Conditional Reasoning in Large Language Models

Internship Thesis

submitted in partial fulfillment of the requirements for the validation of

Année de Recherche Prédoctorale à l'Étranger

***

Diplôme de l'École Normale Supérieure Paris-Saclay

conducted at the *Institute for Logic, Language and Computation*, Universiteit van Amsterdam

---

Written by

Samuel Debray

Département Informatique
École Normale Supérieure Paris-Saclay

Under the supervision of

Dr. Katrin Schulz
and
Prof. Dr. Robert van Rooij

Formal Semantics and Philosophical Logic Unit
Institute for Logic, Language and Computation
Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Universiteit van Amsterdam

Year 2022 – 2023

ABSTRACT

Reasoning with conditionals and causation is a key ability for humans: it allows them to think about alternative scenarios, and it is therefore necessary to make decisions. However, there is no consensusal account of the reasoning patterns that have been reported in the literature. Building on recent advances in artificial intelligence, we propose a new way of adressing the issue of reasoning with causal conditionals by assessing the abilities of large language models like GPT-3 and Meta AI's OPT with 66 billion parameters. We crowd-sourced a large dataset of conditionals involving different kinds of causal relations between the antecedent A and the consequent B (namely A causes B, B causes A, A and B are independent, and A and B are jabberwocky sentences). We first used this dataset to verify, during an online behavioural experiment, that the causal relation between A and B has a strong effect on human reaction time when judging conditional inferences as valid or invalid. We then showed that, although OPT cannot accurately predict the validity of such conditional inferences, it still gets a sense of the causal relation between A and B. Finally, we also found out that, for most causal relations, human and OPT's endorsement rates are similar, and that human and model's cognition can therefore be connected.

Keywords. conditional reasoning, causality, reasoning biases, large language models, NLP

ACKNOWLEDGMENTS

First of all, I would like to thank my advisors, Katrin Schulz and Robert van Rooij. They both have helped me a lot to navigate through the dense literature of conditionals, and have given me the freedom and resources to work on a topic that really interested me.

I would also like to thank all the PhD students with whom I have shared enlightening discussions, and who have given me very precious advice as to how to run my project properly. To think regard, many thanks to Jaap Jumelet, Michael Hanna, Oskar van der Wal and Alina Leidinger.

Thank you to my office mates: Fei Xue, Niccolò Rossi, Tanguy Bozec, Raoul Koudijs. A very special thanks goes to Niccolò, for having made my first few months in Amsterdam a lot merrier than they would have been without him.

Thank you also to the ILLC community broadly speaking. I want to thank in particular Peter van Ormondt, Ewout Arends, Alex Zieglerová and Roos Bouwdewijn, who try to make this weird lab, which is split into very buldings and very different teams, a nice place to work in. Thank you also to all the people with whom I have shared lunch, coffes and nice chats over the course of the year: Dean McHugh, Levin Hornischer, Émile Enguehard, Tom Roberts and Marianna Girlando.

Finally, thank you to all the amazing people – who have become close friends – I have met in KW 18: Sana, Haley, Holy, Tobia, Aya, Nicola, Vincenzo, Andrew and Bernie. I can't wait to see you again, in Amsterdam or elsewhere.

CONTENTS

DECLARATION OF CONTRIBUTION

Definition of the resarch questions. Katrin Schulz and myself

Literature review. myself, with guidance from Katrin Schulz and Robert van Rooij

Choice of methodology. myself, with advice from PhD students (esp. Jaap Jumelet and Michael Hanna)

Design of the experiments. Katrin Schulz and myself

Review of the participations to the experiments. myself, with the help of Andrew McIntyre

Programming. myself

Data analysis. myself

Interpretation of the results. Katrin Schulz and myself

Redaction of the thesis. myself

Proof-reading of the thesis. Katrin Schulz

CHAPTER 1

INTRODUCTION

1.1 Motivations

In the last few years there has been an increasing interest in assessing large language models' (LLM) reasoning skills (Binz & Schulz, 2023; Dasgupta et al., 2022; Helwe et al., 2021) and comparing them to human performance, both from the behavioural (Arehalli et al., 2022; Wilcox et al., 2020, 2021) and computational (Caucheteux & King, 2022; Caucheteux et al., 2021a, 2021b, 2022, 2023; Millet et al., 2022; Pasquiou et al., 2022, 2023; Schrimpf et al., 2021) points of view. Our project is in line with this direction and focuses on reasoning with conditionals.

Reasoning with conditionals is a central focus in various fields such as philosophy, linguistics or psychology and has been studied extensively since the 1970s (see for instance Politzer, 2007; Stenning & van Lambalgen, 2008; van Rooij & Schulz, 2019 for a review). Still, the way humans reason with conditionals remains a widely open question. We believe studying from the point of view of NLP might give us some insight both on the way machines and humans handle this type of reasoning.

It is known from the literature (Cummins et al., 1991; Cummins, 1995; Fernbach & Darlow, 2010; Fernbach & Erb, 2013; Skovgaard-Olsen et al., 2021) that humans show content effects when reasoning with conditions. Our goal is to investigate whether these effects can also be found in machines. More precisely, we want to probe whether large language models like GPT-3 (Brown et al., 2020) or OPT with 66B parameters (Zhang et al., 2022) show the same effects, or if we can train them to predict them from human behavioural data. We believe that answering these questions could help better seize how well humans and machines are aligned. This could also help better understand conditional reasoning in humans by investigating where and when exactly these effects are learnt by models.

More specifically, we shall try to answer the two following questions.

(i) Are GPT-3 and OPT able to predict the logical validity or invalidity of Modus Ponens (MP), Modus Tollens (MT), Denial of the Antecedent (DA) and Affirmation of the Consequent (AC) when presented with conditionals of the form "If A, then B"? Do their performances depend on the relation between A and B or the training (few-shot, see Brown et al., 2020) of the models?

(ii) Are the models able to predict the endorsement rate and reaction time of humans presented with such stimuli? Again, does the performance depend on the relation between A and B or the type of inference?

1.2 Literature review

In this section, we review two strands of the literature. In section 1.2.1, we try to summarise the rich theory of human reasoning with conditionals, which has been developed since the 1970s. We then turn to the recent progress made in artificial intelligence, we review works focusing on Language Models' ability to reason (section 1.2.2) and to predict human data (section 1.2.3). We focus more specifically on the Transformer architecture (see Vaswani et al., 2017 and appendix A.1), which is responsible for

the recent breakthoughts in the field of AI.

1.2.1 The psychology of reasoning with conditionals

It would be foolish to attempt to give here a review of everything that has been going on in the field of conditional reasoning since the 1970s. We will, however, try to give an overview of some theories that seem to have had a large influence on the literature. Again, we do not pretend to cover all the influential theories, but only those that emerged from the subset of the literature that we have read.

The main theories of reasoning with conditionals

Broadly speaking, one can make a distinction between two kinds of conditionals:

(1) a. If you drop this glass, it will break.
b. If Peter got your message, then he will be here this afternoon.

(2) a. If you were in Paris next week, we could meet.
b. If you had prepared for the exam, you would have passed.

(1) is an example of *indicative conditionals* and (2) of *counterfactuals*. Indiactive conditionals are distinguished from counterfactuals by not having one of the modal verbs *would*, *could*, *should*, *ought* as finite verb in the consequent. Of course, this classification is rough are subcategories can be found within each categories, for example, Haegeman (2003) argues for a distinction between event-conditionals (like (1)) and *premise conditionals* (like (3-a)). Similarily, Austin (1956) differentiates the so-called *biscuit conditionals* (like (3-b)) from the others.

(3) a. If you're so clever, why don't you do this problem on your own? (Haegeman, 2003)
b. There are biscuits on the sideboard if you want them. (Austin, 1956)

As for counterfactuals, they typically have a false antecendent and are phrased in the subjunctive mood.

In the literature, indicative conditionals and counterfactuals are often treated separately. To illustrate the apparent semantic difference, philosophers and linguists often use the following example:

(4) a. If Oswald hadn't shot Kennedy, then someone else would have.
b. If Oswald didn't shoot Kennedy, then some else has.
(Adams, 1970)

It is considered that sentence (4-a) is false, at least if you do not believe in conspiracy theory. However, sentence (4-b) is true, as there is someone somewhere who killed Kennedy, whether it be Oswald or whoever else. In this section, we will therefore examine, turn by turn, indicative conditionals and counterfactuals.

Counterfactuals. The first theories of reasoning with counterfactuals date back to the works of Robert Stalnaker and David Lewis (Lewis, 1973b; Stalnaker, 1968). It had been known for long that neither the material (hereafter denoted by $\rightarrow$) nor the strict[1] (hereafter denoted by $\Rrightarrow$) implications from standard logic were good models of counterfactual conditionals. For example, it is well known that the material and strict implications allow, in all systems, to strengthen the antecedent:

$$\phi \rightarrow \psi \vDash (\phi \wedge \chi) \rightarrow \psi \qquad \text{and} \qquad \phi \Rrightarrow \psi \vDash (\phi \wedge \chi) \Rrightarrow \psi.$$

However, counterfactuals do obey the stengthening the antecedent rule, as shown by sentences (5).

(5) a. If I had struck the match, it would have lit.
b. If I had struck the match and had been in an oxygen-free environment, it would have lit.

[1]The strict implication is defined in modal logic as $\phi \Rrightarrow \psi := \Box(\phi \rightarrow \psi)$.

In example (5), the sentence with weak antecedent (5-a) is true, but that with strong antecedent (5-b) is false. There is an abundant literature about finding a proper semantics for the natural language counterfactual operator $\Box\!\!\rightarrow$, but everyone seems to agree that it can be represented in a standard modal logic. The most popular theory (Kratzer, 1981; Lewis, 1973b; Pollock, 1976; Stalnaker, 1968) is based on possible world semantics and Kripke models, which allow for the strict implication to quantify only on *accessible* worlds. The main debate is about what kind of accessbility (or *similarity*) relation is best suited for the system to best model natural language counterfactual implication. For instance, Stalnaker (1968) supports the *uniqueness assumption* – that is that there should be a unique most similar world to every world – while Lewis (1973b) argues against it.

The notion of causality is closely related to that of counterfactual implication. Interestingly, two points of view are confronted. Lewis (1973a), on the one hand, derives causality from the very notion of counterfactuals (which is itself grounded in similarity): ϕ causes ψ iff $\phi \,\Box\!\!\rightarrow \psi$ and $\neg\phi \,\Box\!\!\rightarrow \neg\psi$. Balke and Pearl (1994) interpret conditionals in terms of causation and intervention: given a causal model $\langle B, E, F\rangle$, the counterfactual "If X had been x, then Y would have obtained value y." is defined, for $X, Y \in E$ two endogenous variables, as meaning that y is the solution for Y to the equations obtained from F by replacing all functions corresponding to member sets of X by the constant function $X = x$ (that is, y is the potential response to action $\text{do}(X = x)$).

Indicative conditionals[2]. At first glance, material and strict conditionals do not seem to be as unsuitable to model natural language indicative conditionals as they are for counterfactuals, and one could think that pragmatics could help fill the gaps and explain away standard counterexamples. However, Grice (1989) provided the following example: Yog and Zog play chess and Yog has black nine out of ten times. We don't know who won what game, but we do know that there are no draws, and that of the hundred games they played up to now, Yog won eighty times when he had black and lost all ten times that he had white. Then the following sentences are intuitively true:

(6) a. If Yog had black, then there is a probability of 8/9 that he won.
b. If Yog didn't win, then there is a probability of 1/2 that he didn't have black.
(Grice, 1989)

This example challenges the idea that the indicative conditional implication can be modeled using either the material or the strict implication and, in fact, it even questions the very idea that it can be expresses a logic proposition. Adams (1965) suggests that conditionals should be represented in terms of conditional probability rather than proposition. Stalnaker (1970) proposes that $\Pr(A > B) = \Pr(B \mid A)$ where $>$ denotes the indicative conditional. However, Lewis's triviality result (Lewis, 1976) proves him wrong and raises again the question to know whether indicative conditional actually expresses a proposition.

Among the proponents of the non-propositional account, a few authors argue for a full probabilistic account (Adams, 1996; Oaksford & Chater, 2007; Stevenson & Over, 1995). Some other authors, propose that, in everyday reasoning, the interpretation of conditionals is grounded in the reasoner's knowledge to a much more considerable extent than what is considered in the impoverished contexts used to study formal reasoning. This view is a trade-off between the propositional and probabilistic accounts, and involves causality. Following Cheng (1997), many theories have been proposed to interpret the assertability of a condition "If ϕ, then ψ." in terms of the *causal power* of ϕ to produce ψ $p_{\phi,\psi}$ (Skovgaard-Olsen et al., 2021; van Rooij & Schulz, 2019). Another offshot of Cheng's theory is that of Cummins et al. (1991) and Cummins (1995) (also defended by Fernbach & Darlow, 2010; Fernbach & Erb, 2013) described in section 2.1.

Canonical conditional reasoning tasks

In this section, we briefly review the main tasks that psychologists have designed to assess conditional reasoning in humans. See Politzer (2007) for a more comprehensive review.

[2]For a comprehensive review, see van Rooij and Schulz (2020).

Wason's selection task. This is probably the most popular task (for a (very) detailed discussion, see Stenning & van Lambalgen, 2008, chap. 3). Participants are shown four cards and are given a rule. For example, the cards could be

A	K	3	7

and the rule "*If there is an A on one side, then there is a 3 on the other side.*". People are then asked what card they should turn over in order to verify whether the rule is true, the number of card they turn over should also be minimal. While a material implication account would predict that people should turn cards [A] and [7] over, only 5% of participants typically do so, and 35% only want to turn [A] over and 45% want to turn [A] and [3].

Conditional reasoning task. In this task, participants are shown a conditional "If ϕ, then ψ." and are asked to judge which of the following inferences are valid:

Modus Ponens (MP) $\phi \vDash \psi$;

Modus Tollens (MT) $\neg\psi \vDash \neg\psi$;

Affirmation of the Consequent (AC) $\psi \vDash \psi$;

Denial of the Antecedent (DA) $\neg\phi \vDash \neg\psi$.

The material implication account would predict that people endorse MP and MT and reject AC and DA. However, MP is endorsed by 90% of participants, MT by 70% and DA and AC only by 50% (chance level). This is the task we will focus on in our study of conditional reasoning in humans and machines.

Truth table task. This task is the least used and dates back to the 1950s. Participants are shown a condition "If ϕ, then ψ." and then asked use combinations of $\{\phi, \psi, \neg\phi, \neg\psi\}$ to construct situations in which the statement is true or false. The overwhelming majority of participants chose $\phi \wedge \psi$ to make the statement true, and $\psi \wedge \neg\phi$ to make it false. $\neg\phi$ is seldom chosen. Thus, participants seem to consider the $\neg\phi$ case as irrelevant to the truth value of conditional statements, which challenges the material account.

Assertability task. Contrary to the others, this task focuses on assertability rather than truth value. It is useful for instance to probe the theories involving causal power (see section 1.2.1 and, for instance, Skovgaard-Olsen et al., 2021). Participants are presented with "If ϕ, the ψ." conditionals, possibly in context, and are then asked to estimate the probability of different events (e.g. $\Pr(\psi \mid \phi)$), in order to derive various quantities of interest like the causal power.

1.2.2 Language models' reasoning abilities

In the last couple of years, researchers have become interested in language models' ability to hold a real argument. It is a fact that these models are becoming more and more efficient in formal linguistic tasks, it has been shown for instance that they are able to learn hierarchical structure (Manning et al., 2020) and abstraction (Tenney et al., 2019). However, many studies show that these models fail at tasks involving more than mere formal linguistic abilities, like commonsense or formal reasoning (Binz & Schulz, 2023; Dasgupta et al., 2022; Helwe et al., 2021). For instance, here is an example of a discussion with GPT-3 (Brown et al., 2020) reported by Collins et al. (2022) to illustrate that language models only seem to be thinking because they are able to learn word co-occurrence patterns and can be misled by out-of-distribution queries:

> **Prompt:** Get your sofa onto the roof of your house, without using a pulley, a ladder, a crane...
> **GPT-3 response:** Cut the bottom of the sofa so that it would fit through the window... break the window to make room for the sofa.

Another example of failure in reasoning in out-of-distribution contexts is provided by Kassner et al. (2020) with the use of negation. They point out that BERT (Devlin et al., 2019) answers correctly at the following LAMA (Petroni et al., 2019) question:

> **Prompt:** Marcel Oopa died in the city of [MASK].
> **BERT response:** Paris
> **Expected response:** Paris

however, it delivers the exact same result when the word "not" was added by Kassner et al.:

> **Prompt:** Marcel Oopa did not die in the city of [MASK].
> **BERT response:** Paris
> **Expected response:** any city but Paris

Helwe et al. (2021) and Mahowald et al. (2023) provide a good review of what Transformer-based model are able and unable to do at the moment.

In addition, some articles use tools and tasks developed by psychologists to assess LLMs' abilities. Binz and Schulz (2023), for instance, assessed GPT-3 on several skills – namely decision-making, information search, deliberation, and causal reasoning – using common tests from the cognitive psychology literature. They showed that, while GPT-3 performs extremely well (sometimes better than humans) at some tasks, like multiarmed bandit or vignette-based tasks, it also fails at some others, like causal reasoning tasks. This highlights both the successes of AI, and the fact that language models are still far form human-like performance. Similarily, Dasgupta et al. (2022) evaluated DeepMind's Chinchilla (Hoffmann et al., 2022) on Wason's selection task (see section 1.2.1) and on syllogisms and found that models, just like humans, are better at handling realistic than abstract situations.

In the same spirit, some authors (e.g. Mahowald et al., 2023; McClelland et al., 2020; Nye et al., 2021) propose to adapt the architecture and training (both data and objective function) of LLMs to mimick the human mind, hoping that this would allow such models to approach human-like performance on reasoning tasks. For example, Mahowald et al. (2023) suggest a distinction between formal linguistic abilities, at which LLMs are very good, and functional linguistic abilities, which require additional skills (e.g. pragmatics or commonsense reasoning). They claim that, since both abilities ellicit very different networks in the brain, they should rely on two distinct systems in LLMs. They argue that language models would benefit from having a second system dedictated to functional skills, and propose that this system be either hardcoded in the architecture of the models, or that it be incided throught the training process.

On another note, Transformer models have been shown to be capable of learning from example and perform advanced mathematical tasks such as inverting and diagonalising matrices (Charton, 2021), solving ODE (Lample & Charton, 2019) or predicting qualitative properties of differential systems (Charton et al., 2021). Furthermore, even though the models do not achieve 100% accuracy, Charton (2022) shows that the model still manages to extract many information from its training and that it is still possible to make sense of its answers, which are not absurd hallucinations like it is sometimes argued.

1.2.3 Languge models to predict human data

In the literature, many articles present ways to use LLMs to predict human data. The data can be of two kinds: behavioural (e.g. reaction time) or physiological (e.g. fMRI time series). In this section, we review these two trends.

Prediction behavioural data. Wilcox et al. (2020) proved that modern language models are good predictors of humans' reading time. More precisely they showed that GPT-2's negative log-likelihood is strongly positively correlated to word reading time from several datasets. In addition, they reviewed several models, and put forward a positive correlation between a model's ability to predict the next word of a masked input, and its psychometric predictive power. This score, which they refer to as ΔLogLik, is computed using 10-fold cross-validation on the whole dataset and measures a model's

ability to predict psychometric data (like eye-movement of self-paced reading time) by computing the average over words of the input of the log difference of the predicted and measured data. In a follow-up article, Wilcox et al. (2021) tried to account for magnitude effects but concluded that LLMs systematically underestimate the slowdown in reading pace due to ungrammatical sentences (van Schijndel & Linzen, 2021, see also). Arehalli et al. (2022) inverstigated further the magnitude effect and confirmed that, even when the weights of lexical and syntactic predictability and estimated independently, LLMs still understimate the slowdown in reading pace due to syntactic ambiguities.

Predicting physiological data. Schrimpf et al. (2021) studied how well LLMs can predict fMRI time series during language processing. They compared different models and found out that the intermediate layers of Transformer models predict up to 100% of the signal up to noise ceil in language regions of interest. They also found a positive correlation between a model's prediction score and its ability to predict the next word of a sentence, and that the more a model predicts activity, the better it predicts behavioural parameters like reading time. Interestingly, Schrimpf et al. also found that Transformer models perform above chance even with no training: while models' architecture (e.g. number of layer) is important, the size of their training dataset or vocabulary do not have much influence on the performance. This point, however, has been challenged by other authors like Pasquiou et al. (2022). Similarily, Caucheteux and King (2022) compared several deep language models with the brain using MEG and fMRI imaging. They too found that the middle layers of compositional models like Transformer-based ones predict brain activity with high accuracy and that a model's quality of prediction and ability to predict next word are positively correlated. They thus argue for a partial – there seems to be a plateau for high performance models – convergence of brains and algorithms in natural language processing: the higher a model's performance, the better its prediction of brain fMRI or MEG activity, although architecture plays an important role too. Caucheteux et al. (2022) even proved that the quality of the fit LLMs' internal representations and brain activations during the listening of a story is correlated to the self-assessed level of comprehension of subjects.

On another note, Caucheteux et al. (2021b) tried to replicate the hierarchy of language brought out by Lerner et al. (2011) by using GPT-2 to emulate the brain response. Their results show strong similarities with those of Lerner et al., suggesting that Transformer models – and especially their intermediate layers – are good models of the way language is processed in the brain. Millet et al. (2022) obtained similar results with other Transformer models. In addition, they found that self-supervised learning models perform sightly better than supervised learning ones.

To compare LLMs' and brain's activation, all the above works use the same method (see fig. 1.1): they select a subset of the stimuli and fMRI data (e.g. 80%) to fit a generalised linear model on of the brain's activity given the model's hidden states to obtain a weight matrix W and compute a *brain score* using the rest of the data

$$\text{corr}(Y, W \cdot X) \tag{1.1}$$

where Y is the true brain response and X the hidden states of the model for the held-out stimuli. The final brain score is obtained by averaging the brain scores obtained for each subset of a partition the dataset.

