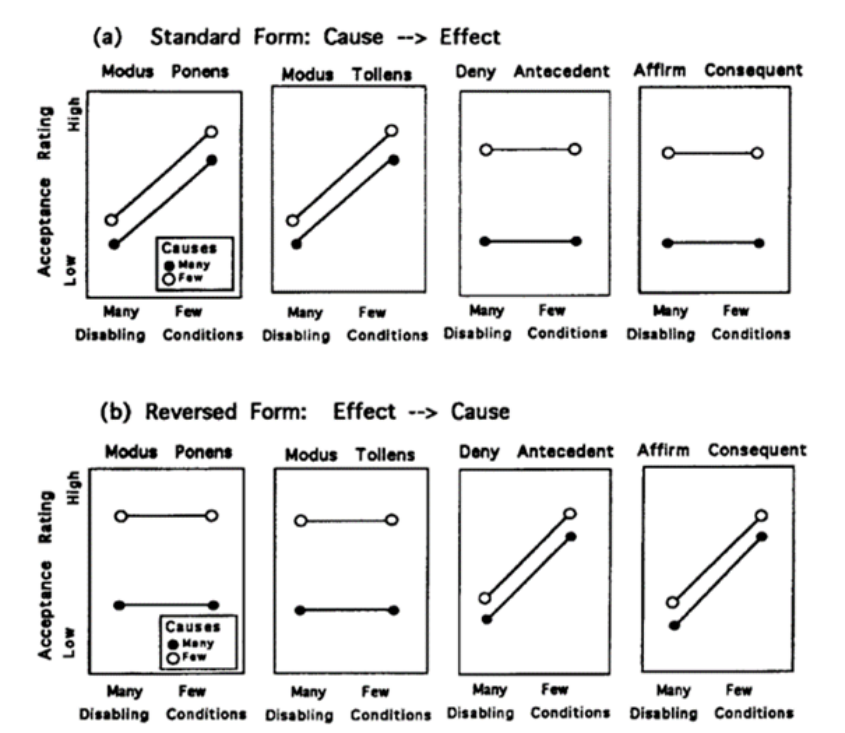


Figure 1: Cummins' hypothesis.

Empirically verified!

Aims:

(i) See whether LLMs are able to predict the validity of conditional inferences.

(ii) If not, do they have the same biases as humans?