1.3 Conventions

Softwares. All analyses were coded in `Python` 3.9. Figures were plotted using either `matplotlib`'s `pyplot` or `seaborn`. Statistical tests were performed using `scikit-learn` and `statsmodels` librairies. NLP tasks were done using the `transformers` package. The online experiments were run on Prolific, and designed either with Qualtrics or with `JavaScript`'s `jsPsych` library (and hosted on a Pavlovia server).

Statistical tests. For all statistical tests, a p-value was computed and the null hypothesis was rejected whenever $p < .05$.

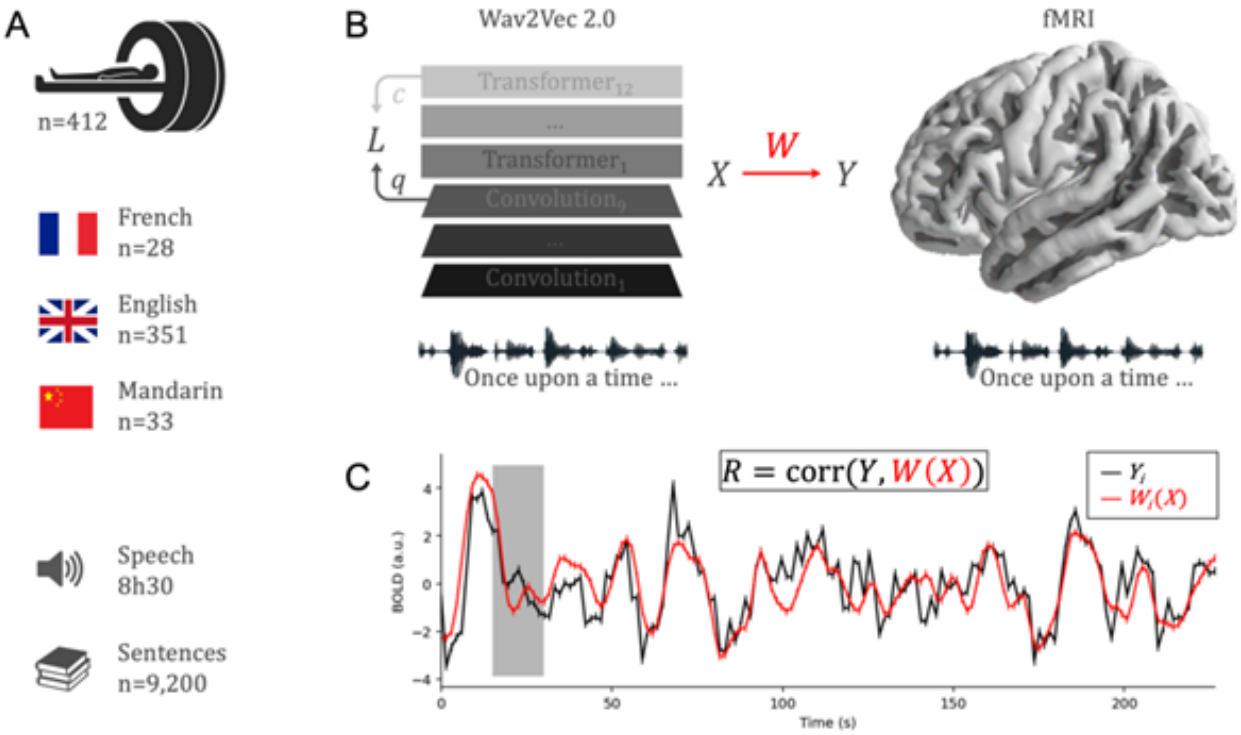


Fig. 1.1. Illustration of the method used to speech representations in brains and deep neural networks. Extract from Millet et al., 2022.

Data availability. All material associated with the work presented in this thesis can be found online, on the following repositories:

- for chapters 2, 4 and 5, `https://github.com/samdebray/ConditionalsLM`;
- for chapter 3, `https://gitlab.pavlovia.org/sdebray/conditionalendorsement`.

For the time being, these repositories are private, be access will be granted upon request.

CHAPTER 2

PRELIMINARY ANALYSES

Goals of this chapter

In this chapter, we want to probe the hypothesis Cummins et al. (1991) and Cummins (1995) formulated on causal reasoning in humans in GPT-3. We thus use Cummins' stimuli and assess GPT-3 performance in order to see whether large language models are able to reason with conditionals, and if they show the same content effects as human. This analysis is exploratory, as there is no literature tackling conditional reasoning in LLMs per se, but Dasgupta et al. (2022) gives reason to hope that we might find the same biases in GPT-3 as in humans.

2.1 Cummins' hypothesis

Cummins et al. (1991) and Cummins (1995) showed that, in the conditional reasoning task (see section 1.2.1) endorsement rates of inferences involving a causal conditional "If ϕ, then ψ." where ϕ is a cause for ψ can be influenced by two factors: the number of alternative causes (other than ϕ) which could trigger ψ, and the number of disabler which, given that ϕ is the case, could prevent ψ from being the case.
To motivate her hypothesis, Cummins et al. (1991) gives two examples:

(1) a. If my finger is cut, then it bleeds.
b. If I eat candies often, then I have cavities.
(Cummins et al., 1991)

Depending on the kind of inference, either the number of disablers or the number of alternative causes is relevant (see fig. 2.1). For instance, take (1) in a modus ponens context (i.e. "If my finger is cut, then it bleeds. It cut my find. Therefore, it bleeds."): while the argument is logically valid in both cases, people should be more inclined to think that it is valid for (1-a) than for (1-b) as they cannot think of many ways of cutting one's finger without bleeding (whereas eating candies often and brushing teeth ten times a day could prevent cavities) ; in other words (1-b) has more disablers than (1-a). Similarily, with affirming the consequent, people should be more incline to endorse the argument in the case of (1-a) than (1-b), even though it is logically invalid in both cases. This comes from the fact there are few things that could cause a finger to bleed other than cutting the skin on it, but one can think of many things that could cause someone to have cavities other than eating candy. In this case so, it is the number of alternative causes which matters.
As shown on fig. 2.2, the predictions made by Cummins are quite close to what is empirically measured.

Later, Fernbach and Darlow (2010) and Fernbach and Erb (2013) proposed a refinement of this model. They suggest that causal powers should be used instead of number of alternatives causes and disabling conditions, following Cheng (1997). More precisely, they propose that the *acceptability of*

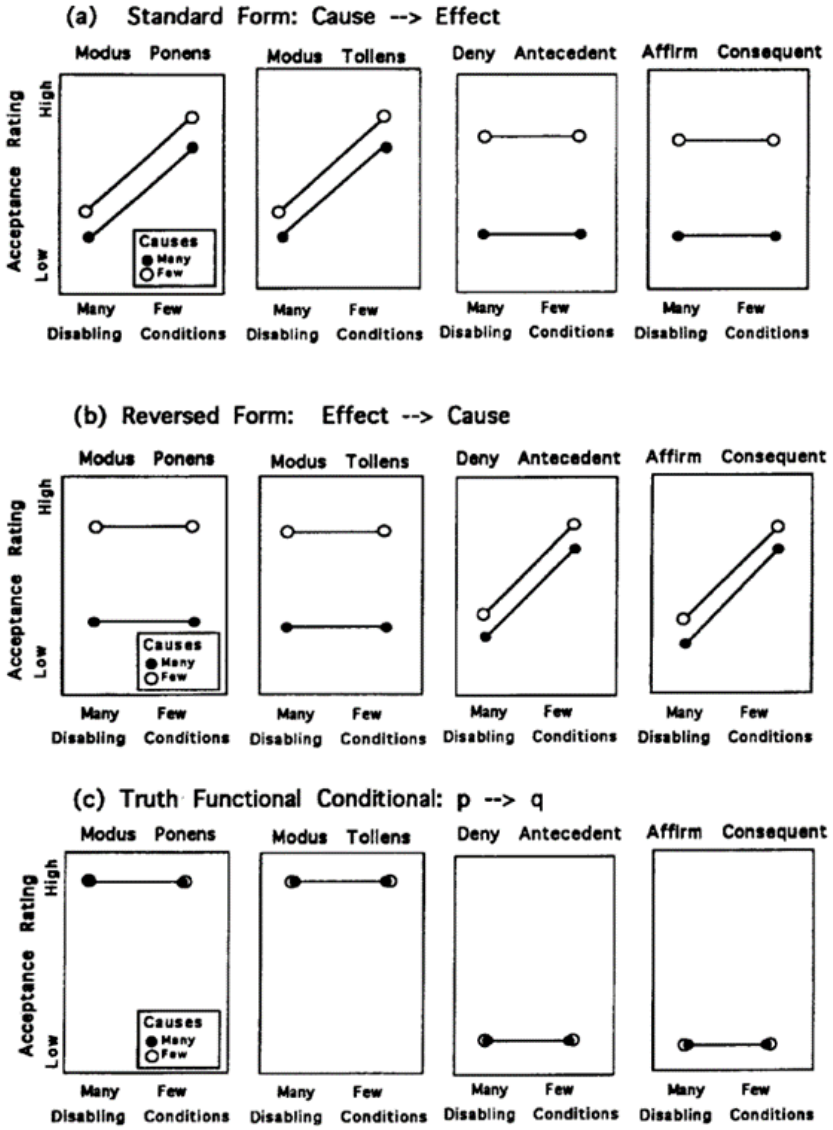


Fig. 2.1. Cummins et al. (1991) and Cummins (1995)'s prediction of acceptability ratings for the four types of inferences, based on the causal analysis (**(a)** and **(b)**, as hypothesised by Cummins), and the formal, material implication, analysis (**(c)**). Extract from Cummins, 1995.

MP and AC be modelled as follows

$$\text{MP} \simeq \Pr(\text{consequence} \mid \text{cause}) = 1 - \left(\sum_{i=1}^{n} A_i - \sum_{i<j} A_i A_j + \sum_{i<j<k} A_i A_j A_k + \cdots + (-1)^{n-1} \prod_{i=1}^{n} A_i \right)$$

$$\text{AC} \simeq \Pr(\text{cause} \mid \text{consequence}) = 1 - (1 - P_c) \frac{W_a}{P_c W_c + W_a - P_c W_c W_a}$$

where W_c is the stength of the cause, i.e. the probability of the cause to successfully trigger the effect, W_a the stength of the alternative causes, P_c the prior on the cause, and $A_i := P_i d_i^*$ with P_i the prior on the i-th disabling cause and d_i^* its disabling strength (i.e. its likelihood of successfully prevent the antecedent cause from bringing about the effect). This approach allows to take into account the strength rather than the count (e.g. a single very probable alternative is more deterrent than many very unlikely ones). These equations can derived assuming that conditional secanarios approximate a *noisy-or common effect model* (Cheng, 1997). The noisy-or model assumes that there are multiple independent causes for a given effect, each of which may or may not be effective on a given trial. The predictions of the data from Cummins (1995) by these equations are better than those of Cummins' model as it accounts for a significant amount of unique variance, beyond what number of alternative

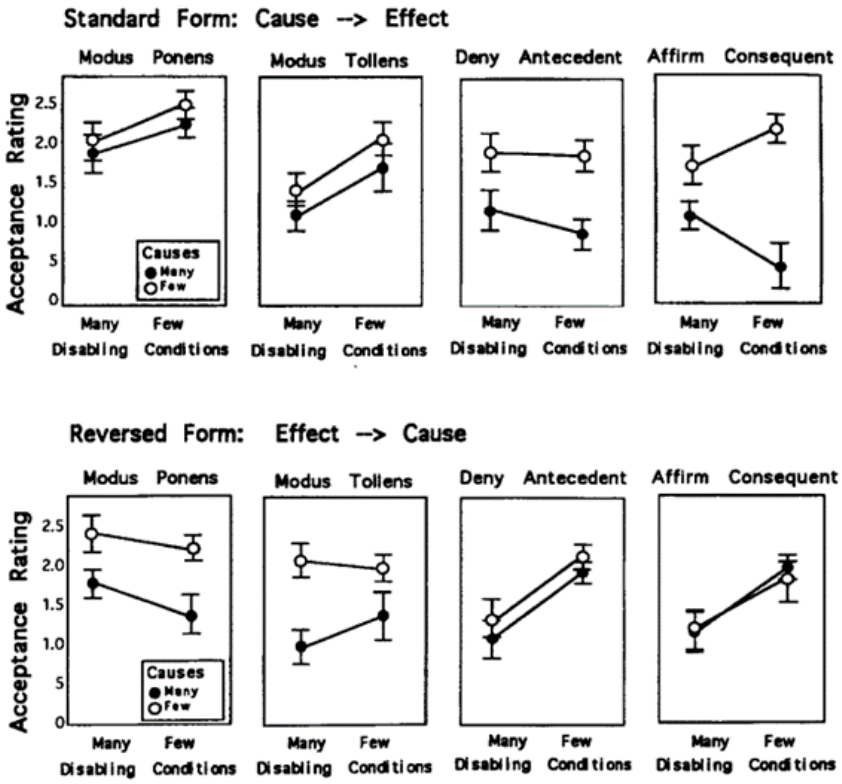


Fig. 2.2. Mean acceptance rate for arguments based on contextualised, causal conditionals presented in their standard (if cause, then consequence) or reversed (if consequence, then cause) form. The rating scale ranged from −3 ("very sure cannot draw this conclusion") to 3 ("very sure can draw this conclusion"), with 0 representing "can't tell". Extract from Cummins, 1995.

causes and disabler can explain (Fernbach & Darlow, 2010).

2.2 Method

For this study, we used the stimuli from Cummins (1995) (see appendix C.1) and evaluated GPT-3's `text-davinci-003` model (Brown et al., 2020) on the causal inference task. There were in total thirty-two statements divided into 2×2 categories (few/many alternative causes and disablers), each containing eight statements. Each statements yielded four inferences (MP, MT, DA and AC), thus summing up to a total of 128 stimuli.

As Brown et al. (2020) suggests that few-shot learning can dramatically improve the performance of GPT-3, We decided to implement k-shot learning for $k \in [\![0, 5]\!]$. The stimuli were presented both in the "If [cause], then [effect]." and the "If [effect], then [cause]." causal directions. In order to show the model a wide variety of examples and to avoid that it only mimics what it has just been shown, the examples were chosen as follows:

- if $k \in [\![1, 3]\!]$: we imposed that all k examples and the actual task had the same direction ("If [cause], then [effect]." or "If [effect], then [cause]."), but different inference types and different premises.
- if $k \in \{4, 5\}$: we also imposed that all $k - 1$ examples and the actual task had the same direction ("If [cause], then [effect]." or "If [effect], then [cause].") different premises, but this time we imposed that the tuple $\langle$inference type, #alt causes, #disablers$\rangle$ were different for all k inferences.

Prompting. In order to fit with the experimental design of Cummins (1995), the prompts were first designed as follows. When $k = 0$, the prompt was phrased as follows (for inference type MP and hypothesis "If fertilizer was put on the plants, then they grew quickly."):

```
Given this statement and this fact, choose the response that best reflects your decision
    about the conclusion?
Statement: 'If fertilizer was put on the plants, then they grew quickly.'
Fact: 'Fertilizer was put on the plants.'
Conclusion: 'Therefore, they grew quickly.'
Response:
- Very sure that I cannot draw this conclusion
- Sure that I cannot draw this conclusion
- Somewhat sure that I cannot draw this conclusion
- Can't tell
- Somewhat sure that I can draw this conclusion
- Sure that I can draw this conclusion
- Very sure that I can draw this conclusion
```

When $k > 0$, the prompt was (for $k = 1$, example inference type MP, inference hypothesis "If fertilizer was put on the plants, then they grew quickly.", evaluation inference type DA and evaluation hypothesis "If the brake was depressed, then the car slowed down."):

```
Given this statement and this fact, choose the response that best reflects your decision
    about the conclusion?
---
Statement: 'If fertilizer was put on the plants, then they grew quickly.'
Fact: 'Fertilizer was put on the plants.'
Conclusion: 'Therefore, they grew quickly.'
Response:
- Very sure that I cannot draw this conclusion
- Sure that I cannot draw this conclusion
- Somewhat sure that I cannot draw this conclusion
- Can't tell
- Somewhat sure that I can draw this conclusion
- Sure that I can draw this conclusion
- Very sure that I can draw this conclusion
=> Very sure that I can draw this conclusion
---
Statement: 'If the brake was depressed, then the car slowed down.'
Fact: 'The brake wasn't depressed.'
Conclusion: 'Therefore, the car didn't slow down.'
Response:
- Very sure that I cannot draw this conclusion
- Sure that I cannot draw this conclusion
- Somewhat sure that I cannot draw this conclusion
- Can't tell
- Somewhat sure that I can draw this conclusion
- Sure that I can draw this conclusion
- Very sure that I can draw this conclusion
=>
```

For subsequent analyses (other that those producing the same plots as Cummins' fig. 2.6 and **??**), we opted for a clearer, binary designed and re-evaluted GPT-3 which different prompts. When $k = 0$, the prompt was phrased as follows (for inference type MP and hypothesis "If fertilizer was put on the plants, then they grew quickly."):

```
Is the following inference valid? Answer with 'yes' or 'no'.
If fertilizer was put on the plants, then they grew quickly. Fertilizer was put on the
    plants. Therefore, they grew quickly.
```

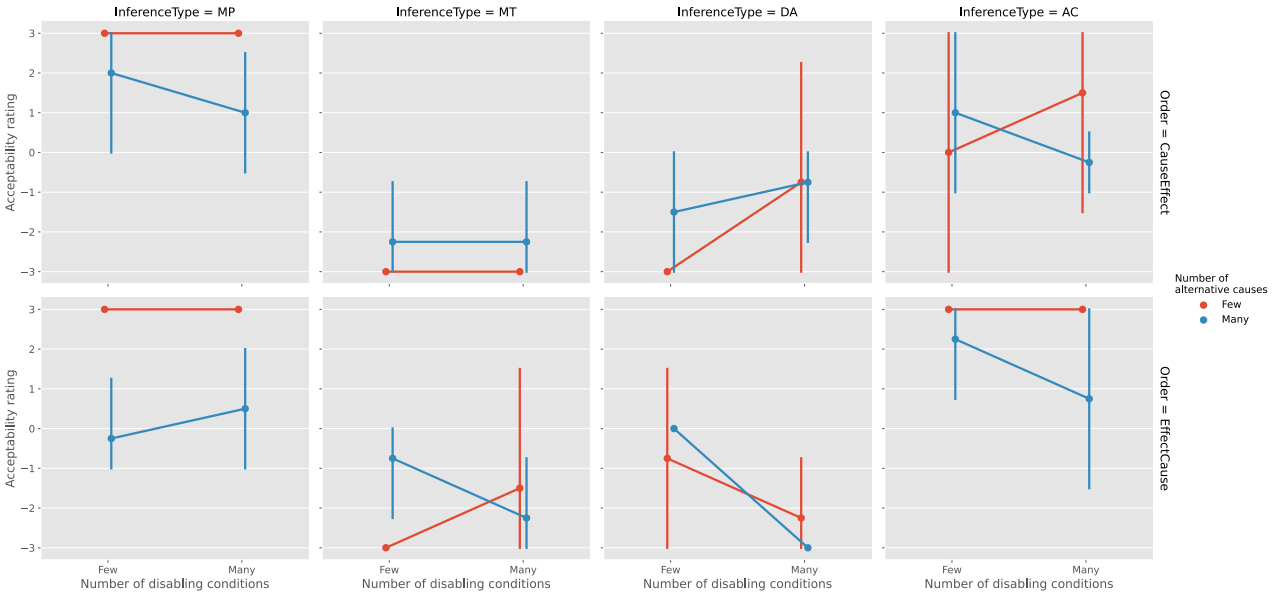
When $k > 0$, the prompt was (for $k = 1$, example inference type MP, inference hypothesis "If fertilizer was put on the plants, then they grew quickly.", evaluation inference type DA and evaluation hypothesis "If the brake was depressed, then the car slowed down."):

```
Is the following inference valid? Answer with 'yes' or 'no'.
If fertilizer was put on the plants, then they grew quickly. Fertilizer was put on the
    plants. Therefore, they grew quickly. => Yes
```

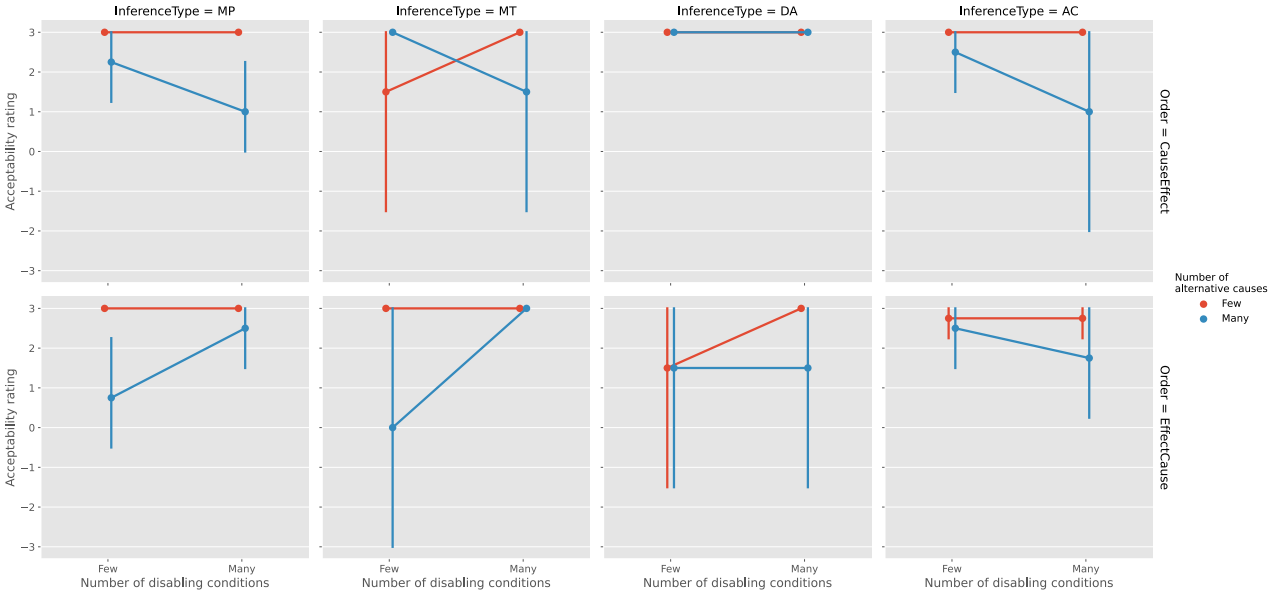
```
If the brake was depressed, then the car slowed down. The brake wasn't depressed. Therefore,
    the car didn't slow down. =>
```

2.3 Results

Unless otherwise stated, the results reported in this section are those obtained for the binary prompts described in section 2.2.