First attempts using GPT-3

Replication of Cummins' results

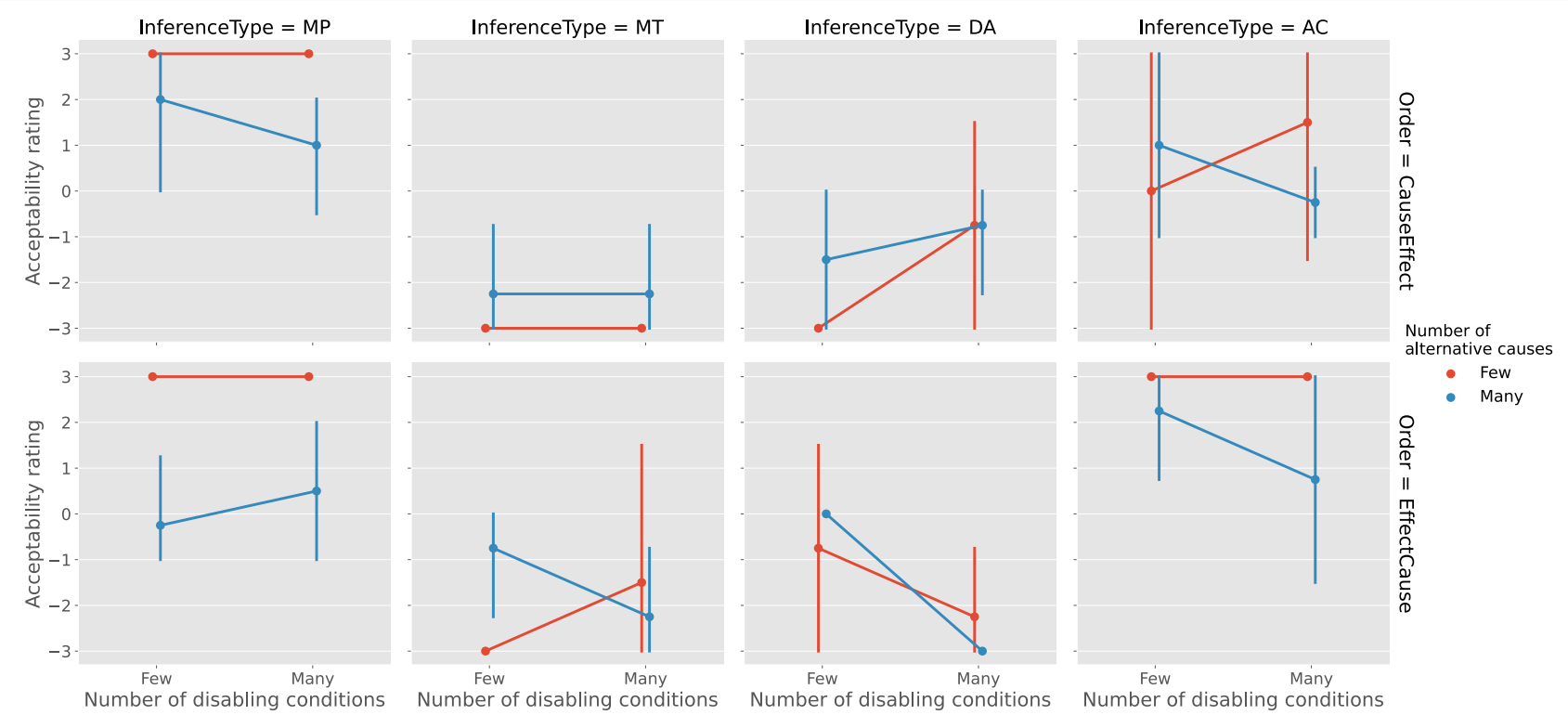


Figure 2: Cummins-like plot with GPT-3.

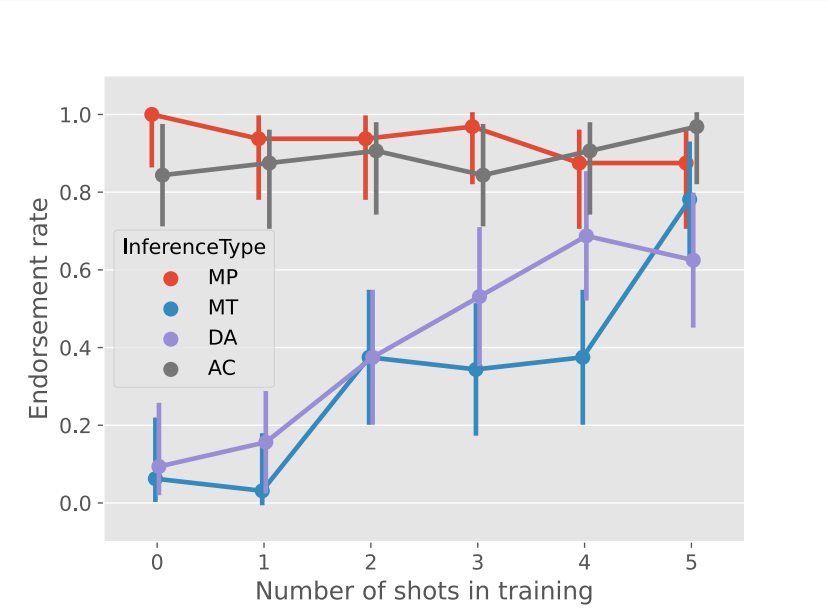
We also tried to improve performance with few-shot learning, but it did not help.

What is actually going on?

Few-shot training

Showing examples with solution in the prompt on the fly (no additional training).

Has been proved to improve GPT-3's performance.



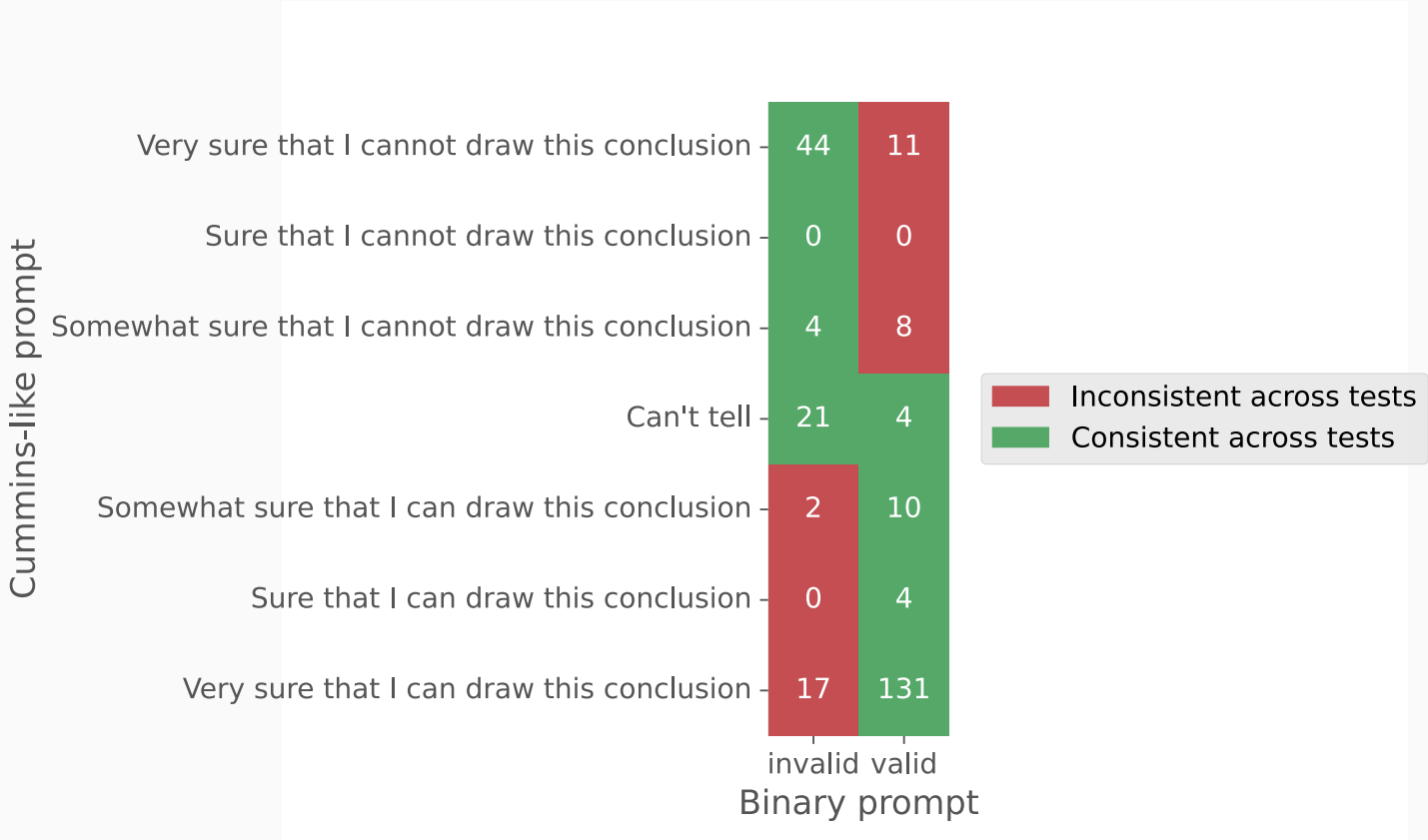
Two explanations:

- purely syntactic (pessimistic);
- model good at forward reasoning and bad at backward (optimistic).

Figure 3: Effect of few-shot learning

Main issues

- Dataset too small (wide CI)
- GPT-3 is too obscure: we can't see what goes on inside...
- Sensitivity to prompting!!! (disagree in 14.8% of cases)



Conditional reasoning in OPT

Where are we now?

We could not probe Cummins' hypothesis in GPT-3.

We redo the same with:

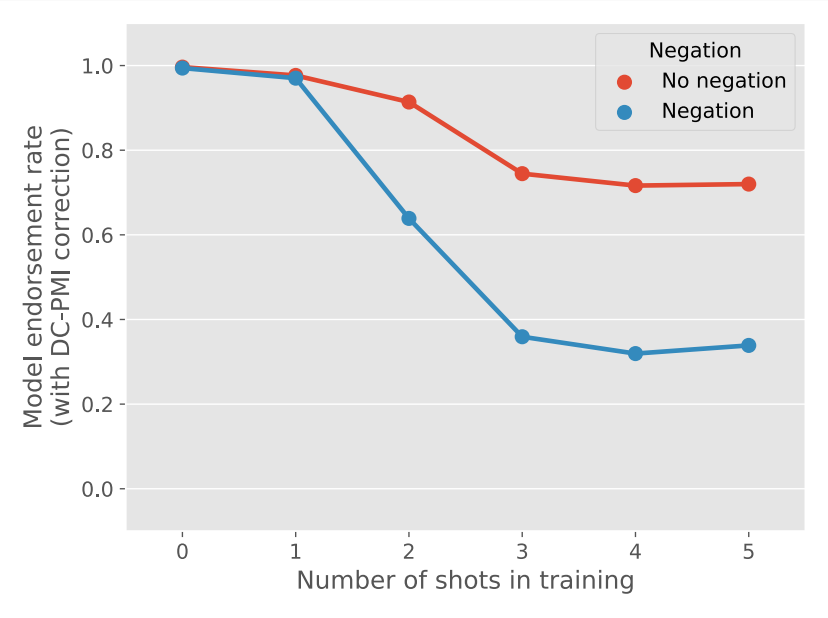
- a larger dataset;
- different causal relations between the antecedent and the consequent (causal, independent, jabberwocky);
- different prompts, over which we average;
- human RTs and endorsement rates for comparison.

Conditional reasoning in OPT

Accuracy and the like

Do we have the same issues as with GPT-3?

Well, actually it's quite the opposite...