(a) Zero-shot training.



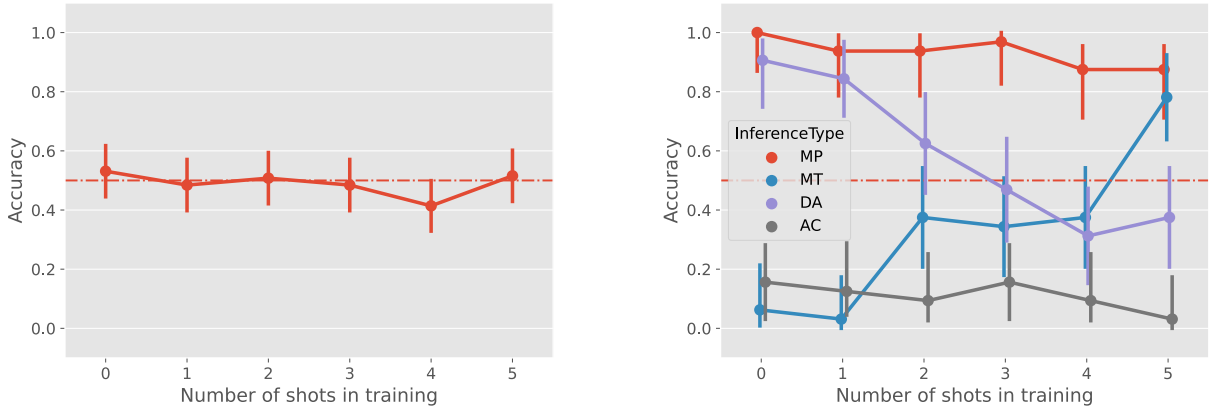
(b) Five-shot training.

Fig. 2.3. Probe of Cummins' hypothesis in GPT-3, with zero-shot and five-shot training (Cummins-like prompt). Vertical bars show 95% Wald confidence intervals.

As shown on fig. 2.3, the predictions of Cummins et al. (1991) and Cummins (1995) do not transpose well to GPT-3 neither with zero-shot training (fig. 2.3a), nor with five-shot training (fig. 2.3b). Moreover, fig. 2.3 (along with fig. 2.4a) makes it clear that GPT-3 does not predict well the validity of conditional inferences. Indeed, because of the very wide confidence intervals, there never seems to be significant differences in acceptability ratings between inferences with few or many alternatives causes and disablers.

In addition, for zero-shot learning, the model seems to judge all MT inferences as invalid and AC inferences as valid. As for five-shot learning, the models classifies almost all inferences as valid. We shall now explore further the model's biases.

First, looked for an effect of the number of shots of learning on model's accuracy (i.e. proportion of correct answers).



(a) Overall. **(b)** Depending on inference type.

Fig. 2.4. GPT-3 accuracy on the conditional reasoning task with Cummins' stimuli against number of shots of training, in the "If [cause], then [effect]." and "If [effect], then [cause]." causal directions. Horizontal lines show chance level and vertical bars show 95% Wald confidence intervals.

Figure 2.4a confirms that the model perfoms at chance level at accurately predicting the validity of a causal inference, no matter the number shots of training. However, as shown on fig. 2.4b, GPT-3's performance seems to depend not only on the number of shots of training first, but also on the type of inference. To probe this hypothesis, we fitted a multivariate logistic regression to predict the model's accuracy based on the inference type, the number of shots of training and their interaction. A significant regression equation was found ($\chi^2 = 427.21$, $p < .001$, see eq. (B.1)), with pseudo $R^2 = .40$ (see table B.1 for full report of the regression results). This regression reveals a significant effect of the number of training shots ($p = .043$) and all inferences types compared to baseline MP (MT: $p < .001$; DA: $p = .019$; AC: $p < .001$). In addition, there is a strong interaction between inference type and number of shots in training (OR $= 3.02$, $p < .001$ for MT $>$ MP per additional training shot). Interestingly, depending on the type of inference, few-shot learning does not always improve the model's performance. For instance, the previous regression showed that the odds of the model being correct decrease by $1 - e^{\beta_4+\beta_6} = 45\%$ for each additional shot of training in the DA inference type. This effect is visible on fig. 2.4b, where the model perfoms high above chance (90% accuracy) on DA for zero-shot learning, but below chance for four-shot learning.

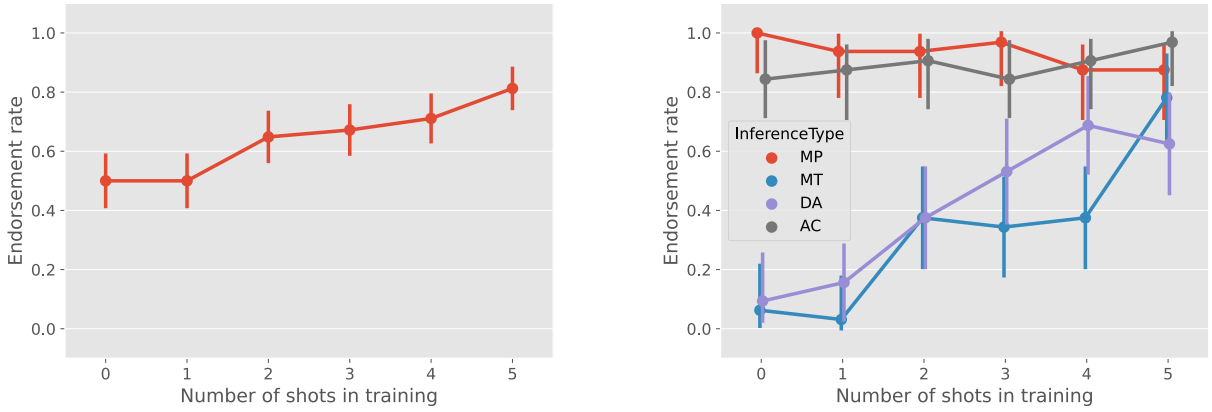
Let us now examine whether the model shows human-like content effects on reasoning with conditionals. Since the points on fig. 2.3 were obtained by averaging only 8 measures, we did not have enough statistical power to perform an ordinal regression with the model's acceptance rate as dependent variable and inference type, number of alternative causes and number of disablers as independent variables. However, we still wanted to investigate semantic content effects, and thus looked for an effect of the causal direction of the hypothesis. Indeed, we could expect that it would be more difficult to reason from cause to effect than from effect to cause, and we may also observe a significant interaction with the inference type: it is easier to reason backward (MT and AC) if the hypothesis is in the "If [cause], then [effect]." direction and to reason forward (MP and DA) is the hypothesis is in the other direction.

One can see on fig. 2.5 that the causal direction of hypothesis does not seem to have any significant effect on the model's accuracy. To confirm this observation, a multivariate logistic regression was calculated to predict the model's accuracy based on the type of inference, the number of shots of training and their interactions. A significant regression equation was found (eq. (B.2), $\chi^2 = 337.61$,

Fig. 2.5. Accuracy of GPT-3 on Cummins' stimuli against number of shots of training depending on the hypothesis' causal direction for each inference type. Horizontal lines show chance level and vertical bars show 95% Wald confidence intervals.

$p < .001$), with pseudo $R^2 = .32$ (see table B.2 for full report of the regression results). No effect of any term of the regression involving the causal direction of the hypothesis was declared significant in the regression. Therefore, GPT-3 does not seem to exhibit any causal content effect on reasoning with conditionals.

However, by observing fig. 2.4 and fig. 2.6 we hypothesise that GPT-3's behaviour is strongly impacted by syntax. More specifically, it seems that: (i) with small number of training shots, the model accurately predicts the validity of forward inferences (MP, DA) and poorly predicts that of backward inferences (MT, AC); (ii) with small number of training shots, the model judges that inferences containing a negation (MT and DA) are invalid and that the others are valid; (iii) as the number of training shots increases, the model learns that all inferences are valid.



(a) Overall. **(b)** Depending on inference type.

Fig. 2.6. Proportion of inferences classified as valid by GPT-3 against number of shots of training, in the "If [cause], then [effect]." and "If [effect], then [cause]." causal directions. Vertical bars show 95% Wald confidence intervals.

To make these effects clearer, we also ploted the dependent variables in (i) and (ii) against the number of shots of training depending on the independent variables in fig. B.1. We also fitted a multivariate logistic regression to confirm (i) and (ii) and significant regression equations were found (pseudo $R^2 = .35$, $\chi^2 = 353.01$, $p < .001$ for (i); pseudo $R^2 = .28$, $\chi^2 = 294.55$, $p < .001$ for (ii)). We found a significant effect of negation on model's predicted validity (OR $= 0.01$, $p < .001$ which means that, for zero-shot learning, the odds of the model predicting an inference is valid are 99% lower if the inference contains negation than if it does not) and a strong interaction between the presence of negation and the number of shots of training (OR $= 1.95$, $p < .001$). Similarily we found a significant effect of inference direction on model's accuracy (OR $= 0.01$, $p < .001$) and a strong interaction between inference direction and number of shots of training (OR $= 2.21$, $p < .001$; see table B.3 and table B.4

for full results). To probe (iii), we fitted a simple logistic regression with dependent variable number of training shots and independent variable number of shots of training. The results are reported in table 2.1. The table shows that there is a significant effect on number of shots of training (OR = 1.33,

Independent Variable	Estimate	Standard Error	Odds Ratio	p-value	Significance
Number of training shots	0.29	0.046	1.33	<0.001	***
Intercept	−0.101	0.13	0.90	0.44	

Table 2.1. Logistic regression analysis of stimuli's validity predicted by GPT-3 with independent variable the number of shots of training. Pseudo $R^2 = .04$, $\chi^2 = 39.77$, $p < .001$. Significance code: ***$p < .001$; **$p < .01$; *$p < .05$.

$p < .001$) and that the model's odds of predicting an inference is valid increases by 33% with each additional shot of training (see fig. 2.6a).

Finally, it must be noted that GPT-3 is very sensitive to prompting.

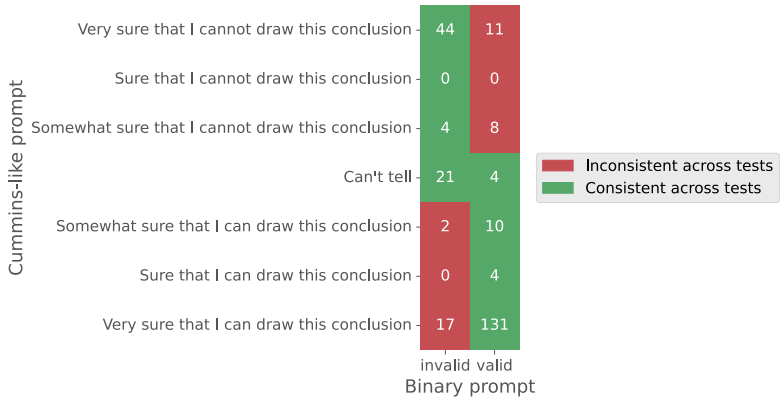


Fig. 2.7. Distribution of GPT-3 validity judgement in the two prompting settings presented in section 2.1. Cells coloured in red indicate that the model was not consistent across the two settings.

On fig. 2.7, one can see that there are 19 inferences that GPT-3 classified as invalid in binary prompt setting and valid in the Cummins-like prompt setting, and also 19 inferences that it classified as valid in binary prompt setting and invalid in the Cummins-like prompt setting.

2.4 Discussion

In this chapter, we presented a probe of Cummins' hypothesis of causation (Cummins et al., 1991; Cummins, 1995) in GPT-3 (Brown et al., 2020). Qualitatively, Cummins' hypothesis on the effect on causality on performance doesn't seem to hold for GPT-3. The model actually does not show any significant content effect, and is largely influenced by prompting and syntax. With zero-shot training, GPT-3 judges all inferences which feature negation as invalid and considers the others valid and, as the number of shots of training increases, it learns to classify all inferences as valid. Importantly, GPT-3 performs at chance level at predicting the validity of causal inferences no matter the number of shots of training.

There are two ways that we can interpret the fact that, with zero-shot training, GPT-3 judges all inferences which feature negation invalid and considers the others valid. First, this could mean that the model only processes syntactic features and relies on the presence of negation to predict validity. The second, more optimistic, interpretation is that the model is actually good at forward reasoning with zero-shot learning (see eq. (B.4)). Further investigation is needed to find out which interpretation is the right one. However, the fact that few-shot learning moderates this trend can be considered a clue in favour of the less optimistic, purely syntactic interpretation.

Finally, we showed that GPT-3 was inconsistent across the two prompt settings for 14.8% of the stimuli we presented it, which is highly unsatisfactory.

In addition to this prompt issue, the main problem we met when analysing GPT-3's performance on the stimuli from Cummins (1995) is the reduced size of the dataset. This led to poor quality regressions and very wide confidence intervals. As a consequence, we deemed necessary to have a larger dataset, which is why we decided to create our own conditional reasoning dataset. Furthermore, having only access to the model's textual output is a strong limitation, as it only yields discrete independent variables, and limitates the range of analyses we are able to perform. For this reason, we turned to another (very) large language model: MetaAI's OPT with 66B parameters (Zhang et al., 2022).

CHAPTER 3

THE CONDITIONAL REASONING DATASET

Goals of this chapter

We reported in chapter 2 studies that prove that humans showed content effects when reasoning with causal conditionals. However, we also saw that the data collected by Cummins is not sufficient to find these effects in large language models such as GPT-3. In this chapter, we will explain how we collected a large dataset of conditional statements featuring different kinds of relations between the antecedent and the consequent that would help us achieve two goals:

- get a coarser-grained analysis of human content effects on reasoning, showing that it is harder to reason with conditionals featuring no causation at all than with direct causal ones for instance;
- gain more statistical power by averaging over a larger number of inferences.

Our hypothesis here is that it is harder for humans to reason with non-causal than with causal conditionals (by increasing level of difficulty: causal $>$ confound $>$ correlated $>$ independent), and that the direction of the causal relation between the antecedent and the consequent also has an effect: it should be easier to reason from cause to effect than the reverse.

3.1 Generation of conditional statements

As explained in section 1.1 and as a follow-up of the analyses presented in chapter 2, we decided to run a bigger grained study than that of Cummins et al. (1991). Namely, we would like to confirm that humans show content effects when reasoning with conditionals featuring different causal relations between the antecedent and the consequent[1]. Prior to running any study, we identified different types of possible causal relations between the antecendent (A) and the consequent (C):

- A causes C (causal cause–effect);
- C causes A (causal effect–cause);
- A and C have a common cause and neither A causes C nor C causes A (confound);
- A and C have no causal relation whatsoever, but the probability of C being the case when A is the case is higher than that of C being the case unconditional on A: $\Pr(\text{C} \mid \text{A}) \gg \Pr(\text{C})$ (correlated);
- A and C have no causal relation whatsoever and $\Pr(\text{C} \mid \text{A}) \simeq \Pr(\text{C} \mid \neg\text{A}) \simeq \Pr(\text{C})$ (independent);
- A and C feature jabberwocky[2] words (jabberwocky, for control: it is purely syntactic and doesn't involve prior causal knowledge as the words are unknown).

Our prior hypothesis was that it would be increasingly difficult for humans and machines to judge the validity of inferences based on conditionals featuring the above relations (easiest for A causes C and most difficult for independent).

In order to evaluate language models and humans on the same task, we needed to decide on a

[1] In Cummins et al. (1991), all conditionals were directly causal, but the causal strength varied across conditionals (e.g. number of disablers or alternative causes).

[2] See `https://www.poetryfoundation.org/poems/42916/jabberwocky`.

dataset that we would use. We needed to be able fine-tune a large language model if necessary. This task requires a large number of examples (typically around $10,000$) so that the model can learn from a wide variety of data. Such a big dataset didn't exist, neither in the literature of deep learning (causal conditionals have not been studied a lot in NLP) or in psychology. Therefore, we had to create our own.

Jabberwocky clauses. The jabberwocky sentences were generated automatically. More specifically, we used Markov chains to generate pseudowords based on trigrams tabulated from a list of $370,105$ English words, using a `R` script provided by Christophe Pallier (`https://github.com/chrplr/openlexicon/tree/master/scripts/pseudoword-generation-by-markov-on-trigrams`). We generated 510 statements by first sampling at random and independently: a mood (indicative or subjunctive), an article (definite or indefinite), two negation forms ("not" or "n't"), two first names from the list of the 200 most common first names in the US (100 male and 100 female), and four jabberwocky words (each of length L given by the random variable $L = 6 + X$ where $X \hookrightarrow \mathcal{B}(9, .5)$). We then used this information to create the antecedent, the consequent, their affirmation (for conditional sentences) and negation. The template for indicative conditionals is "If [name] [jabberwocky]s [article] [jabberwocky], then the [jabberwocky] is [jabberwocky]ed" (e.g. "If Amy secaliperas a sikvery, then the coulowee is dettededed."), and for subjunctive conditionals it is "If [name] hadn't/had not [jabberwocky]ed [article] [jabberwocky], then the [jabberwocky] wouldn't/would not be [jabberwocky]ed." (e.g. "If Jordan had nontalianed the lishr, then the celesurupl would be hypothembleed.").

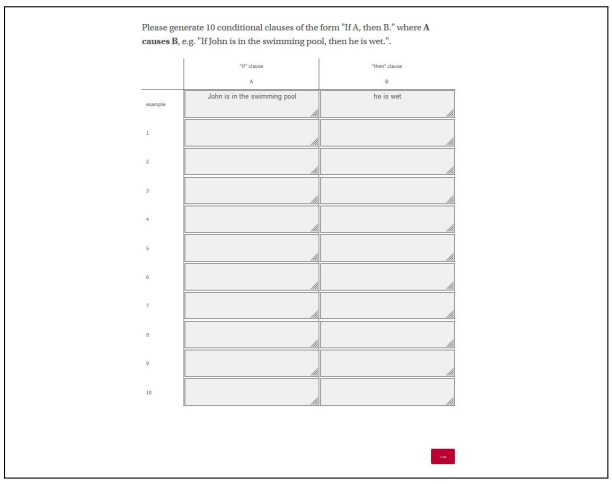
Other clauses. To generate the other conditionals for the other causal conditions, we ran an online experiment on Prolific. We required that the participants be native English speakers and have a Prolific approval rate of at least 95%. Participants were paid £3 for their participation. The answers were reviewed manually to exclude unsuitable sentences (spelling mistakes, ungrammatical sentences, wrong causal condition, etc.).

	# participants	Causal		Confound		Correlated		Independent	
		# collected	% kept	# collected	% kept	# collected	% kept	# collected	% kept
1st pilot	16	166	72.3	93	22.5	79	13.9	130	80.0
2nd pilot	14	152	73.0	80	38.8	—	—	108	86.1
1st run	52	492	63.0	—	—	—	—	475	70.1
2nd run	10	66	77.3	31	19.4	—	—	—	—

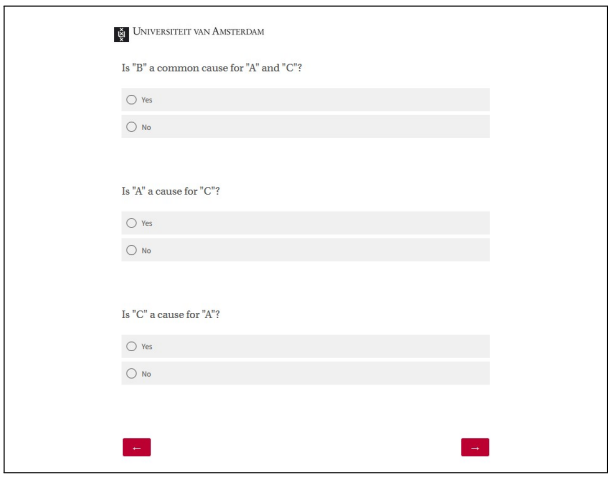
Table 3.1. Summary of the studies run to collect the conditional reasoning dataset.

For each causal condition, we asked participants to generate ten clauses. They were allowed not to fill all ten fields, and we provided an example each time, to make sure that they understood the instruction well (see fig. 3.1). table 3.1 summarises all the studies we ran to collect the dataset.

We first ran a pilot which included all four conditions (we only included the causal cause–effect condition as we could obtain the effect–cause condition simply by swapping the antecendent and the consequent). Unsurprisingly, we found that participants had a very hard time coming up with clauses in the correlated condition (these are known to be very difficult to imagine since they violate Reichenbach's common cause principle, see Reichenbach, 1956). Since participants spent a lot of time of the correlated task and were not able to generate suitable clauses anyway, we decided to exclude this task from the subsequent surveys and that the final dataset would not contain many clauses in this condition. We thus ran a second pilot which showed that the confound task was quite difficult and time-consuming as well, so we decided to exclude this task from the first run of the survey as well. After the first run, we still missed a few hundred clauses to make it to $10,000$ inferences (each clause induces at least four inferences: one in each condition MP, MT, DA and AC), so we ran a second survey including only the confound and causal tasks. We modified the confound task and asked participant to explicitely check that the common cause was a cause for both the antecendent and the consequent and that the latters were not causally related at all (see fig. 3.1b). We included the causal



(a) Causal task.