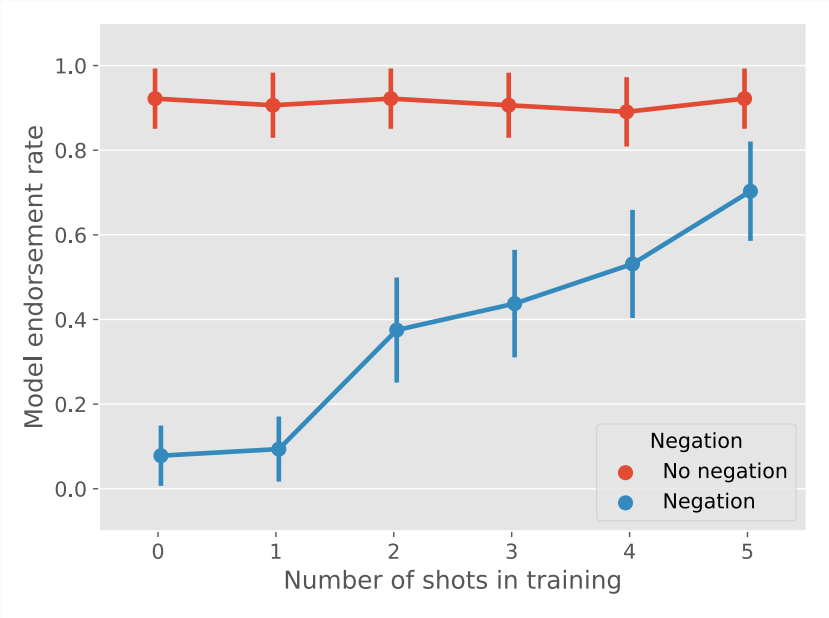
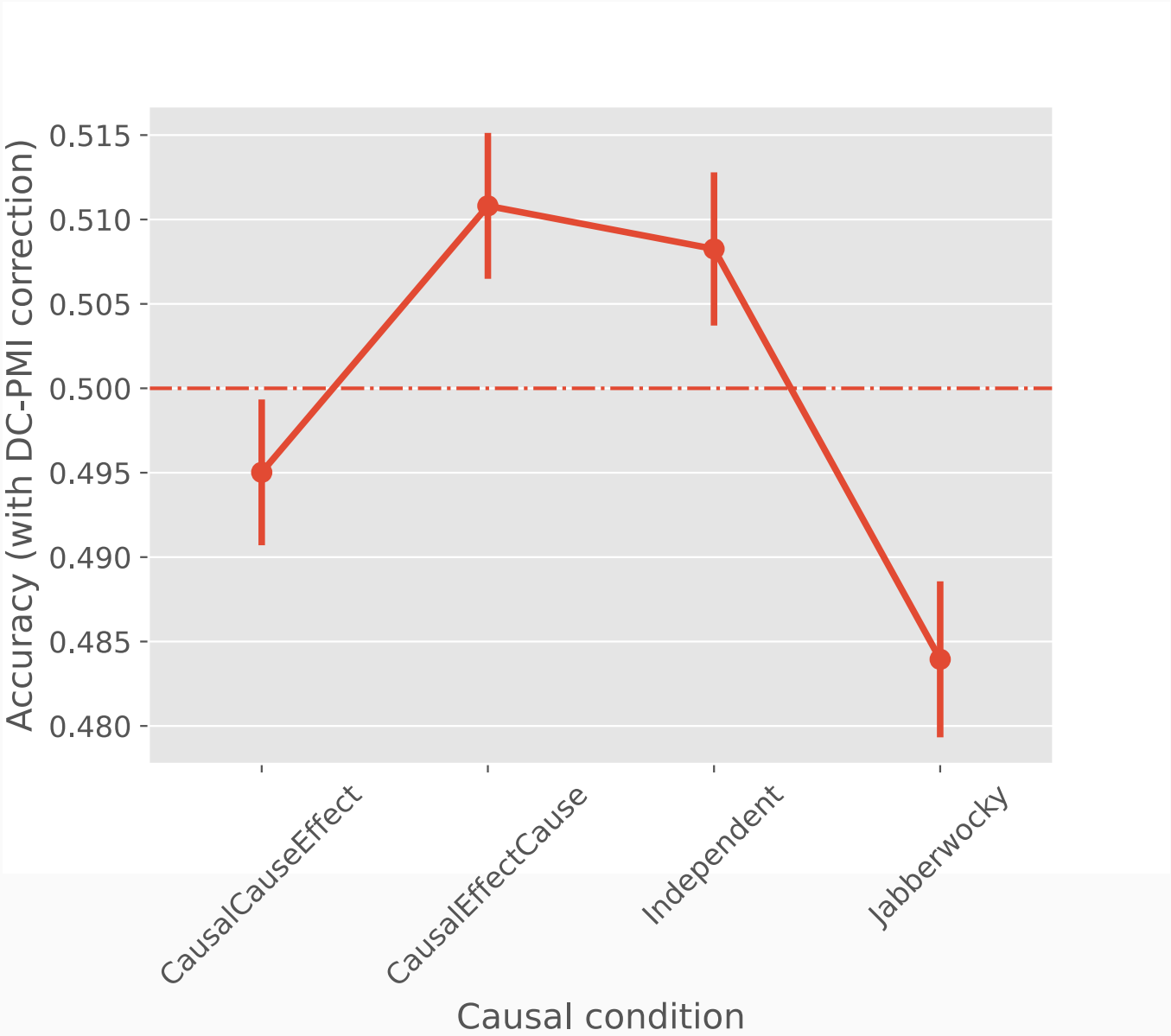


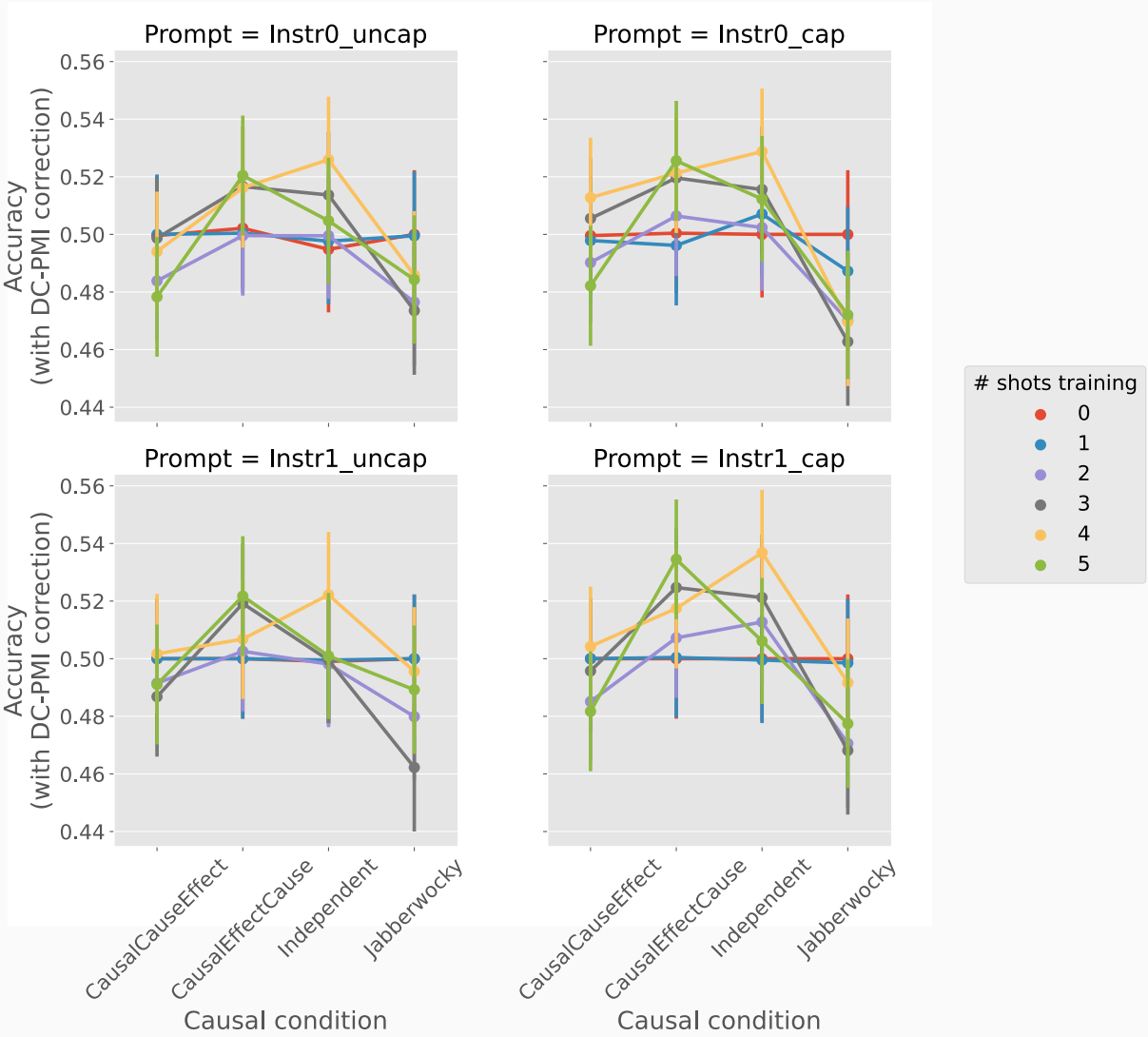
Figure 4: Effect of few shot training and negation

OPT performs at chance to predict validity...



OPT performs at chance to predict validity...(cont.)

And few-shot training or instructions do not help:



Wait, what is a Transformer model?

What we know so far

A type of architecture of artifical neural networks.

We used two different models: GPT-3 (from Open AI) and OPT with 66B parameters (from Meta).

$$\begin{array}{lcccl} \text{Input} & \xrightarrow{\text{Model}} & \text{Logits} & \xrightarrow{\mathcal{L}:\, x \mapsto -\log(x)} & \text{NLL} \\ x_1 & & \Pr_\theta(x_1) & & -\log(\Pr_\theta(x_1)) \quad := \text{NLL}(x_1) \\ x_2 & & \Pr_\theta(x_2|x_1) & & -\log(\Pr_\theta(x_2|x_1)) \quad := \text{NLL}(x_2) \\ \vdots & & \vdots & & \vdots \\ x_t & & \Pr_\theta(x_t|x_{<t}) & & -\log(\Pr_\theta(x_1|x_{<t})) \quad := \text{NLL}(x_t) \end{array}$$

The output score is

$$\text{output}(X) := \frac{1}{t}\sum_{i=1}^{t} \text{NLL}(x_i), \qquad X = x_1 \cdots x_t$$

it captures the model's surprisal on the input statement.

Something happens!