(b) Confirmation screen from the confound task from the 2nd run.

Fig. 3.1. Screenshot of two different tasks from the online survey.

task in this second run so as to make sure we would have enough clauses in the end and that would not have to run a third survey.

All in all, we collected and kept 588 clauses in the (direct) causal condition, 530 in the independent condition, 58 in the confound condition and 10 in the correlated condition.

Generation of inferences from the statements. As we only asked the participants to the survey to provide the antecedent and the consequent of the conditional clauses, we had to generate the inferences from these. We first tried to do it automatically by defining the premise as "If [antecedent], then [consequent]." and, depending on the type of inference, defining the hypothesis and conclusion as "[antecedent]." and "[consequent]." (for MP) or "[negation-antecedent]." and "[negation-consequent]." (for DA). To do so, we needed to get the negations of the antecedent and the consequent provided by the participants. We attempted to use OPT-66B for this purpose, but we found that neither few-shot training nor prompt hacking was sufficient for OPT to perform this task properly. For instance, the model very often failed at finding the negation of a negative sentence like "Peter doesn't like cats.": in these cases, the model just left the sentence unchanged. In addition, we realised that the automatic procedure to generate inferences was problematic as it sometimes produced ill-formed inferences. This was the case especially for subjunctive conditionals and for clauses written in the future. For instance, if the premise is "If Harry had emptied the dishwasher, Meghan wouldn't have had to do it.", the automatic procedure would have defined "Meghan wouldn't have had to do it." as the hyptothesis and "Harry had emptied the dishwasher." as the conclusion for AC, but the argument would not make sense; and the same goes with premise "If Peter pushes the button, then the light will turn on." for backward inferences (MT and AC). However, the automatic procedure would not be able to detect these clashes (or it would require some machine learning, but we already saw that this would at least require human supervision) and would thus produce many ungrammatical or awkward sentences.

We found no systematic way of fixing these issues, and we therefore had to review individually every inference to check the grammar and the tenses. Since we are not native English speakers, there may still be some issues in the dataset, but we think we solved the vast majority of them.

3.2 Validity judgement of conditional inferences

Now that we had collected the dataset of conditional inferences. We needed to run an experiment to collect human endorsement rates and reaction times (RTs) for a subset of these inferences in order to be able to compare the reasoning pattern of humans and models.

3.2.1 Method

For this second experiment, we had to meet some constraints and to make sure we would collect enough judgements per inference. We thus chose arbitrarily 48 conditional clauses (12 in each condition causal cause–effect, causal effect–cause, independent and jabberwocky) from the whole dataset, and used the $4 \times 48 = 192$ inferences they yield for the online experiment. So as to make sure these sentences did not contain any mistake or clumsy syntax, we asked a native English speaker to review and correct all 192 inferences.

The contraints that had to be met were budget (which did not allow us to recruite more than fifty participants for 25 ± 5 min) and statistical power (as we wanted to average RTs and endorsement rates over at least twelve observations). We aimed for an average duration of 25 ± 5 min so as to remain within the attention span of participants and measure RTs as accurately as possible.

The experiment was coded using the `jsPsych` module of `JavaScript` and hosted on a Pavlovia server. Participants were recruited on Prolific and paid £3 for their participation. We required that all fifty participants take the study on a laptop, be native English speaker, have a minimum Prolific approval rate of 90%, have graduated from High School and have not participated in the previous studies on conditional clauses generation presented in section 3.1.

The experiment started with a short survey asking the whether the participant had a training in logic, their age, last degree obtained and major. Then some instructions were displayed and the main block started. It consisted in sixty-four repetitions of the same judgement task: one for each inference. The inferences were randomised, and we only imposed that each participant judged four inferences per causal condition and inference type (i.e. MP, MT, DA and AC). The judgement task started with the display of a fixation cross on the screen for 1.000 ms, then the participant was asked to judge whether the argument was valid (see fig. 3.2).

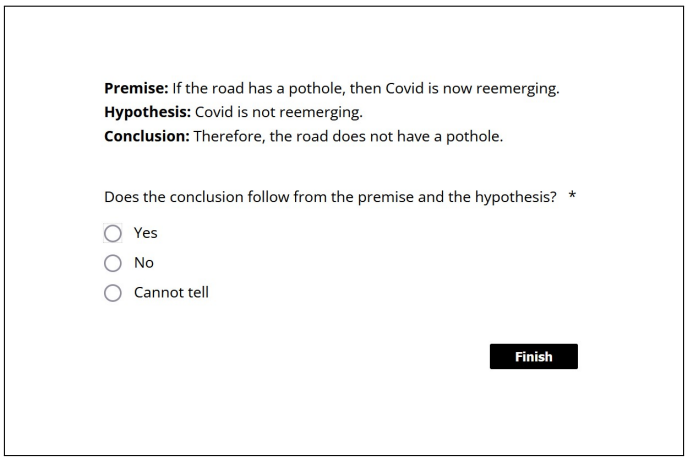


Fig. 3.2. Screenshot of the screen asking participants to judge whether an inference was valid (inference type: MT, causal condition: independent).

3.2.2 Results

Exclusion criteria. Before running any analysis, we computed the mean RT μ across all judgements and the associated standard deviation σ and systematically excluded all judgements made within less than $\mu - 2\sigma$ or more than $\mu + 2\sigma$. In other words, we performed a two-tailed z-test on RTs and imposed that $|z| \leq 2$ for a judgement not to be considered an outlier.

Sanity checks. First, we made sure that we received enough judgements for each inferences. On fig. 3.3, one can see that all inferences received at least eight different judgements, and that 75% of them received more than fourteen judgements ($\mu = 16.5$, $\sigma = 3.8$).

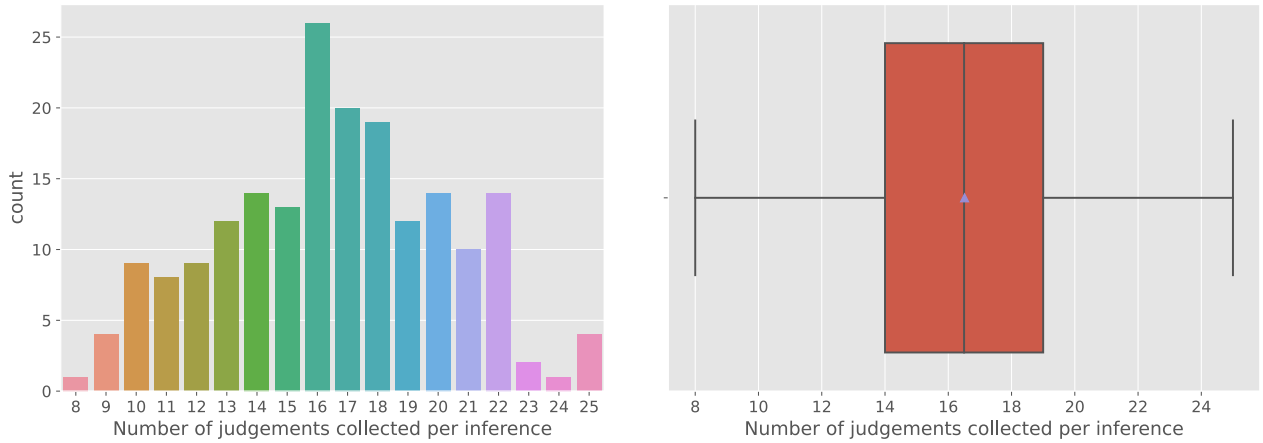


Fig. 3.3. Distribution of the number of judgements collected per inference.

RTs. As the inferences were chosen at random for the whole dataset, we could not match their length. Therefore, the first step in analysing reaction times is to see whether these are predicted by the length of the premise of the argument presented.

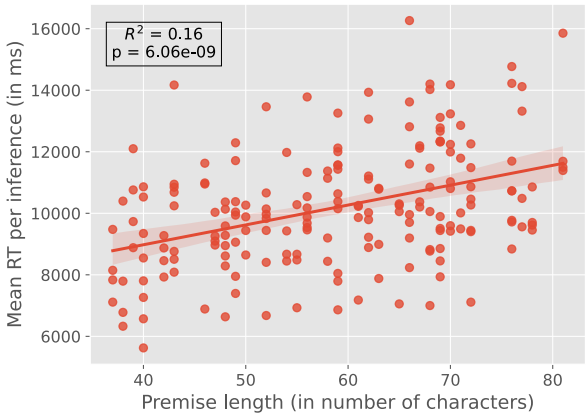


Fig. 3.4. Relationship between mean RT per inference and length of premise.

fig. 3.4 shows that premise length is not a good predictor of the mean RT for a given inference ($R^2 = .16$, $F(1, 190) = 37.12$, $p < .001$). We then ran a two-way type I ANOVA on mean RT per inference with first independent variable premise length and second independent variable causal relation between antecendent and consequent. We found a statistically significant main effect of causal relation on RT, $F(3, 184) = 9.90$, $p < .001$ (see fig. 3.5).

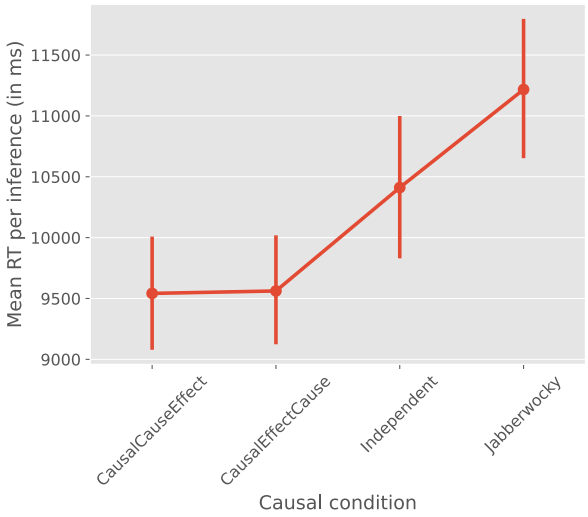


Fig. 3.5. Mean RT per inference against causal relation between antecedent and consequent. Vertical bars show 95% confidence intervals.

We then ran a post hoc Tukey HSD test, with family-wise error rate (FWER) .05 and found that the mean RT for causal conditionals (both in the cause–effect and effect–cause settings) was significantly lower than that of jabberwocky conditionals ($p < .001$ both times). However, the difference of means between other groups could not be declared significant (see fig. 3.6).

As one can see on fig. 3.5, the means for the cause–effect and effect–cause settings of the causal group seem to be very similar. However, we expect to observe an interaction between the direction of the inference (forward when reasoning from antecedent to consequence and backward otherwise) and the causal direction of the premise (cause–effect or effect–cause), as show on fig. 3.7. We thus ran a two-way type III ANOVA on mean RT per inference restricted to (direct) causal premises, with

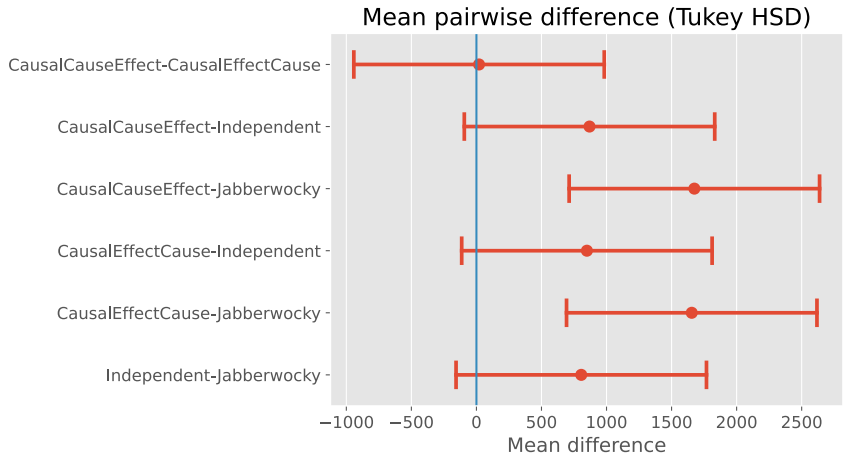


Fig. 3.6. Tukey HSD test, with family-wise error rate (FWER) .05, for pair-wise comparison of difference of mean RTs between all causal conditions.

independent variables direction the inference and causal direction of the premise.

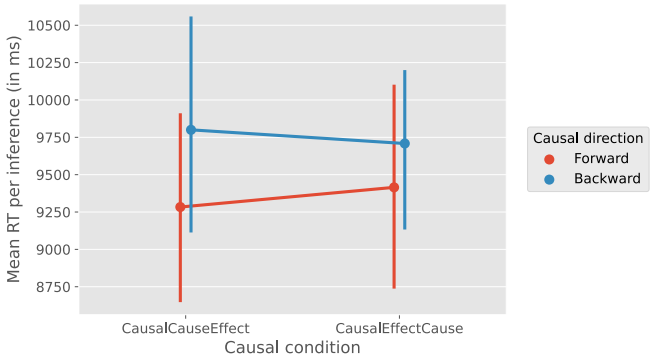


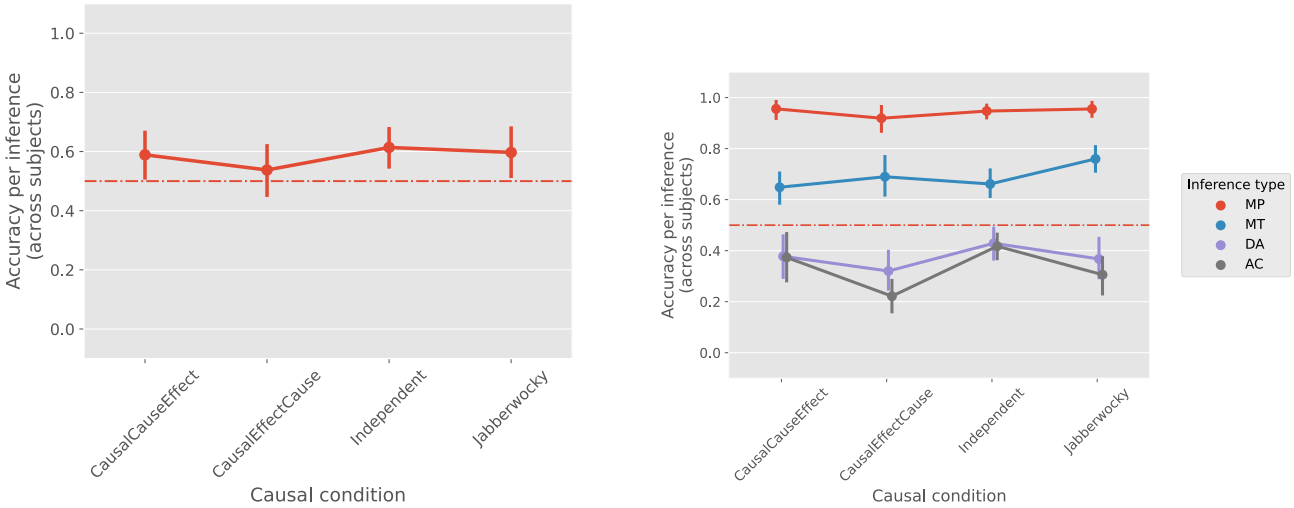
Fig. 3.7. Mean RT per inference for direct causal inferences against causal direction of the premise, depending on direction of inference. Vertical bars show 95% confidence intervals.

We did not find a significant interaction between the two independent variables, $F(1, 92) = 0.11$, $p = .74$.

Accuracy of prediction. We also wanted to test whether the causal relation between the antecendent and consequent in the premises had an effect on the accuracy of prediction, that is the proportion of participants who correctly classified the inference as valid or invalid.
We first ran a one-way ANOVA on accuracy of prediction with independent variable the causal condition of the premise. We found no significant effect of the causal condition ($F(3, 188) = 0.67$, $p = .57$, see fig. 3.8a). However, we then realised that performance depended a lot on the kind of inference involved (MP, MT, DA or AC). We thus ran a two-way type I ANOVA with first independent variable the inference type and second independent variable the causal condition (see fig. 3.8b). We found significant effects of inference type ($F(3, 176) = 305.69$, $p < .001$) and causal condition ($F(3, 176) = 3.95$, $p = .009$), as well as a significant interaction between the two independent variables ($F(9, 179) = 2.36$, $p = .015$).

3.3 Discussion

We showed that the results from Cummins (1995) could be extended in the following sense: it gets more difficult for humans to judge the validity of a conditional inference as the complexity of the



(a) Overall. **(b)** Depending on the inference type.

Fig. 3.8. Accuracy of prediction against causal condition. Vertical bars show 95% confidence intervals, vertical line shows chance level.

causal relation between the antecedent and the consequent in the premise increases (direct causal > independent > jabberwocky). Further work is needed to include other causal relation, such as confounds or correlated. These were not included in the study as we saw in section 3.1 that statements featuring these relations are very difficult for people to come up with.

In particular, we found a very significant effect of the causal relation between the antecedent and the consequent on reaction times. However, there was no effect of this variable on endorsement rates. This fact could be explored further by asking participants to tell how confident they are about their judgement, as Cummins (1995) did. Indeed, endorsement rates as we used them in this study do not take into account the difficulty for people to produce a judgement (only RTs account for this), but our hypothesis is precisely about how difficult it is for people to comprehend the inferences.

Finally, the pairwise differences in RTs between the causal conditions were not always significant, but the reason for this might be the design of the experiment itself. Indeed, the experiment had to meet some constraints, but we would have got much better results by asking the participants to come in the lab and control their environment. In addtion, a click-and-validate design is not very suitable when measuring reaction times but, again, it was all we could do with the resources we had. In any case, a different design would probably have enabled to collect more accurate RTs that would have shown a stronger effect of causal category, especially since Tukey HSD test is quite conservative. However, the fact that we still found an effect despite the noise in our data proves that the effect of causal category is very strong.

CHAPTER 4

CONDITIONAL REASONING IN OPT

Goals of this chapter

We saw in chapter 2 that GPT-3 only performs at chance level to predict the validity of conditional inferences. We thus ran another study trying to investigate further the performance of LLMs on condition reasoning. We hope to get a better insight by:

- using an open-source model, namely OPT with 66 billion parameters, so that we can access its scores and hidden-states and analyse them along with its textual output;
- using several different prompts, so that the results can be averaged to get rid of the high dependency to prompts we found in chapter 2;
- have a coarser grained approach, by using not only direct causal conditionals, but conditionals featuring a range of causal relations (including causal independence and jabberwocky statements) thanks to the dataset from chapter 3;
- using many more conditionals, to be able to gain statistical power.

By doing so, we hope to get a better understanding of how OPT handles reasoning with causal conditionals. In concrete terms, we will measure different scores (including accuracy of prediction), and assess the model's performance and how the causal relation between the antecedent and the consequent of the premise affect these scores.

4.1 Method

For this study, we decided to use OPT with 66 billion parameters (Zhang et al., 2022), available from HuggingFace, and the dataset collected in chapter 3. Like in chapter 2, we used k-shot learning in prompts for $k \in [\![0, 5]\!]$. In order to present the model with a wide variety of conditions, the k example inferences and the main inference were chosen at random with the constraint that they must all feature a different causal relation between the antecedent and the consequent of the premise (causal cause–effect, causal effect–cause, confound, correlated, independent and jabberwocky). As we had collected less statements in the confound and correlated than in the other conditions (see chapter 3), we only used the associated inferences in the training shots, but never as the main inference. That is to say, *the analyses only include the causal cause–effect, causal effect–cause, independent and jabberwocky conditions.*

We also used several prompts to control for the large effect of prompting we found in chapter 2. The different prompts for 0-shot learning are presented in fig. 4.1 with inference "If Alice went for a walk in the rain, then she is wet." and inference type MP. These prompt settings where chosen because we noticed that instructions and capitalisation to impact quite heavily models' performance.
When $k \geq 1$, instruction 0 was put in the plural form and each example was separated from the previous one with two line breaks. Thanks to few-shot learning and to the use of several prompts, each inference was evaluated twenty-four times by the model.

```
Carefully evaluate this logical argument.

Premise: "If Alice went for a walk in the
    rain, then she is wet."
Hypothesis: "Alice went for a walk in the
    rain."
Conclusion: "Therefore, she is wet."
Question: Is this argument valid?
- yes
- no
Answer:
```

(a) Instruction 0, uncapitalised.

```
Carefully evaluate this logical argument.

Premise: "If Alice went for a walk in the
    rain, then she is wet."
Hypothesis: "Alice went for a walk in the
    rain."
Conclusion: "Therefore, she is wet."
Question: Is this argument valid?
- Yes
- No
Answer:
```

(b) Instruction 0, capitalised.

```
Determine whether the conclusion follows from
     the premise and the hypothesis.

Premise: "If Alice went for a walk in the
    rain, then she is wet."
Hypothesis: "Alice went for a walk in the
    rain."
Conclusion: "Therefore, she is wet."
Question: Is this argument valid?
- yes
- no
Answer:
```

(c) Instruction 1, uncapitalised.

```
Determine whether the conclusion follows from
     the premise and the hypothesis.

Premise: "If Alice went for a walk in the
    rain, then she is wet."
Hypothesis: "Alice went for a walk in the
    rain."
Conclusion: "Therefore, she is wet."
Question: Is this argument valid?
- Yes
- No
Answer:
```

(d) Instruction 1, capitalised.

Fig. 4.1. Different prompt settings for inference "If Alice went for a walk in the rain, then she is wet." (causal cause–effect) in MP setting.

We loaded the model in half-precision using `pytorch`'s `float16` standard. To run inferences with OPT-66B, we used the `AutoModelForCausalLM.from_pretrained()` function provided by the `transformers` library, along with the associated tokeniser. The evaluation script was run on a GPU node from the Dutch Snellius national supercomputer. This node was equiped with two Intel Xeon Platinum 8360Y CPUs and four NVIDIA A100 GPUs, and we used the automatic device mapper provided by the `accelerate` library.