Another measure:

$$\text{certainty}(X) = |\,\text{NLL}(X_{\text{yes}}) - \text{NLL}(X_{\text{no}})| = \left|\log\left(\frac{\Pr(X_{\text{yes}})}{\Pr(X_{\text{no}})}\right)\right|.$$

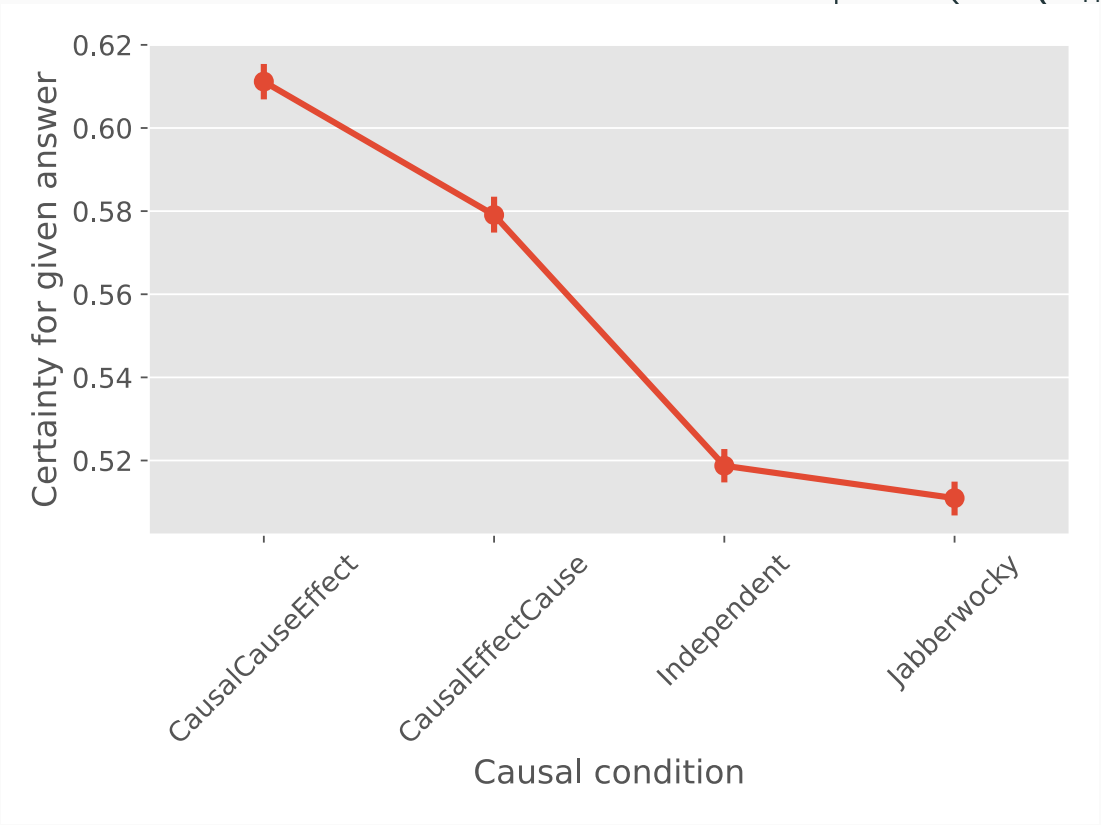
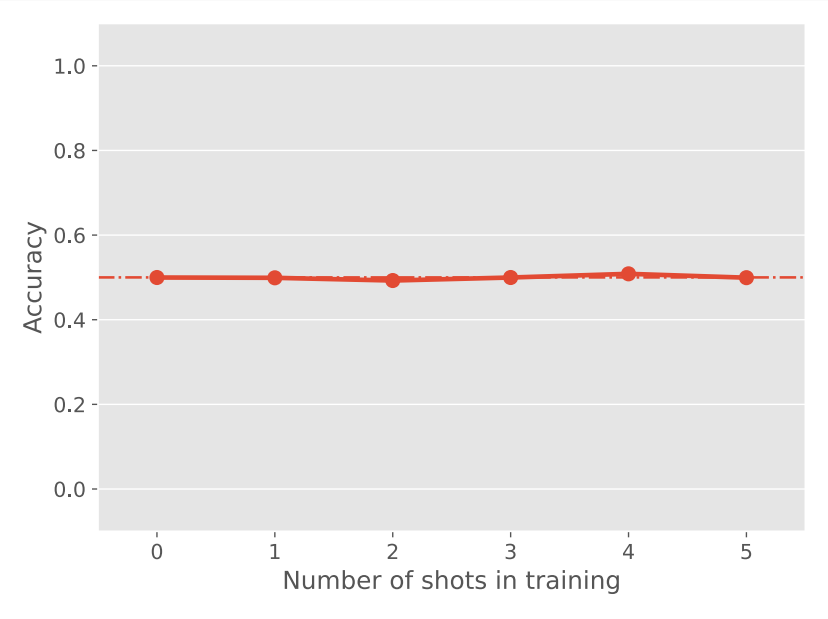


Figure 5: Effect of causal condition on certainty

All pair-wise differences significant!

Effect of prompting and few-shot learning



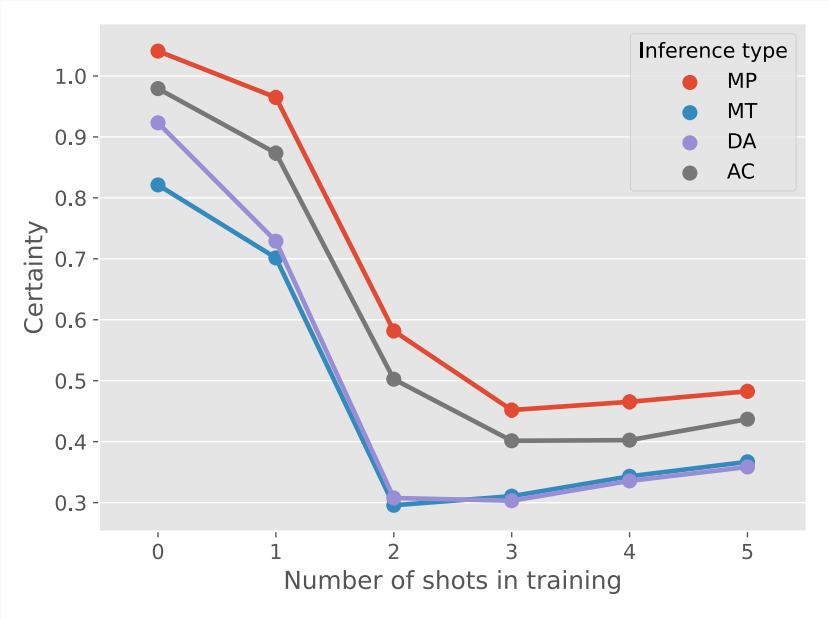


Figure 6: Effect of few-shot learning on performance

→ No effect of FSL on accuracy...

→ ...but certainty keeps evolving

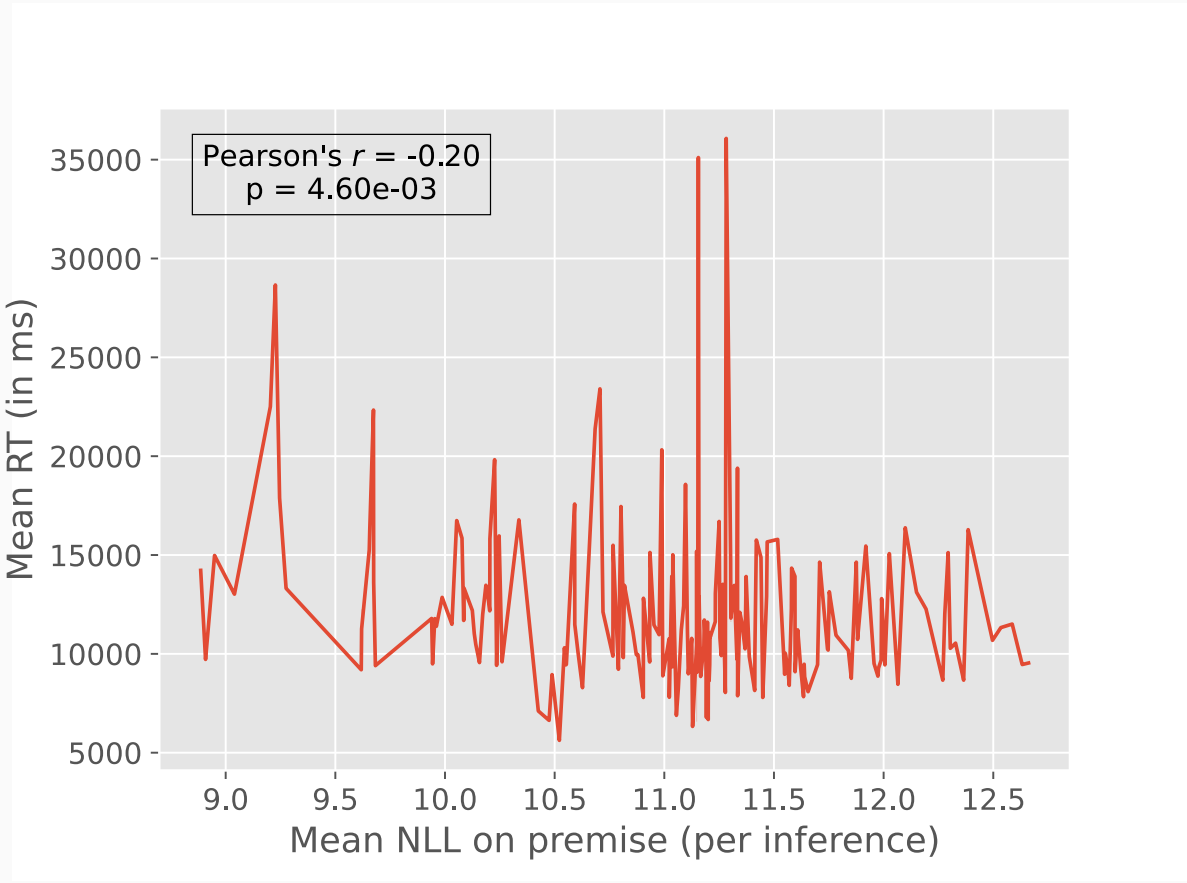
So the model must be picking up on something!