In addition to the negative log-likelihood (NLL) associated with the completion tokens (␣`yes` and ␣`no` for the uncapitalised prompt settings, and ␣`Yes` and ␣`No` for the capitalised) after processing of the whole input, we also saved the following parameters:

- the mean NLL for the whole prompt of the main inference (i.e. excluding the instruction and the training examples when applicable);
- the mean NLLs for the premise, hypothesis and conclusion of the main inference (excluding the quotation marks and final full stop);
- the NLL associated with completion tokens in a baseline prompt context, where the baseline prompt is simply `Answer:`, to apply the DC-PMI correction in the analyses (Holtzman et al., 2022).

We defined two model predictions:

- *without* DC-PMI correction, the model prediction was "valid inference" if $\text{NLL}_{\text{yes}} \leq \text{NLL}_{\text{no}}$ and "invalid inference" otherwise;
- *with* DC-PMI correction, the model prediction was "valid inference" if $\text{NLL}_{\text{yes}} - \text{NLL}_{\text{baseline yes}} \leq \text{NLL}_{\text{no}} - \text{NLL}_{\text{baseline no}}$ and "invalid inference" otherwise.

In this chapter, we will always consider the predictions with DC-PMI correction, as suggested by Holtzman et al. (2022).

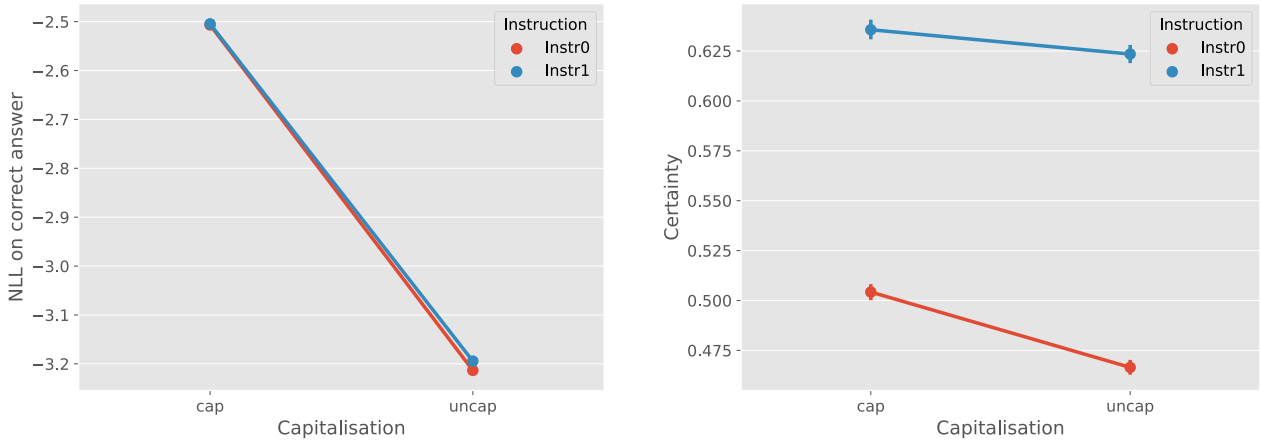
We identified four quantities that could inform us about the model's handling of conditonal

inferences. The first one is the *endorsement rate*: it is the percentage of inferences from a given group that the model judged as valid (across all or within each prompting and few-shot learning settings). The *accuracy* is similar to the endorsement rate but corresponds to the percentage of inferences classified correctly (as valid or invalid). The third one is the *NLL on the correct answer* (that is `␣yes` or `␣Yes` for valid inferences and `␣no` or `␣No` for invalid ones), which measures the model's surprisal for this answer. Finally, the *certainty* is the positive difference between the NLLs on the two completions, and it measures how certain the model is about its choice. Both NLL on the correct answer and certainty are computed *with* the DC-PMI correction.

4.2 Results

Sanity checks. We first assess the effects of the methodological choices we made in section 4.1.

Effect of prompting. As we saw in chapter 2, prompting has a strong influence on GPT-3's performance. Here, we assess whether it also is the case for OPT-66B.



(a) Effect of prompting on NLL on correct answer.

(b) Effect of prompting on certainty.

Fig. 4.2. Effect of prompting on OPT-66B's performance at predicting the validity of conditional inferences. Vertical bars show 95% confidence intervals.

As one can see on fig. 4.2, there seems to be a significant effect of both capitalisation and instruction on the certainty and the NLL on the correct answer. To probe this hypothesis, we ran a two-way type III ANOVA on the NLL on the correct answer with independent variables capitalisation, instruction and their interactions (see fig. 4.2a). We found a significant effect of both variables (for capitalisation: $F(1, 212636) = 75179.65$, $p < .001$; for instruction: $F(1, 212636) = 17.15$, $p < .001$), as well as a significant interaction ($F(1, 212636) = 10.82$, $p = .001$). Similarily, we ran a type III ANOVA on certainty with independent variables capitalisation, instruction and their interactions (see fig. 4.2b). Again, we found a significant effect of both variables (for capitalisation: $F(1, 212636) = 197.24$, $p < .001$; for instruction: $F(1, 212636) = 6577.37$, $p < .001$), as well as a significant interaction ($F(1, 212636) = 51.65$, $p < .001$).

Influence of DC-PMI correction. We also wanted to see whether DC-PMI correction really had an impact on the accuracy on OPT-66B's predictions, as suggested by Holtzman et al. (2022). One can see on fig. 4.3 that DC-PMI does not seem to have any effect whatsoever on the accuracy of prediction. In fact, the model seems to perform at chance no matter the correction or number of shots in training. We tried to fit a logistic regression to account for the effect of DC-PMI correction on accuracy, but we found no significant regression equation.

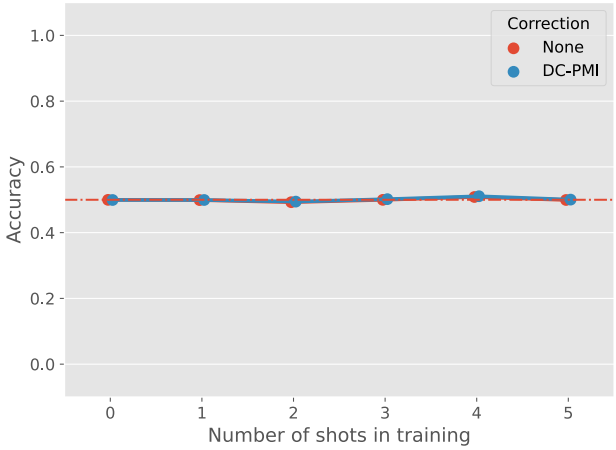


Fig. 4.3. Effect of DC-PMI correction on OPT-66B's accuracy of prediction, depending on number of shots in training. Vertical bars show 95% Wald confidence intervals.

Effect of negation. We saw in chapter 2 that the presence of negation in the inferences (no negation in MP and AC, negation in MT and DA) have dramatic effect on the model's endorsement rate. We wanted to test this in OPT-66B as well.

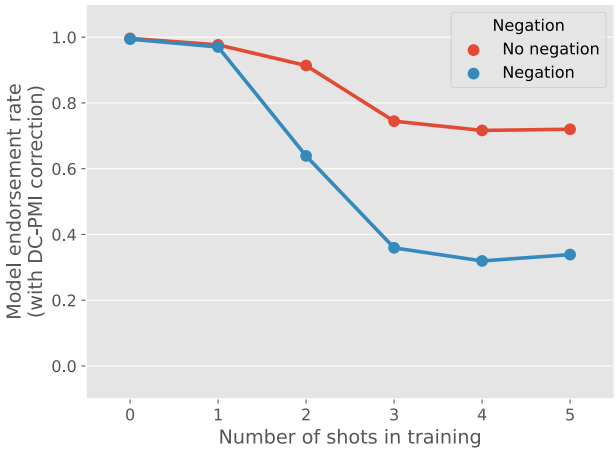


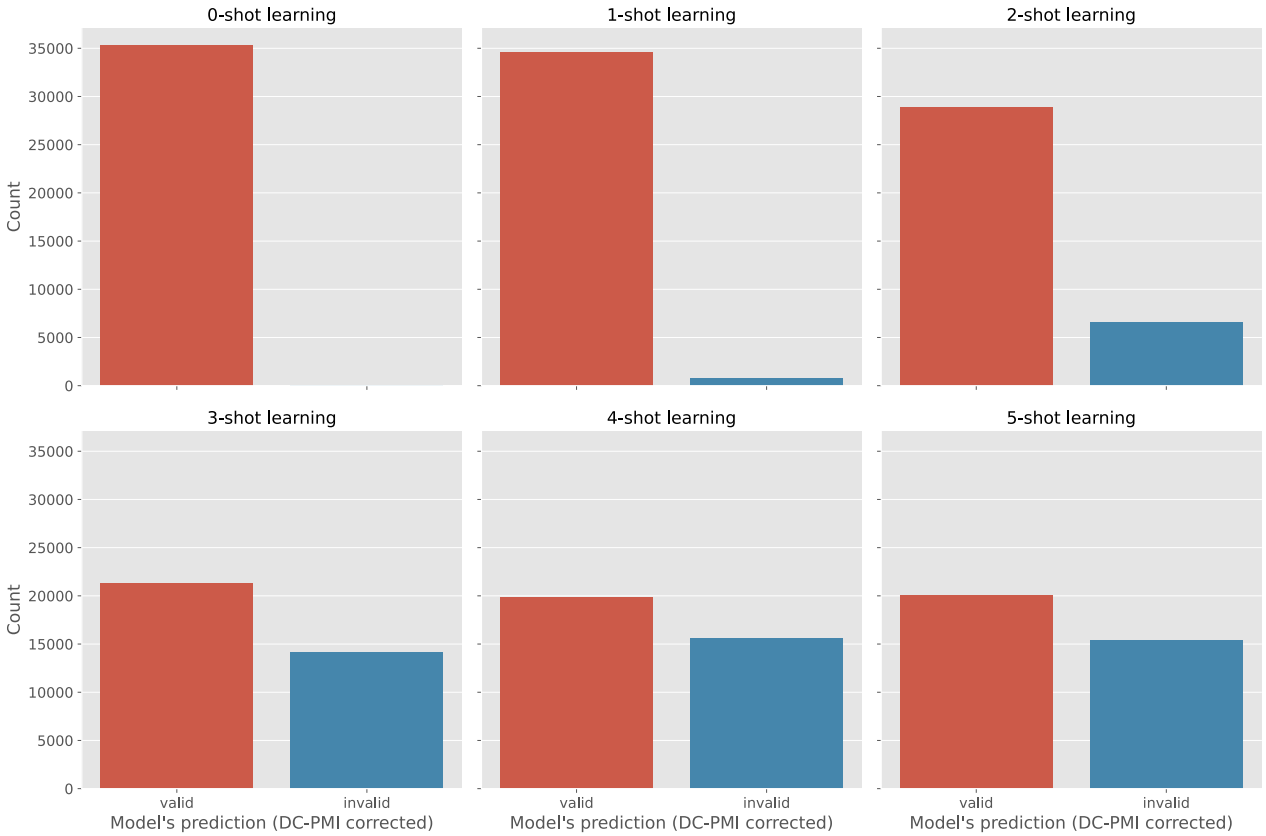
Fig. 4.4. Endorsement rate against number of shots in training depending on whether inferences feature negation or not. Vertical bars show 95% Wald confidence intervals.

One can infer from fig. 4.4 that there is a strong effect of negation. Indeed, for those inferences that contain negation, the endorsement rate flips when the number of shots of training increases. We fitted a multivariate logistic regression with independent variable the endorsement rate (with DC-PMI correction) and independent variables the number of shots in training and the presence of negation to confirm this impression. A significant regression equation was found ($\chi^2 = 63087.40$, $p < .001$) with pseudo $R^2 = .25$ (see table B.5). This revealed a significant effect of the number of shots in training ($p < .001$) and of the presence of negation ($p < .001$). In addition, the interaction is also significant ($p < .001$).

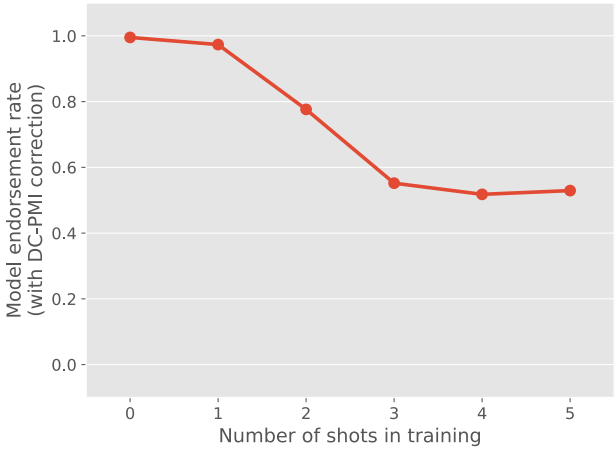
Training dynamics. One can also see on fig. 4.4 that it seems that the model starts by judging that all inferences are valid, that differences are only introduced after a few shots of training. This trend is also made clear on fig. 4.5.

We then refined fig. 4.5b to see whether the behaviour we observed depended on the type of inference or the causal condition of the premise.

We see on fig. 4.6a that OPT-66B becomes less sure about the validity of the different inferences.



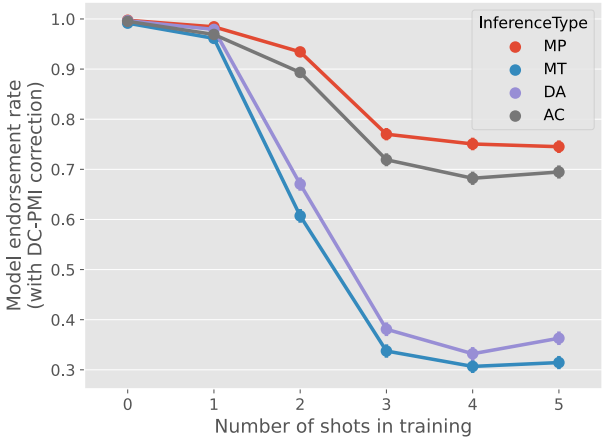
(a) Number of inferences classified as valid and invalid (with DC-PMI correction) depending on number of shots in training.



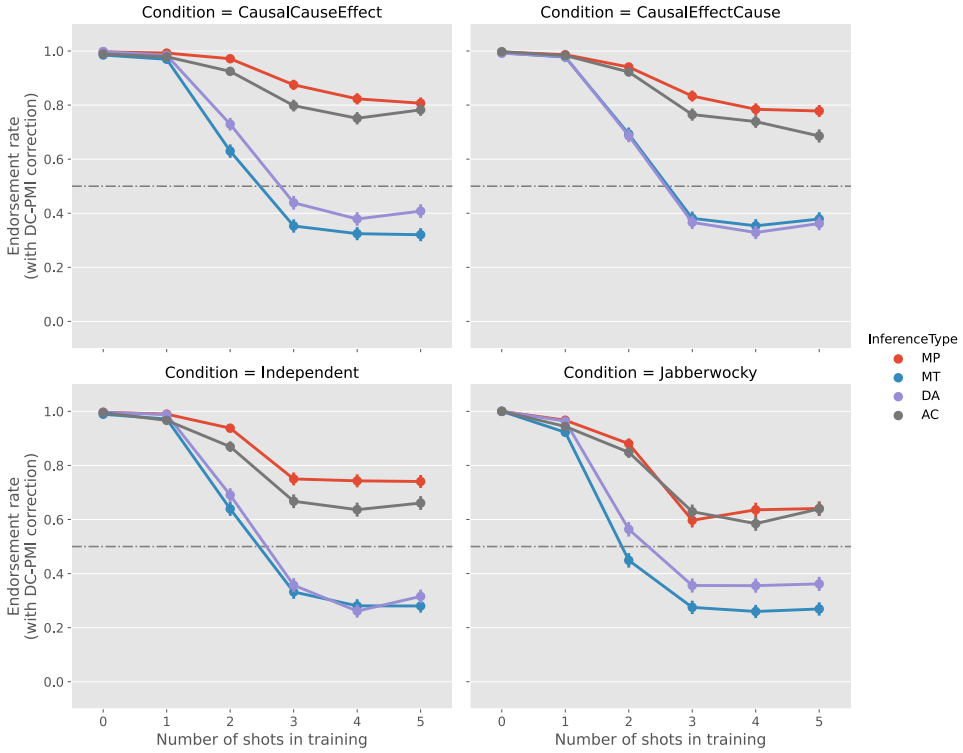
(b) Endorsement rate (with DC-PMI correction) against number of shots in training.

Fig. 4.5. Evolution of model's endorsement rate with number of shots of training. Vertical bars show 95% Wald confidence intervals.

To confirm this impression, we fitted a multivariate logistic regression of the endorsement rate with independent variables the number of shots in training, the type of inference and their interaction. A significant regression equation was found ($\chi^2 = 63605.74$, $p < .001$) with pseudo $R^2 = .25$ (see table B.6). We found significant effects of all inferences types compared to baseline MP ($p < .001$) and of number shots in training ($p < .001$), and significant interactions between number of shots and all inference types except AC ($p < .001$ each time). As described in table B.6, with each additional shot of training the odds of the model considered an inference as valid decrease by 56.8% for MT > MP and by 56.4% for DA > MP. This pattern is the same for all causal conditions, as shown on fig. 4.6b.



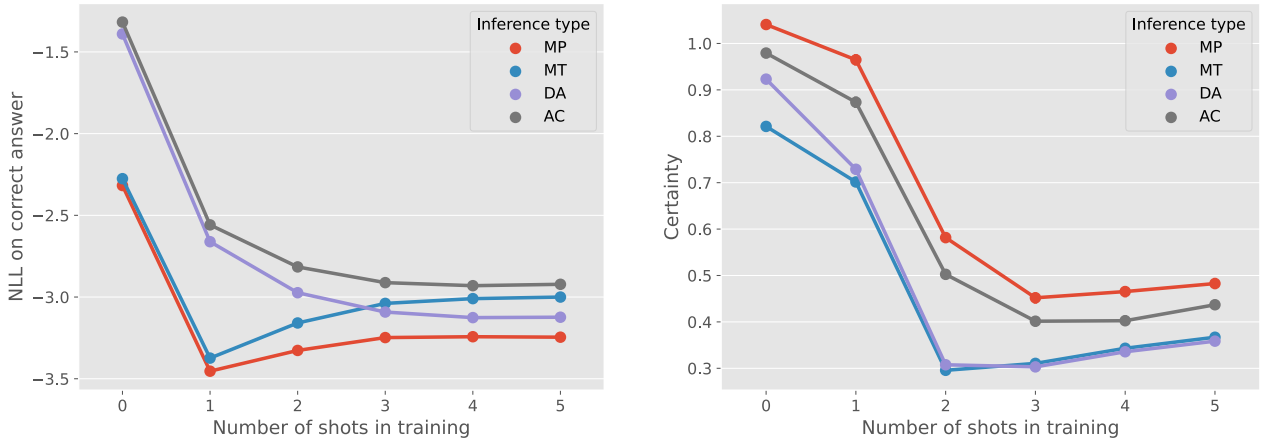
(a) Inference type.



(b) Causal condition.

Fig. 4.6. Evolution of model's endorsement rate with number of shots of training, depending on inference type (fig. 4.6a) or causal condition (fig. 4.6b). Vertical bars show 95% Wald confidence intervals.

When looking at the NLL on correct answer and certainty of the model, we found that the former doesn't change a lot after the first shot of learning. However, the latter is still improving also for later shots of learning, as shown on fig. 4.7.
These effects were confirmed by type III two-way ANOVAs on NLL on correct answer and certainty with independent variables number of shots of training and inference type. Both times, a significant effect of number of shots of training (for NLL on correct answer: $F(5, 212616) = 51244.60$, $p < .001$; for certainty: $F(5, 212616) = 20047.38$, $p < .001$) and of inference type (for NLL on correct answer: $F(3, 212616) = 19034.72$, $p < .001$; for certainty: $F(3, 212616) = 3838.14$, $p < .001$) were found. Interactions between the two independent variables were also significant (for NLL on correct answer: $F(15, 212616) = 2453.93$, $p < .001$; for certainty: $F(15, 212616) = 121.84$, $p < .001$).



(a) NLL on correct answer (with DC-PMI correction).

(b) Certainty (with DC-PMI correction).

Fig. 4.7. Effect of few-shot learning on NLL on correct answer (fig. 4.7a) and certainty (fig. 4.7b). Vertical bars show 95% confidence intervals.

Accuracy. We then turned to assess whether accuracy is significantly different between the different causal conditions. As one can see on fig. 4.8, there seems to be an effect of causal condition on accuracy.

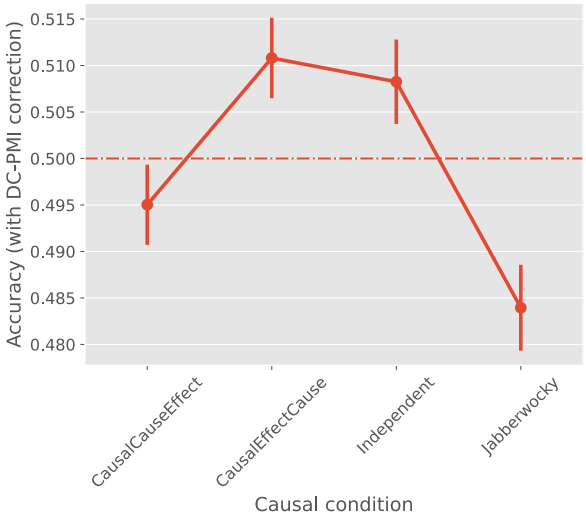


Fig. 4.8. OPT-66B's accuracy (with DC-PMI correction) against causal condition of the premise. Vertical bars show 95% Wald confidence intervals.

However, the accuracy remains roughly at chance level for all conditions (between 48% and 51.5%). In order to probe this hypothesis, we fitted a logisitc regression of the model's accuracy with the causal condition as independent variable. A significant regression equation was found ($\chi^2 = 96.25$, $p < .001$) with pseudo $R^2 = 0$ (see table B.7). A significant effect of all causal conditions against baseline (causal effect–cause) was found ($p < .001$ each time).

NLL on correct answer. We then looked for an effect of causal condition on the NLL on correct answer.

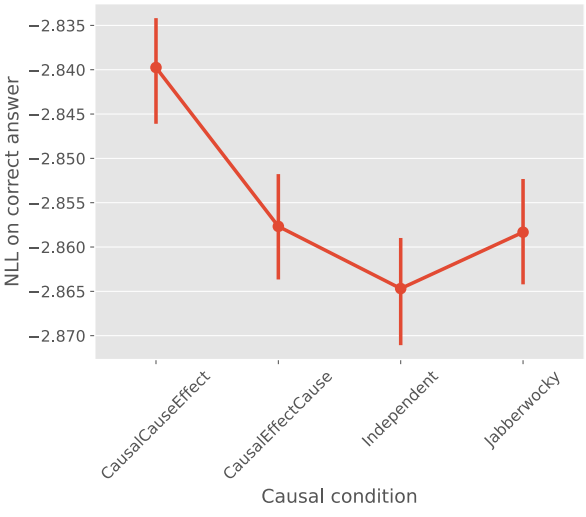


Fig. 4.9. Mean NLL on correct answer against causal condition of premise. Vertical bars show 95% confidence intervals.

As one can see on fig. 4.9, there seems to be an effect of the causal condition of the premise. The model gets less surprised for the causal categories causal cause–effect > causal effect–cause > independent and then gets more surprised by the jabberwocky category. To confirm this impression, we ran a one-way ANOVA with dependent variable NLL on the correct answer and independent variable the

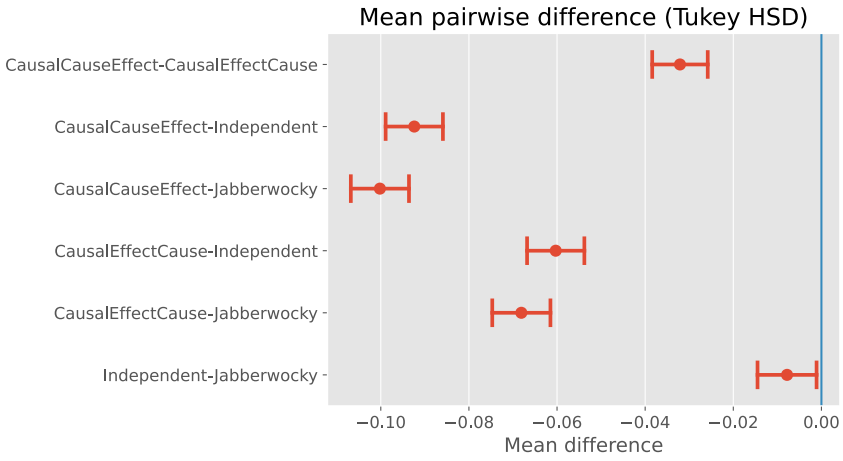


Fig. 4.10. Tukey HSD test, with FWER .05, for pair-wise comparison of difference of NLL on correct answer between all causal conditions.

causal condition of the premise. We found a significant effect of causal condition ($F(3, 212636) = 13.45$, $p < .001$). We then ran a post hoc Tukey HSD test with FWER .05 and found that the NLL on correct answer for the causal cause–effect condition was significantly higher than for all other conditions. However, the differences between the other groups could not be declared significant (see fig. 4.10).

Certainty. We finally did the same analysis for certainty.

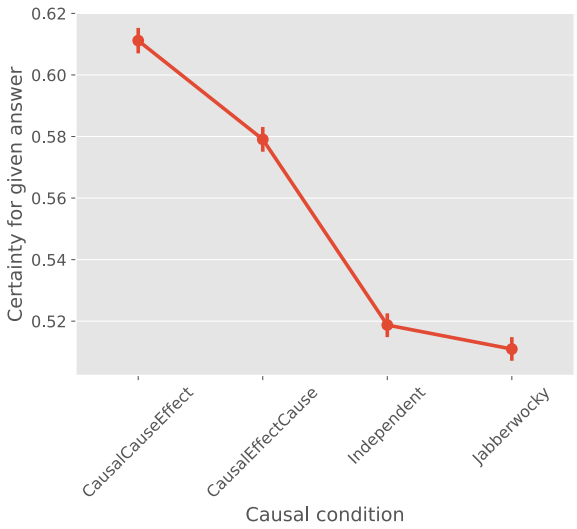


Fig. 4.11. Mean certainty against causal condition of premise. Vertical bars show 95% confidence intervals.

As one can see on fig. 4.11, there seems to be an effect of the causal condition of the premise. The model becomes less certain for the causal categories causal cause–effect > causal effect–cause > independent > jabberwocky. To confirm this impression, we ran a one-way ANOVA with dependent variable certainty and independent variable the causal condition of the premise. We found a significant effect of causal condition ($F(3, 212636) = 720.61$, $p < .001$). We then ran a post hoc Tukey HSD test with FWER .05 and found that all pair-wise difference were significantly different from 0 (see fig. 4.12).

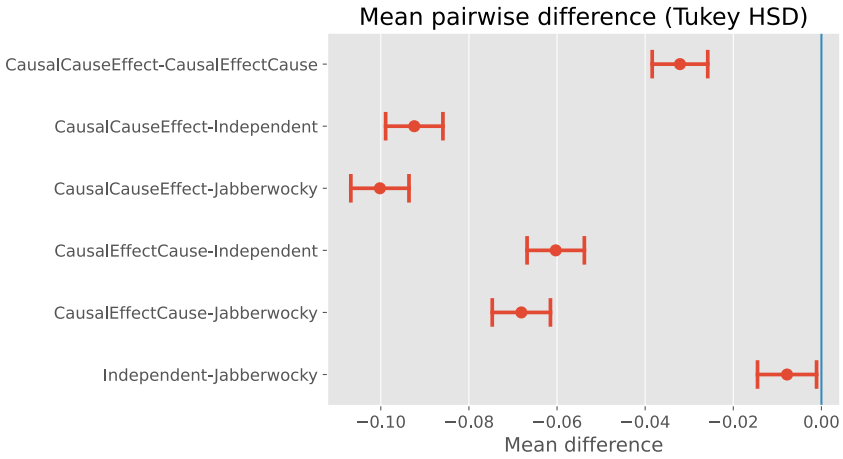


Fig. 4.12. Tukey HSD test, with FWER .05, for pair-wise comparison of difference of certainty between all causal conditions.

4.3 Discussion

In section 4.2, we found that OPT-66B's performance depended on the prompt, both on the capitalisation of the choices and on the instruction prompted before the task. We expected to find this, similarily to chapter 2, and it is actually the reason why we decided to use several prompt settings after analysing the results from chapter 2. The conclusions we will draw in the rest of this section won't depend on prompting (as we will average over all settings), but they still need to be taken with a grain of salt: it might the case that other prompt settings significantly alter OPT-66B's performance.

As far as DC-PMI is concerned, we did not find any effect of the correction on the model's accuracy. These results are surprising insofar as Dasgupta et al. (2022) found that DC-PMI correction was beneficial for models' predictions, which is also what Holtzman et al. (2022) suggested. Although we saw that it did not have much importance, the analyses for this chapter were run with DC-PMI correction in order not to deviate from our original plans.

We also found that few-shot learning had an effect on the model's behaviour, as suggested by Brown et al. (2020). Indeeed, even though few-shot training did not improve the model's accuracy, it did affect both the mean NLL on correct answer and certainty. We saw that there were significant differences between 0, 1 and 2-shot learning, and that there seems to be a plateau after the third shot of training, although certainty seems less stabilised that NLL on correct answer. However, few-shot learning did not have the same impact on both variables: the suprisal of the model on the correct answer decreases as the number of shots increases, while the certainty decreases. Thus, few-shot training seems to help the model pick up on some cues, but these are not related to logical accuracy. Further analyses are required to understand what exaclty happens there.

Regarding the effect of negation, our results are the opposite from those of chapter 2. Indeed, we saw that GPT-3 always classified inferences which do not feature negation as valid, and increases its endorsement rate for those featuring negation as the number of shots of learning increases. For OPT-66B, the trend is the other way around: it starts by classifying all inferences as valid, and decreases its endorsement rate for all types of inferences as the number of shots in training grow. The decrease is greater of those inferences that feature negation that for the others. This observation highlights the differences between large language models, which may have very different behaviours. In addition, the fact that this trend is the same for all causal conditions shows that, *at the level of endorsement rate and accuracy, the model doesn't make big distinctions between causal conditions.* This is made even clearer by the fact that OPT-66B performs at chance at predicting the validity of conditional inferences no matter the causal condition.

Finally, we showed that, although the model performs at chance for predicting the validity of inferences, the causal condition of the premise has a significant effect both on the NLL on correct answer and on the certainty of the model. The fact that the model's suprisal decreases with the conditions causal cause–effect > causal effect–cause > independent > jabberwocky is very puzzling, and requires further investigation to be understood. However, the certainty of the model changes exactly as we would have expected, as the model becomes less certain about its answer as the causal relation between the antecedent ans the consequent becomes less clear. In particular, the model is quite certain about causal inferences (like humans), and uncertain about independent and jabberwocky sentences. This observation gives us confidence about the model's ability to get a remote sense of causality; but is also points to the fact that it is not really picking up on the syntactic pattern of the logical inference.

We saw that the model is a poor predictor of the validity of conditional inferences, but we know that humans are too. We shall now turn to assess whether OPT is a good predictor of human reasoning patterns.

CHAPTER 5

OPT VERSUS HUMAN SUBJECTS

Goals of this chapter

We saw in chapter 3 that human reaction times were significantly different across the different causal categories of conditionals. In chapter 4, we that, although OPT-66B is not a good predictor of the validity of conditional inferences (but neither are humans), it still picks up on some cues. Our hypothesis is that it might be the case that the model actually picks up on human reasoning patterns. In this chapter, we will evaluate whether the model's and human endorsement rates and RTs are correlated and we will try to identify the best settings (in terms of prompting and few-shot learning) to obtain an optimal correlation.
Since we understand the architecture of the model better than that of the human brain, comparing human and models will give us useful information on what cognitive features are necessary to apprehend causality and conditional reasoning.

5.1 Method

To compare humans and OPT-66B, we will use the data from chapters 3 and 4.

The analyses presented in the next section are obtained from the raw data from section 3.2 (after exclusion of outliers). We also included all judgements collected from OPT-66B, in all different prompting and few-shot learning settings. We decided to include all few-shot learning settings in order to model the fact that inferences are presented sequentially to human participants, who thus gain familiarity with the task over the course of the experiment.

Unless otherwise stated, the dataset is restricted to the inferences presented to human subjects during the online experiments from chapter 3.

5.2 Results

First, we tried to probe whether human reaction times were correlated with the mean surprisal of the model for the premise.
We found that the mean NLL on premise (averaged across all 24 presentations of each inference) and the mean RT per inference are weakly *negatively* correlated ($r(190) = -0.20$, $p = .004$). We also computed Pearson's correlation coefficient between mean RT and mean NLL on hypothesis, conclusion and full prompt, but none was significant (see table 5.1).

We also tried to find the best settings to get a good correlation between mean RT and mean NLL on premise. As shown on fig. 5.2, there is always a significant weak negative correlation between the two variables, but the r coefficient does not vary a lot across the settings ($\mu = -0.19$, $\sigma = 0.02$).

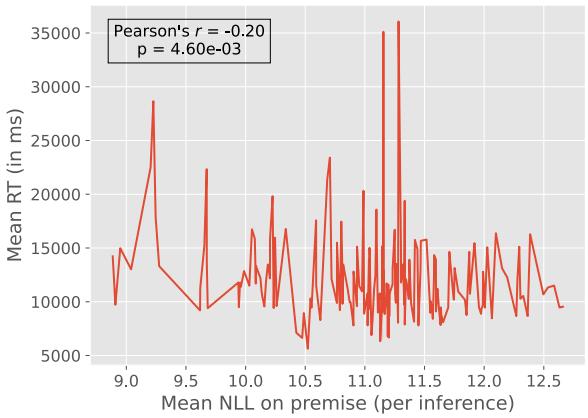


Fig. 5.1. Pearson's correlation between the mean human RT per inference and mean OPT-66B's NLL on the premise (averaged over all few-shot training and prompt settings). Note that the model's surpisal increases as the NLL increases (that is, higher values of NLL mean higher surprisal).

Variable	df	Pearson's r	p-value
NLL on hypothesis	190	0.01	0.90
NLL on conclusion	190	−0.01	0.72
NLL on full prompt	190	0.07	0.35

Table 5.1. Pearson's correlation between mean human RT per inference and OPT-66B's mean NLL on hypothesis, conclusion and full prompt.

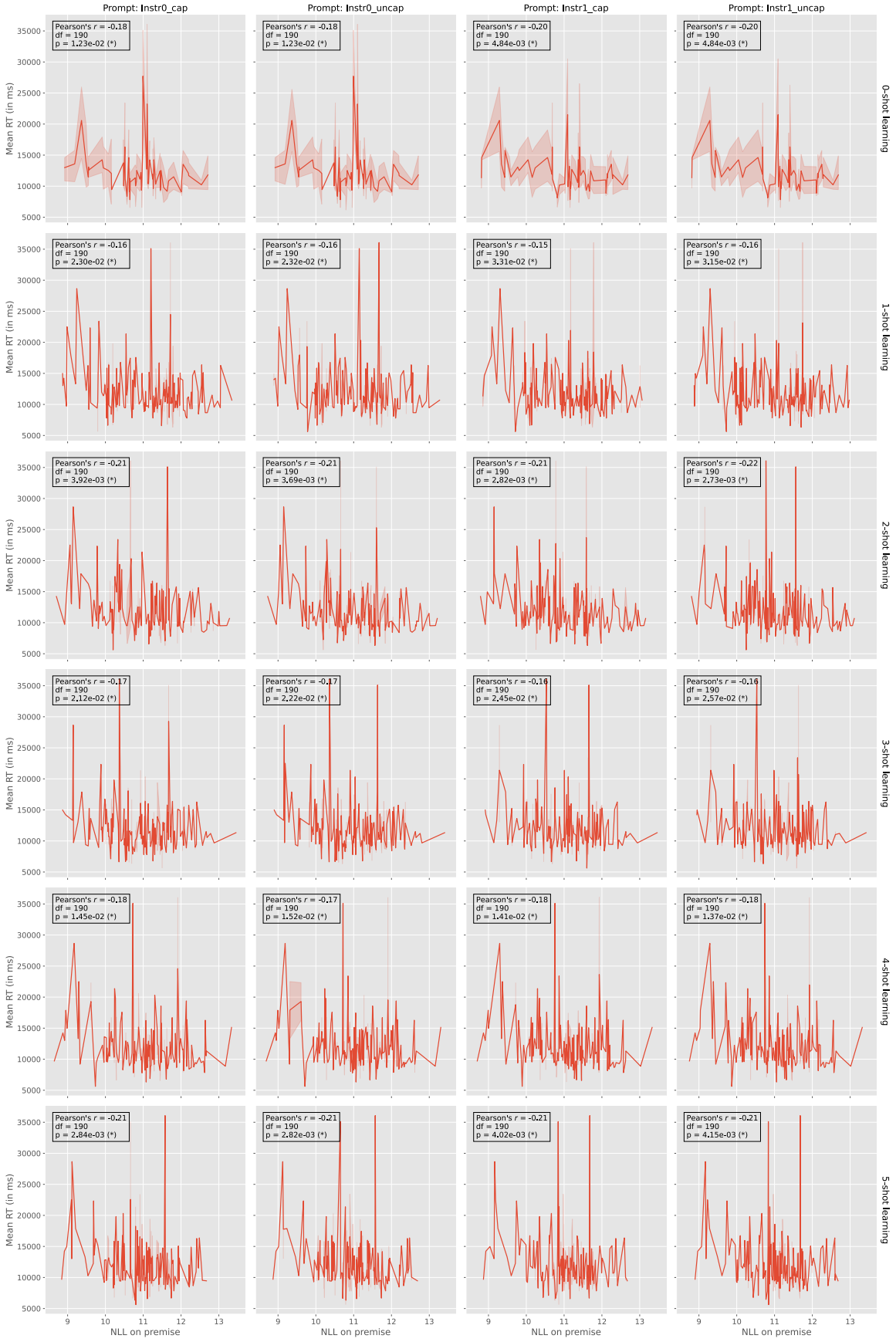


Fig. 5.2. Pearson's correlation between the mean human RT per inference and mean OPT-66B's NLL on the premise for each few-shot training and prompt settings.

We then turned to analyse endorsements rates. We first compared the endorsement rates (averaged across all judgements of each inference) of humans and OPT-66B both with and without DC-PMI correction.

(a) Uncorrected. **(b)** DC-PMI correction.

Fig. 5.3. Pearson's correlation between human and OPT endorsement rate, with and without DC-PMI correction.

We found, each time, a strong correlation between human and OPT-66B's endorsement rates (without DC-PMI correction: $r(190) = .41$, $p < .001$; with DC-PMI correction $r(190) = .41$, $p < .001$; see fig. 5.3).

In order to take this analysis a step further, and to see whether humans and OPT-66B had the same reasoning patterns and biases, we ran 2×2 contingency chi-square tests to probe whether there were significant differences between:

- model endorsement rates with and without DC-PMI correction ("M-M" in fig. 5.4);
- human and model endorsement rates with DC-PMI correction ("H-MDCPMI" in fig. 5.4);
- human and model endorsement rates without DC-PMI correction ("H-M" in fig. 5.4).

The tests were run on the raw data, without averaging over the different judgements of a same inference, and on *all inferences from the dataset* (even those only presented to the model). Bonferroni correction was applied to avoid type I error due to multiple comparisons, and the significance threshold was set to $p < 1.04 \times 10^{-3}$ in order to get an overall alpha level of $\alpha = .05$

As one can see on fig. 5.4, we were never able to reject the null hypothesis for the difference between humans and model (with or without DC-PMI correction) except for MP inferences involving jabberwocky statements, and for all AC inferences. It remained the case even when we did not apply a Bonferroni correction and took the usual significance threshold $p < .05$. In addition, we always found a significant differences between endorsement rates with and without DC-PMI correction.

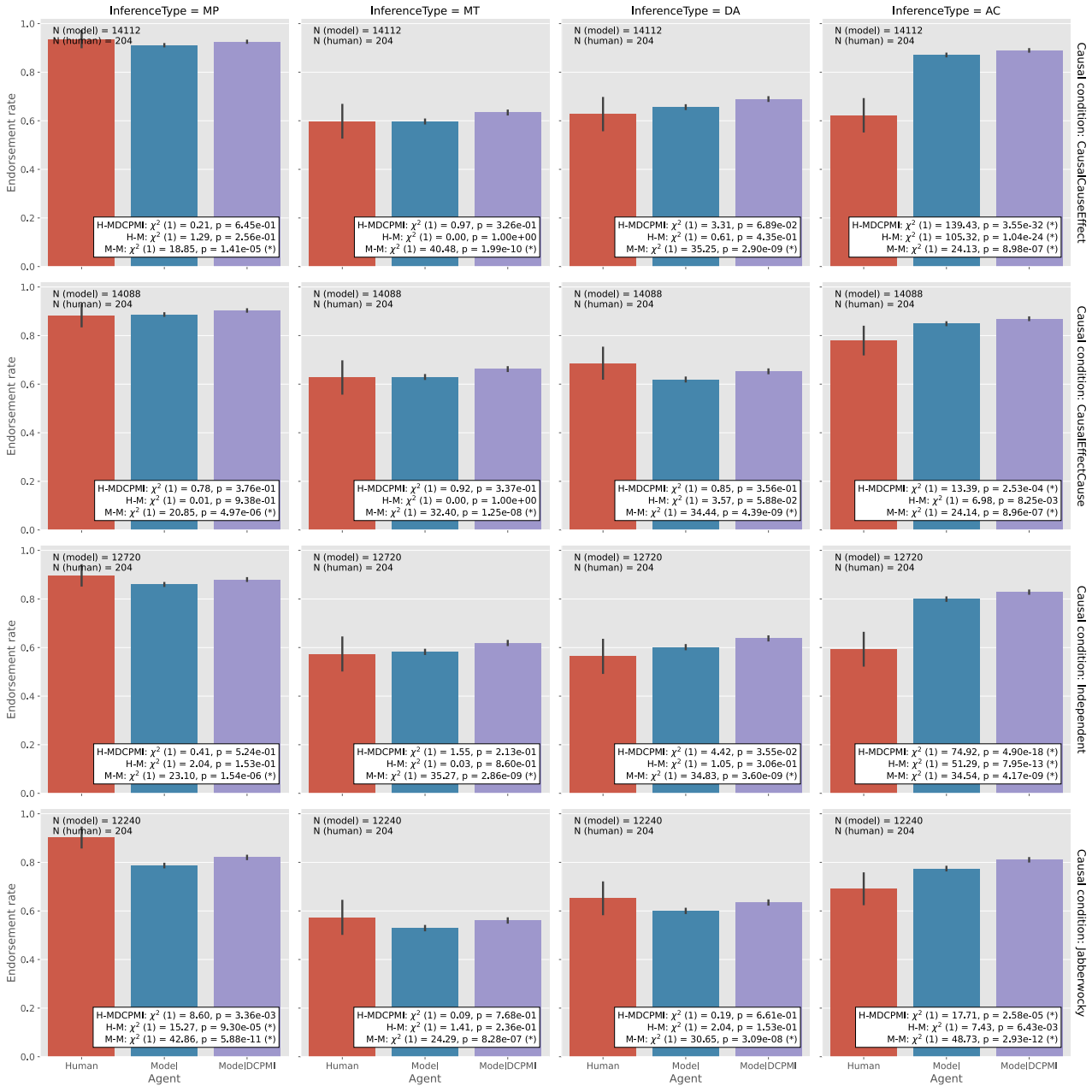


Fig. 5.4. Contingency chi-square tests between human and model endorsement rates, on all inferences without averaging, for each inference type × causal condition setting. Vertical bars show 95% Wald confidence intervals. Because of the multiple comparisons, a Bonferroni correction and the significance threshold should be set to $p < 1.04 \times 10^{-3}$ to get an overall alpha level of $\alpha = .05$.