Conditional reasoning in OPT

Comparison with humans

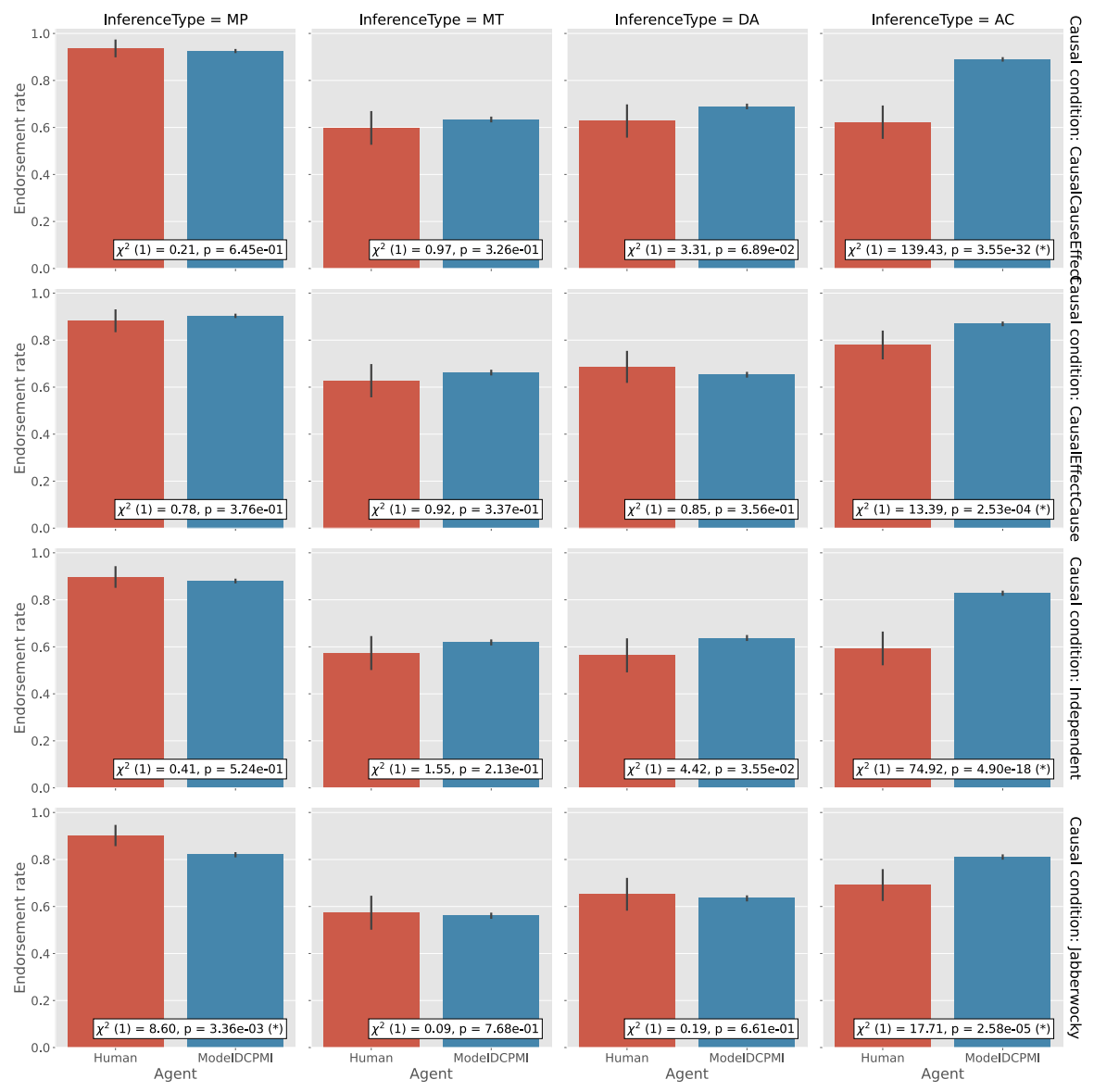
Correlation between RT and NLL

Weak correlation between mean RT and NLL on premise:



But the correlation is negative!

OPT actually shows human-like content effects



Same endorsement rates as human except:

- for all AC inferences
- for jabberwocky MP

Concluding remarks

Take home messages

- Huge differences between models, prompt settings, training, etc. Everything must be taken with a pinch of salt.
- OPT and GPT-3 perform at chance to predict validity of conditional inferences.
- However, OPT has same endorsement rates as humans (most of the times).
- OPT's certainty is very sensitive to causal condition: maybe the model gets a tiny sense of it?

And now what?

An option: use all the AI machinery (fine-tuning, bigger models, fancy analyses) to make these results better?

No!

Our results show that, a priori, the architecture of off-the-shelf Transformers does not allow it to perform conditional reasoning, no matter the causal condition.

But we could dig further to understand:

- the pattern with negation and few-shot learning,
- why the endorsement rates for AC and MP jabberwocky are different.

Questions?