5.3 Discussion

Contrary to we might have expected given the findings Wilcox et al. (2020, 2021), we only found a weak (though significant) correlation between human RTs and OPT-66B's NLL on the stimuli from our dataset. Suprisingly however, this correlation is negative, meaning that the higher the human RT is, the lower the model surprisal. Again, this puzzling fact could be explored further by running a finer-grained experiment to collect RTs. Indeed, by getting a more precise estimation of the RTs, and getting reading times for each word of the prompt, one would able to fit the same GAM as Wilcox et al. and might obtain results more significant and more interpretable. Another interesting follow-up of this analysis (which would still need to have RTs for individual words) would be to use diffusion decision models (Krajbich & Rangel, 2011; Ratcliff et al., 2016) to model both the reading and the decision time, and only try to correlate NLL with one or the other. Overall, we can only draw very little conclusions for the analyses we ran on RT because of the low size effects (the RT measures are probably very noisy) and the lack of significant results.

Regarding endorsement rates, we found that in all but one category, humans and OPT-66B had the same reasoning patterns and biases with MP, MT and DA inferences. The difference observed for the jabberwocky sentences in MP inferences might be due to the fact that the model sees the syntactic patterns (as proved by its performance elsewhere), but does not really pick up on them outside of its training distribution (the jabberwocky category was introduced so that the model would encouter words for the first time). As for the higher endorsement rate by OPT than by humans for AC inferences, it is probably related to the fact that OPT tends to judge inferences which do not feature negations (i.e. MP and AC) as valid, especially with 0, 1 and 2-shot learning (see section 4.2).

Finally, the systematic significant difference between endorsement rates with and without DC-PMI correction again shows the importance of surface from competition, as highlighted by Holtzman et al. (2022).

CHAPTER 6

GENERAL DISCUSSION

Our investigation of conditional reasoning with causation in large language models went through different stages. We first tried to probe Cummins' (Cummins, 1995) hypothesis that numbers of alternative causes and disablers influence endorsement rates in GPT-3. We then collected a bigger dataset than that provided by Cummins, and assessed OPT-66B's abilities to reason with causal conditionals, both in an absolute sense and as compared to humans.

6.1 Preliminary results

In chapter 2, we presented a probe of Cummins' hypothesis of causation (Cummins et al., 1991; Cummins, 1995) in GPT-3 (Brown et al., 2020). Qualitatively, Cummins' hypothesis on the effect on causality on performance doesn't seem to hold for GPT-3. The model actually does not show any significant content effect, and is largely influenced by prompting and syntax. With zero-shot training, GPT-3 judges all inferences which feature negation invalid and considers the others valid and, as the number of shots of training increases, it learns to classify all inferences as valid. Importantly, GPT-3 performs at chance level at predicting the validity of causal inferences no matter the number of shots of training, meaning that few-shot training does not improve performance contrary to what we might expect (Brown et al., 2020).

This chapter provided the beginning of an answer to the question of whether large language models are able to perform conditional reasoning with causation. However, due to the wide confidence intervals and lack of information about GPT-3's internal representations, we had to use another LLM, namely OPT-66B, and to create a bigger dataset. In order to explore causation further, this dataset would gather conditionals featuring a wide range of causal relations between their antecedents and consequents.

6.2 The Conditional Reasoning dataset

In chapter 3, we presented the online experiments we have run to collect a dataset of conditional statements involving different kinds of relations between the antecedent and the consequent (section 3.1). We then reported on the online experiment designed to collect reaction times (RTs) and endorsement rates for some of the conditionals from our dataset (section 3.2). section 3.2 shows that the results from Cummins et al. (1991) can be extended to the bigger picture: the causal relation between the antecedent and the consequent has a significant effect on RTs. However, contrary to what we could have expected from Cummins' results, endorsement rates or accuracy of prediction are not affected by the causal relation.

We also collected a large dataset of conditionals which can be reused by others.

Once this dataset was created and we made sure that our hypothesis on human reasoning patterns with causal conditionals was verified, we could turn to probe conditional reasoning in OPT-66B, which

is open source and would provide us more details about its inner representations.

6.3 OPT's conditional reasoning abilities

In chapter 4, we presented the assessment of OPT-66B's ability to reason with conditionals from the dataset we collected. We showed that OPT always performs at chance at judging the validity of conditional inferences, no matter the prompt, the number of shots of training and the causal relation between the antecedent and the consequent of the premise.

However, we also found that each of these factors had an effect on two other variables, namely the model's NLL on the correct answer and its certainty. In particular, we showed that the certainty decreases with the causal categories direct causal > independent > jabberwocky.

All these results together show that, although OPT-66B is not able to accurately predict the validity of conditional inferences, it is somewhat sensitive to causation.

The fact that the model's certainty keeps increasing as the number of shots of training increases also suggests that OPT might be picking up on some cues. We thus evaluated whether it was actually picking on the human reasoning patterns – which we know also deviate from logical predictions.

6.4 OPT as a predictor of human behaviour

In chapter 5, we compared the data collected from humans in chapter 3 and from OPT-66B in chapter 5. We showed that there existed a weak negative correlation between human reaction time for an inference and OPT's NLL on the premise. We also found that OPT's and human endorsement rates are very similar, except for inferences of type DA. This seems to prove that, although the model is a poor predictor of inferences' accuracy (see chapter 4), its handling of causal conditionals is somewhat related to that of humans, at least as far as endorsement rates are concerned.

We thus proved that OPT-66B's NLL does not predict human RTs, but that its endorsement rates were similar to those of humans, except for DA and MP jabberwocky inferences.

6.5 Conclusions, limitations and future directions

6.5.1 Conditional reasoning in LLMs and humans: what have we learnt from this project?

In this project, we created a large dataset of conditionals featuring different causal kinds of relations between the antecedent and the consequent. We also verified that human reaction times increase as the causal relation becomes less clear.

We assessed two different state-of-the-art large language models, namely OpenAI's GPT-3 and Meta AI's OPT with 66 billion parameters, and found significant differences between the two. However, both models performed at chance when predicting the validity of causal inferences, even when using few-shot learning. We showed that, nonetheless, OPT is sensitive to causality as its certainty is influenced by the causal relation between the antecendent and the consequent of the premise, in the same way as human reaction times. This gives hope that the model might get a remote sense of what causality is, although it is not able to accurately predict the validity of causal conditional inferences.

Finally, we also proved that, in most inference type × causal condition categories, human and model's endorsement rates were similar, thus starting to build a bridge between human and LLM's handling of causality and conditionals.

6.5.2 Limitations

It must be noted that all the results we presented were obtained from only one large language model (GPT-3 in chapter 2 and OPT-66B in chapters 4 and 5). Therefore, all the claims we made are valid for these models and not for LLMs broadly speaking. In addition, we used OPT-66B in half-precision (in order to only use one computational node on the cluster) and only had access to the free version of

GPT-3 (without scores or hidden states). Thus, the poor performance of these models might be due to these technical factors.

Regarding the Conditional Reasoning dataset, all the data was reviewed manually. This process doesn't come without a margin of error, and there might still be some in the dataset. Besides, since we are not native English speakers, there might be some ill formed sentences in the dataset. Overall though, we believe the dataset is clean and perfectly usable.

Finally, the collection of reactions times for the behavioural experiment from chapter 3 was done online. For this reactions, the RTs are not very precise, as we could not place the participants in a controlled environment and monitor their attention.

6.5.3 Possible continuations

The first continuation of this project would be to compare the endorsement rates with collected from human participants with the literature and see how well they match with previous results.

Some work can also be done to use the dataset we collected more extensively by fine-tuning a model for conditional reasoning and help it gain an insight of the syntactic nature of the conditional inference task. There are many proposals to improve models' training, prompting and even architecture (Lake et al., 2017; Mahowald et al., 2023; Nye et al., 2021) which could make LLMs better at predicted the validity of conditional inferences and pick up more about causality.

Finally, to be able to make a general claim about LLMs, one could use the analysis pipeline we developed to assess different language models and compare them, as we first intended to do.

APPENDIX A
MATHEMATICAL TOOLBOX

A.1 The Transformer architecture

This section briefly presents the Transformer architecture, see Vaswani et al. (2017) for an in-depth description of this architecture and HuggingFace[1] for an online documentation.

The Transformer is a class of neural networks relying solely on a *self-attention* function, which enables it to do computations in parallel – whereas most state-of-the-art neural networks only perform sequential computations because they read their input linearly and adapt their behaviour at time t according to their state at time $t-1$.

A Transformer model consists of two stacks: one for encoding and one for decoding. Each stack consists of N (typically $N = 6$) layers equiped with self-attention mechanisms, and the decoder also performs attention on the output of the encoder stack. The Transformer architecture is depicted on figure fig. A.1 on the following page.

The Transformer architecture is suited for two kinds of language modelling. Causal Language Modelling (CLM), first, consists in analysing a sentence from the left to the right: the model tries to predict the first token, then the second token using the actual first token, and so on. In other words, CLM processeses inputs undirictionally. Masked Language Modelling, on the other hand (MLM), predicts each token given all of the others, whether they be located before or after the token to be predicted. This latter approach is bidirectional: it considers the input as a whole and only masks the token that has to be predicted. An example of a CLM model is GPT-3 and an example of MLM model is BERT.

As OPT was used in the project as a CLM model, the rest of this section only applies to such models. Given a training vocabulary $(w_1, \dots, w_n)$ and a tokenised sequence $X = (x_1, \dots, x_t)$, the output layer of a CLM model returns a score matrix S of size $t \times n$ where $S_{i,j}$ is the score (in an arbitrary metric) assigned to the i-th token being w_j. The model then applies the softmax function to each line of S to transform the scores into a conditional probability distribution $t \mapsto \Pr_\theta(t \mid x_{<i})$. Finally, it computes, for each token x_i, a score $\mathrm{NLL}(x_i)$ called *negative log likelihood* which is defined as

$$\mathrm{NLL}(x_i) := -\log(\Pr_\theta(x_i \mid x_{<i})); \tag{A.1}$$

this score is obtained by applying the negative log loss function $\mathcal{L}\colon x \mapsto -\log(x)$ to the inner product $\left\langle S_i^\top \middle| (\delta_{(w_k = x_i)})_{1 \le k \le n}^\top \right\rangle$ where S_i is the i-th line of S. The final score output by the model, *perplexity*, is defined as the mean of the negative log likelihood over all tokens of the input

$$\mathrm{Perplexity}(X) := \exp\left(-\frac{1}{t}\sum_{i=1}^{t} \mathrm{NLL}(x_i)\right). \tag{A.2}$$

Note that this approach is equivalent to taking the exponentiation of the cross-entropy between the data and model predictions.

[1] `https://huggingface.co/`

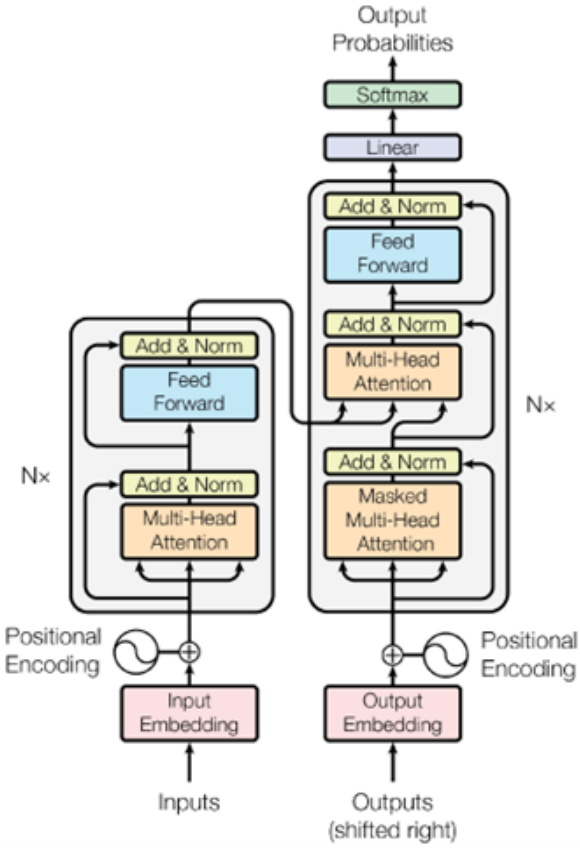
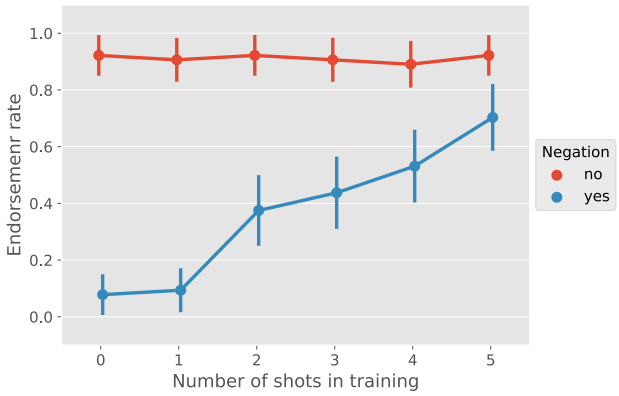


Fig. A.1. Architecture of Transformer models. Extract from Vaswani et al., 2017.

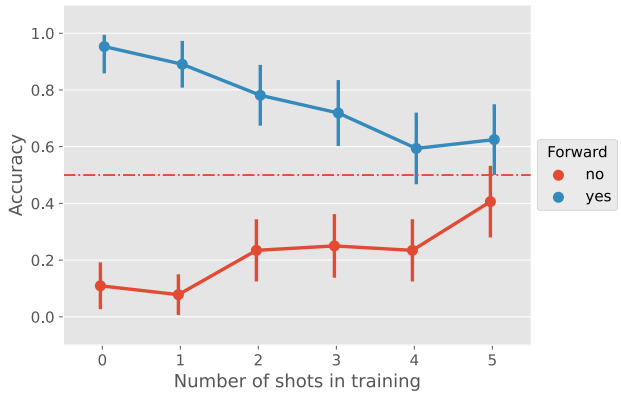
APPENDIX B

SUPPLEMENTARY MATERIAL

B.1 Supplementary figures



(a) GPT-3's predictions against number of shots of training depending on the presence of negation in the inferences.



(b) GPT-3's accuracy against number of shots of training depending on the inferences' direction (forward or backward).

Fig. B.1. Syntactic effects on GPT-3's predictions and accuracy on the stimuli from Cummins (1995). Horizontal lines show chance level and vertical bars are 95% Wald confidence intervals.

B.2 Statistical tests reports

Independent Variable	Estimate	Standard Error	Odds Ratio	p-value	Significance
Number of training shots	−0.382	0.19	0.68	0.043	*
MT > MP	−6.52	0.82	0.001	<0.001	***
DA > MP	−1.82	0.78	0.16	0.019	*
AC > MP	−5.40	0.80	0.004	<0.001	***
MT × number of shots	1.10	0.22	3.02	<0.001	***
DA × number of shots	−0.212	0.22	0.81	0.33	
AC × number of shots	0.186	0.23	1.20	0.423	
Intercept	3.75	0.70	42.72	<0.001	***

Table B.1. Logistic regression analysis of GPT-3's accuracy on Cummins' stimuli with independent variables number of shots of training, inference type (baseline is MP) and their interaction. Pseudo $R^2 = .40$, $\chi^2 = 427.21$, $p < .001$. Significance code: $^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$. Odds ratios should be interpreted with other independent variables equal to baseline, see eq. (B.1).

Independent Variable	Estimate	Standard Error	Odds Ratio	p-value	Significance
Number of training shots	−0.0782	0.078	0.92	0.31	
MT > MP	−3.98	0.56	0.02	<0.001	***
DA > MP	−2.81	0.55	0.06	<0.001	***
AC > MP	−5.20	0.60	0.005	<0.001	***
EffectCause > CauseEffect	−0.972	0.68	0.38	0.15	
Direction × number of shots	0.0396	0.11	1.04	0.71	
Direction × MT	−0.0782	0.69	0.92	0.11	
Direction × DA	0.915	0.69	2.50	0.18	
Direction × AC	0.771	0.77	2.16	0.32	
Intercept	3.34	0.55	28.20	<0.001	***

Table B.2. Logistic regression analysis of GPT-3's accuracy on Cummins' stimuli with independent variables causal direction of hypothesis (baseline is "If [cause], then [effect]."), and number of shots of training, inference type (baseline is MP) and their interaction with causal direction of hypothesis. Pseudo $R^2 = .32$, $\chi^2 = 337.61$, $p < .001$. Significance code: $^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$. Odds ratios should be interpreted with other independent variables equal to baseline, see eq. (B.2).

Independent Variable	Estimate	Standard Error	Odds Ratio	p-value	Significance
Number of training shots	−0.0221	0.11	0.98	0.83	
With > without negation	−4.68	0.41	0.01	<0.001	***
Negation × number of shots	0.667	0.13	1.95	<0.001	***
Intercept	2.39	0.32	10.89	<0.001	***

Table B.3. Logistic regression analysis of GPT-3's predicted validity of Cummins' stimuli with independent variables presence of negation in the inference (baseline is absence of negation), number of shots of training and their interaction. Pseudo $R^2 = .35$, $\chi^2 = 353.01$, $p < .001$. Significance code: $^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$. Odds ratios should be interpreted with other independent variables equal to baseline, see eq. (B.3).

Independent Variable	Estimate	Standard Error	Odds Ratio	p-value	Significance
Number of training shots	−0.443	0.080	0.64	<0.001	***
Backward > Forward	−4.65	0.39	0.01	<0.001	***
Direction × number of shots	0.791	0.11	2.21	<0.001	***
Intercept	2.408	0.28	11.11	<0.001	***

Table B.4. Logistic regression analysis of GPT-3's accuracy on Cummins' stimuli with independent variables direction of inference (forward for MP and DA, backward for MT and AC; baseline is forward), number of shots of training and their interaction. Pseudo $R^2 = .28$, $\chi^2 = 294.55$, $p < .001$. Significance code: $^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$. Odds ratios should be interpreted with other independent variables equal to baseline, see eq. (B.4).

Independent Variable	Estimate	Standard Error	Odds Ratio	p-value	Significance
Number of training shots	−0.594	0.006	0.55	<0.001	***
Negation > No negation	−0.804	0.030	0.45	<0.001	***
Negation × number of shots	−0.243	0.008	0.78	<0.001	***
Intercept	3.508	0.024	33.38	<0.001	***

Table B.5. Logistic regression analysis of OPT-66B's endorsement rate with independent variables presence of negation in the inference (baseline is no negation), number of shots of training and their interaction. Pseudo $R^2 = .28$, $\chi^2 = 294.55$, $p < .001$. Significance code: $^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$. Odds ratios should be interpreted with other independent variables equal to baseline, see eq. (B.6).

Independent Variable	Estimate	Standard Error	Odds Ratio	p-value	Significance
MT > MP	−1.11	0.044	0.36	<0.001	***
DA > MP	−0.94	0.045	0.43	<0.001	***
AC > MP	−0.41	0.049	0.73	<0.001	***
Number of training shots	−0.61	0.010	0.54	<0.001	***
MT × number of shots	−0.23	0.012	0.81	<0.001	***
DA × number of shots	−0.22	0.012	0.82	<0.001	***
AC × number of shots	0.024	0.013	1.05	0.054	
Intercept	3.73	0.037	41.88	<0.001	***

Table B.6. Logistic regression analysis of OPT-66B's endorsement rate with independent variables inference type (baseline is no MP), number of shots of training and their interaction. Pseudo $R^2 = .25$, $\chi^2 = 63605.74$, $p < .001$. Significance code: $^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$. Odds ratios should be interpreted with other independent variables equal to baseline, see eq. (B.7).

Independent Variable	Estimate	Standard Error	Odds Ratio	p-value	Significance
Causal eff–c > Causal c–eff	0.063	0.012	1.07	<0.001	***
Independent > Causal c–eff	0.053	0.012	1.05	<0.001	***
Jabberwocky > Causal c–eff	−0.044	0.012	0.96	<0.001	***
Intercept	−0.020	0.008	0.98	0.018	*

Table B.7. Logistic regression analysis of OPT-66B's accuracy with independent variable causal condition of the premise (baseline is causal cause–effect). Pseudo $R^2 = 0$, $\chi^2 = 96.25$, $p < .001$. Significance code: $^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$. Odds ratios should be interpreted with other independent variables equal to baseline, see eq. (B.8).

B.3 Statistical models' equations

Hereafter, we denote by logit the following function:

$$\text{logit}: \left| \begin{array}{ccc} [0,1] & \longrightarrow & \mathbf{R} \\ x & \longmapsto & \ln\left(\dfrac{x}{1-x}\right) \end{array} \right. .$$

- Equation of the logistic regression reported in table B.1:

$$\begin{aligned}\text{logit}\,\Pr(\text{Accurate}) = \beta_0 + \beta_1\delta_{\text{MT}} + \beta_2\delta_{\text{DA}} + \beta_3\delta_{\text{AC}} + \beta_4\text{nbShots}+ \\ \beta_5(\delta_{\text{MT}} \times \text{nbShots}) + \beta_6(\delta_{\text{DA}} \times \text{nbShots}) + \beta_7(\delta_{\text{AC}} \times \text{nbShots}).\end{aligned} \quad \text{(B.1)}$$

- Equation of the logistic regression reported in table B.2:

$$\begin{aligned}\text{logit}\,\Pr(\text{Accurate}) = \beta_0 + \beta_1\delta_{\text{EffectCause}} + \beta_2\delta_{\text{MT}} + \beta_3\delta_{\text{DA}} + \beta_4\delta_{\text{AC}} + \beta_5\text{nbShots}+ \\ \beta_6(\delta_{\text{EffectCause}} \times \text{nbShots}) + \beta_7(\delta_{\text{EffectCause}} \times \delta_{\text{MT}}) + \beta_8(\delta_{\text{EffectCause}} \times \delta_{\text{DA}}) + \beta_9(\delta_{\text{EffectCause}} \times \delta_{\text{AC}}).\end{aligned} \quad \text{(B.2)}$$

- Equation of the logistic regression reported in table B.3:

$$\text{logit}\,\Pr(\text{Endorse}) = \beta_0 + \beta_1\delta_{\text{Negation}} + \beta_2\text{nbShots} + \beta_3(\delta_{\text{Negation}} \times \text{nbShots}). \quad \text{(B.3)}$$

- Equation of the logistic regression reported in table B.4:

$$\text{logit}\,\Pr(\text{Accurate}) = \beta_0 + \beta_1\delta_{\text{Backward}} + \beta_2\text{nbShots} + \beta_3(\delta_{\text{Backward}} \times \text{nbShots}). \quad \text{(B.4)}$$

- Equation of the logistic regression reported in table 2.1:

$$\text{logit}\,\Pr(\text{Accurate}) = \beta_0 + \beta_1\text{nbShots}. \quad \text{(B.5)}$$

- Equation of the logistic regression reported in table B.5:

$$\text{logit}\,\Pr(\text{Endorse}) = \beta_0 + \beta_1\delta_{\text{Negation}} + \beta_2\text{nbShots} + \beta_3\delta_{\text{Negation}} \times \text{nbShots}. \quad \text{(B.6)}$$

- Equation of the logistic regression reported in table B.6:

$$\begin{aligned}\text{logit}\,\Pr(\text{Endorse}) = \beta_0 + \beta_1\delta_{\text{MT}} + \beta_2\delta_{\text{DA}} + \beta_3\delta_{\text{AC}} + \beta_4\text{nbShots}+ \\ \beta_5(\text{nbShots} \times \delta_{\text{MT}}) + \beta_6(\text{nbShots} \times \delta_{\text{DA}}) + \beta_7(\text{nbShots} \times \delta_{\text{AC}}).\end{aligned} \quad \text{(B.7)}$$

- Equation of the logistic regression reported in table B.7:

$$\text{logit}\,\Pr(\text{Accurate}) = \beta_0 + \beta_1\delta_{\text{Causal effect–cause}} + \beta_2\delta_{\text{Independent}} + \beta_3\delta_{\text{Jabberwocky}}. \quad \text{(B.8)}$$

APPENDIX C

LIST OF SENTENCES

C.1 Cummins' sentences

C.1.1 Few disabling conditions, few alternative causes

- If Joe cut his finger, then it bled.
- If Larry grasped the glass with his bare hands, then his fingerprints were on it.
- If the gong was struck, then it sounded.
- If the doorbell was pushed, then it will ring.
- If Joe's finger bled, then he cut his finger.
- If Larry's fingerprints were on the glass, then he grasped it with his bare hands.
- If the gong sounded, then it was struck.
- If the doorbell will ring, then it was pushed.

C.1.2 Many disabling conditions, few alternative causes

- If the trigger was pulled, the the gun fired.
- If the correct switch was flipped, then the porch light went on.
- If the ignition key was turned, then the car started.
- If the match was struck, then it lit.
- If the gun fired, then the trigger was pulled.
- If the porch light went on, then the correct switch was flipped.
- If the car started, then the ignition key was turned.
- If the match lit, then it was struck.

C.1.3 Few disabling conditions, many alternative causes

- If Alvin read without his glasses, the he got a headache.
- If Mary jumped into the swimming pool, the she got wet.
- If the apples were ripe, then they fell from the tree.
- If water was poured on the campfire, then the fire went out.
- If Alvin got a headache, then he read without his glasses.
- If Mary got wet, then she jumped into the swimming pool.
- If the apples fell from the tree, then they were ripe.
- If the campfire went out, then water was poured on the fire.

C.1.4 Many disabling conditions, many alternative causes

- If fertilizer was put on the plants, then they grew quickly.
- If the brake was depressed, then the car slowed down.

- If John studies hard, then he did well on the test.
- If Jenny turned on the air conditioner, the she felt cool.
- If the plants grew quickly, then fertilizer was put on them.
- If the car slowed down, then the brake was depressed.
- If John did well on the test, then he studied hard.
- If Jenny felt cool, then she turned on the air conditioner.

C.2 Conditional Reasoning Dataset sentences

C.2.1 Causal ("If [cause], then [effect]." version)

- If I study, then I should do well in my exam.
- If Alice went for a walk in the rain, then she is wet.
- If Sara was drinking a lot of alcohol, then she is drunk.
- If Paul is chopping vegetables, then he is using a knife.
- If Mary eats enough food, then her hunger is satisfied.
- If Simon takes his clothes off, then he will be naked.
- If John doesn't pay his power bill, then the lights are turned off.
- If there is no power, then the kettle won't work.
- If Jen eats a banana, then she is full.
- If the sun is shining, then it is warm.
- If John is in the swimming pool, then he will get wet.
- If Antony is at work, then he is earning money.
- If John's alarm doesn't go off, then he will be late for work.
- If the sun is shining, then John is warm.
- If John hangs up his washing, then it will dry.
- If the kettle had boiled, then the water was hot.
- If Harry wakes up late for work, then he is in trouble.
- If John spills hot water on himself, then he burns himself.
- If Sarah does not study for her test, then she will fail.
- If the car engine is on, then the car uses fuel.
- If Sarah is in the desert, then she is hot.
- If Peter is sunbathing, then he is relaxed.
- If John stands in the rain, then he is wet.
- If the little girl gets ice cream, then she is happy.
- If Peter is swimming while drunk, then he will drown.
- If John is sunbathing for ages, then he gets sunburn.
- If Mary cuts her finger, then she is bleeding.
- If Mary is in the classroom, then she is learning.
- If Emma drops her glass, then it will break.
- If Annie is cooking breakfast, then she is hungry.
- If Sarah continues to eat even when full, then she will become obese.
- If John has eaten lunch, then he is full.
- If Sarah runs daily, then she is healthy.
- If Jack is running to the store, then he will be tired.
- If John is cooking, then he is hungry.
- If Peter walks in the rain without a coat, then he is wet.
- If John is asleep, then he is snoring.
- If Joe is eating, then he is full.
- If Amanda graduated university, then she has a degree.
- If Pete's team scores the most goals, then they will win the football match.
- If it is raining outside, then there is water on the street.
- If John leaves his heating turned off, then he will be cold.

- If Stacey sleeps through her alarm, then she will be late to work.
- If Adrian showers, then he will be clean.
- If John drives too fast, then he will wreck his vehicle.
- If it is sunny out, then flowers will grow.
- If Joe touches a flame, then he burns himself.
- If Sally had known John was coming, then Sally would have baked a cake.
- If the the mug falls off the table, then the drink will spill.
- If John takes medicine, then he will get better.
- If the sun explodes now, then we will all cease to exist.
- If Patrick does not put baking powder in the cake batter, then the cake won't rise.
- If the sun has gone down, then it is night time.
- If the driver is in traffic, then he is hooting his car horn.
- If John stubbs his toes, then he is hurt.
- If there is a knock at the door, then the dog barks.
- If James fought at school, then he will be suspended.
- If Claire didn't get enough sleep, then she is tired.
- If the car brakes aren't working, then we will crash.
- If Tom drinks too much beer, then he has a headache.
- If Tiddles the cat is happy, then he is purring.
- If Cedric practices soccer everyday, then he is one of the best players in the team.
- If Mary was drinking alcohol, then she is drunk.
- If Mary is in the shower, then she is wet.
- If Marcus breaks the rules, then he will get in trouble.
- If Samantha is studying late into the night, then she will be sleepy.
- If Susan put her hand in the fire, then she is burnt.
- If John is running, then he is panting.
- If Tom is playing, then he is happy.
- If I plant seeds, then there will be fruit.
- If the girl needed glasses, then she couldn't see.
- If the bike gets a puncture, then I can't ride it.
- If Darren made a very funny joke, then everyone is laughing.
- If Mary had a serious car accident, then she is in intensive care.
- If the hairdresser doesn't show up for work, then the girl won't be able to get a haircut.
- If Christina watches a movie, then she is entertained.
- If John walks in the rain, then he needs an umbrella.
- If a fly lands on a venus fly trap, then it is eaten.
- If Xanthe has an infection, then she is taking antibiotics.
- If John goes for a jog, then he is exercising.
- If John stays up all night, then he will be tired tomorrow.
- If John hasn't eaten all day, then he will be hungry.
- If Amy gets bullied everyday for ten years, then she will need therapy.
- If Jamie is running, then he is burning calories.
- If Jill woke up late this morning, then she is running late.
- If Tino is in the garden, then his hands are muddy.
- If it rains, then the ground will get wet.
- If Marry is in the garden, then she is muddy.
- If John turns on the washing machine, then his clothes will be cleaned.
- If the sun is shining, then it feels warm.
- If Mary eats McDonalds everyday for a year, then she will become fat.
- If John drinks lots of soy sauce, then he will develop hypernatremia.
- If Sue eats too many calories, then she puts on weight.
- If Sandra applies hair growth oil to her hair daily, then she has long healthy hair.
- If Mary mary drank alcohol last night, then she has hangover.

- If the spoon fell in the mud, then it is dirty.
- If Carrie doesn't water the plants, then the plants will die.
- If Bob is salsa dancing, then he is sweaty.
- If John is running, then he is sweating.
- If Kate does not switch on the light at night, then she is in the dark.
- If John is sun bathing, then he is sweating.
- If Katie is tired, then she goes to sleep.
- If Danny is late, then he will miss the train.
- If Danny worked for the government, then he got paid.
- If Anne is gardening, then she is outside.
- If John renewed his passport, then he is able to travel outside of his own country.
- If the dog barked, then the child cries.
- If Nate stands next to the heater, then he is hot.
- If John doesn't grocery shop, then he will have nothing to eat for dinner.
- If I drop an open bottle of milk, then the milk will spill.
- If Tim steps on thin ice, then the ice breaks.
- If Clive is in the pub, then he is socialising.
- If Brannan is crying, then his mother will comfort him.
- If John left the house late, then he is late to work.
- If John doesn't put petrol in his car, then his car won't work.
- If the children had been playing for hours, then they were tired.
- If I eat too many kebabs, then I will become obese.
- If John is caught outside in the rain without an umbrella, then he is wet.
- If Angel is late to the bus stop, then the bus has left her.
- If John turns on the heating, then he is warm.
- If John eats his dinner, then he will be full.
- If it is snowing, then it is winter.
- If John gets caught cheating, then he gets punished in some way.
- If Bob is hungry, then he should eat.
- If Morgan is scared of the dark, then he refuses to go in without any light.
- If Kevin puts his hand on the induction burner stove after use, then he will be burnt.
- If Spot the dog dropped his ball in the sea, then Spot's ball is lost.
- If Cedric does not stretch after training, then he suffers from muscle cramps.
- If Alexander was driving at night, then he is sleep-deprived.
- If the steak is on the grill, then it is hot.
- If Gary reads twenty pages a day, then he will finish thirty books a year.
- If Marlin rides his bike everyday, then he is fit.
- If John is lying down, then he is tired.
- If Peter does not turn off the bath taps, then he is flooding the room.
- If John is drinking beer, then he is getting drunk.
- If Leah is thirsty, then she has a drink.
- If Max fails the exam, then he can't get into the college he wants.
- If the cat starts to run, then the dog will chase the cat.
- If John doesn't feed his goldfish, then it will die.
- If a stranger walks into my yard, then my dog will bark.
- If Jane presses a button on the remote, then the TV turns on.
- If Harry is famous, then he is being followed by photographers.
- If John studies a lot, then he will get high grades.
- If John watches a sad movie, then he will become emotional.
- If Fred likes animals, then he bought a pet.
- If my cat doesn't eat well, then she will lose weight.
- If the clocks go back an hour, then the days feel longer.
- If Jane forgets to pay her internet account, then the account gets closed.

- If the car's wheels stop turning, then the car stops.
- If John closes his eyes, then he cannot see.
- If it snows, then the road will be slippery.
- If the washing machine is operating, then the clothes are wet.
- If Joseph is speeding, then he will crash.
- If Katrin catches rabies, then she will die.
- If Edmond has a bad eyesight, then he needs his glasses to read.
- If Sally doesn't get here by 3pm, then I won't be going out with her.
- If Karen skips breakfast, then she doesn't focus well at work.
- If Mick has the television on really loud, then he is going deaf.
- If Cindy does not study after school, then she does not pass her tests.
- If John said something stupid, then he is embarrassed.
- If the chicken is in the freezer, then it is cold.
- If Leanne puts out bird seed, then the birds will visit the garden.
- If Alicia rolled around in the snow, then she is cold.
- If John is hearing a joke, then he is laughing.
- If Kate walks in the sun, then she is hot.
- If John is eating, then he is getting full.
- If Leo is at the beach, then he is sandy.
- If I go running regularly, then I will get fit.
- If the walls have just been painted, then they are wet.
- If John goes to college, then he is educated to a higher level.
- If the shop is closed, then the children aren't able to buy sweets.
- If John walks through the desert, then he is thirsty.
- If my alarm sounds, then I will wake up.
- If Kyle eats moudly bread, then he gets sick.
- If Norah sprayed some perfume, then she can smell it.
- If it is spring, then flowers bloom.
- If John trains his dog, then it will do tricks.
- If Bill was stuck in traffic, then he was late for work.
- If Jennifer falls from a fifty-metre building, then she will break several bones and die.
- If the river banks burst, then there could be a flood.
- If there is a full moon, then the night is bright.
- If the woman drinks a lot, then she will need to go to the toilet.
- If Jill is on her treadmill, then she is running.
- If John ingests poison, then he will be sick.
- If John turned his ignition on, then his car started.
- If the little boy is happy, then he will smile.
- If Luke gambles, then he will forfeit is inherence.
- If Peter owns a mobile phone, then he has to charge it.
- If Doug the dog goes out and plays in the mud puddles, then he will get mud everywhere in the house.
- If Sam doesn't wear a coat in the snow, then he is cold.
- If Sarah has fallen in some mud, then she is dirty.
- If Sarah stays in the sun too long, then she gets sunburnt.
- If Andrew has washed himself, then he is clean.
- If the clothes are in the washing machine, then they are clean.
- If Steve saves £10 a day, then he will have £1800 in 6 months.
- If Yannis eats junk food daily, then he is overweight.
- If John is driving, then he is concentrating.
- If John is falling, then he is scared.
- If the food has just come out of the oven, then it is hot.
- If Anne is sick, then she is coughing.

- If Mary is in the bath, then she is wet.
- If the snow is heavy, then the children won't be able to go to school.
- If Alexandra works out, then she is tired.
- If I over cook food, then it will be burnt.
- If Pavil flicks the light switch, then light turns on.
- If Adele is in the park with her parents, then she is playing.
- If Kai the cat is ready to pounce, then he is stalking some prey.
- If John runs every day, then he will become healthier.
- If Emma was stood next to the loudspeaker, then her ears hurt.
- If Becky dives in the swimming pool, then she splashes other people in the pool.
- If Mary is unsure on a task, then she asks for help.
- If John throws a glass on the floor, then the glass shatters into many pieces.
- If John is cold, then he is shivering.
- If John turned his microwave on, then his food cooked.
- If a Pete is rich, then he can buy what they want.
- If Matthew studies hard for the test, then he will get an A+.
- If the car travels too fast around a corner, then it loses its grip on the road and crashes.
- If Karen is a nurse, then she helps sick people.
- If the floor is wet, then Dona slips on the floor.
- If Charlie is eating the last ice cream, then there is no more left.
- If the piano key is struck harder, then the sound is louder.
- If Lizzie is hungry, then she is getting impatient waiting for her food.
- If Claudia went shopping on foot, then she is tired.
- If the laundry is on the clothes line, then they are dry.
- If Susie feeds the pigeons every day, then the pigeons will become tame.
- If Selene sits in the sauna after the gym, then she is hot.
- If john is working, then he is getting tired.
- If the roof of the house is fixed, then it won't leak.
- If Paul is in the Arctic, then he is cold.
- If John goes to the gym everyday, then his personal fitness improves.
- If there is traffic on the road, then Dylan will be late home.
- If Samuel drinks water, then he is hydrated.
- If I don't wait for a green light to cross the road, then I will get hit by a moving car.
- If Erin goes outside in the winter without a coat, then she is cold.
- If Antony's phone is ringing, then he needs to answer it.
- If it is summer, then it will be sunny.
- If John drops a glass, then it will break.
- If James had a sore throat, then he went to the doctor.
- If Lily procrastinates from work by playing games, then she won't get much work done.
- If John has gone out for a run, then he is excericising.
- If John is poorly, then he will go to the doctors.
- If John is going to a party, then he is nervous.
- If John is very old, then he is highly wrinkled.
- If John took paracetamol, then his headache went.
- If the little girl falls down, then she will get hurt.
- If Anderson sleeps late, then he will wake up late.
- If John wears nothing in Antarctica, then he freezes to death.
- If the wind is strong, then it causes the windows to slam shut.
- If James is standing at the busstop, then he is waiting for a bus.
- If Bob doesn't wash his clothes, then they start to smell.
- If Jan has a high temperature, then she is feeling poorly.
- If Jane is drinking wine, then she is drunk.
- If Victor watches horror films, then he is scared.

- If John is digging, then he is getting dirty.
- If Toby is in the bath, then he is clean.
- If the dog is on the lead, then it won't run away.
- If Mary goes to sleep earlier than usual, then she feels much more rested.
- If the sun is shining, then the snow will melt.
- If I lock the door, then nobody can enter.
- If Jordan wants to cross the road, then he looks and listens for vehicles.
- If the tea mug is left unwashed for a week, then it is moudly.
- If Nick drinks coffee at night, then he won't sleep well.
- If it is winter, then it will be cold.
- If John undercooks chicken, then he will get sick.
- If Paul was freezing, then he zipped up his coat.
- If a meteor with the circumference akin to that of London falls into France, then many people will lose their lives.
- If Paul has eaten too much food, then he is feeling very full.
- If Etienne is travelling in a strange place with no map, then he gets lost.
- If John walks, then he will move.
- If Kevin is with his children, then he is having fun.
- If John has broken his leg, then he cannot run.
- If John pressed his laptop's on/off button, then his laptop turned on.
- If the light switch is turned off, then the light will go off.
- If Matthew drinks vodka, then he will wake up with a hangover.
- If John flies to the Moon, then he is weightless most of the time.
- If Linda runs every day, then she gets fitter.
- If Fido the dog is bored, then he eats my shoes.
- If Danny studied mechanical engineering, then he knows how to fix cars.
- If John does not have a job, then he is broke.
- If he is worried, then John is shaking.
- If Heather jumps in a muddy puddle, then she gets dirty.
- If I take my bins out, then my rubbish will get taken.
- If I am out when the delivery driver knocks, then he can't deliver my parcel.
- If John completes a Prolific survey, then he is rewarded.
- If iIt is raining outside, then puddles will form.
- If Mary jumps off a wall, then she will hit the ground.
- If the cat falls in the bathtub, then it is wet.
- If Sharon is outside, then she is gardening.
- If the photo frame fell off the wall, then the glass broke.
- If John doesn't budget, then he will run out of money.
- If Simon heard some bad news, then he cried.
- If Ben spends 24 hours playing on his playstation, then his body will feel sore.
- If Sarah adopted a cat, then there is cat hair everywhere at Sarah's place.
- If John stands in snow, then he is cold and wet.
- If John puts his foot on the brakes, then the car will stop.
- If John eats too many sweets, then his teeth decay faster.
- If Patricia was far away, then Andy had to scream to get Patricia's attention.
- If it is snowing heavily, then we can make a snowman.
- If the ice cream isn't eaten quickly, then it melts.
- If Cedric puts on sunscreen, then he is protected from the sun's harmful rays.
- If Thomas watches sad films, then he is crying.
- If John is playing a game, then he is happy.
- If Joe has a blanket, then he is warm.
- If John goes out in the rain, then he will get wet.
- If John does not eat his dinner, then he will be hungry.

- If John has an argument with his friend, then his friend will be angry.
- If John does not make his bed in the morning, then he will be annoyed that his bed is not made at night.
- If John needs a haircut, then he will have to go to the barbers.
- If John's friends call round, then he will have to go out with them.
- If Samuel makes an experiment, then he will find interesting results.
- If the pedestrian traffic light is green, then I will cross the road.
- If John started the fire, then the house is burning.
- If Janet is in the supermarket, then she is buying groceries.
- If Sam is near a liquor bottle, then he will drink it.
- If Mary has a job, then she gets paid a wage.
- If John gets water in his ear, then he will develop an earache.
- If Harry is running, then he has an elevated heart rate.
- If Dave is yawning, then he is tired.
- If James wears a seatbelt, then he will be safe.
- If John is eating, then he is hungry.
- If the sun is shining, then Sarah gets sunburnt.
- If elephants have become extinct, then there are no elephants in Africa.
- If John is not in the swimming pool, then he is not wet.
- If Harry is tired, then he will sleep.
- If Paul is at work, then he is busy.
- If Chloe is listening to music, then she is dancing.
- If David is a pilot, then he can fly planes.
- If John were to be sick, then he wouldn't go into the swimming pool.
- If Fred is walking in the rain, then he's wet.
- If Sarah turns off the light, then the children will be in darkness.
- If we raise interest rates, then inflation will slow down.
- If the cat is dead, then the cat cannot climb the tree.
- If John swims in the pool, then he is exercising.
- If it is sunny outside, then I will wear sunglasses.
- If Cindy studied, then she got good grades.
- If the house is abandoned, then it will become decrepit.
- If Josh's car's battery is flat, then the car won't start.
- If a vehicle is a dark colour, then it absorbs more light.
- If Will is working, then he's earning money.
- If the waiter drops the food, then he will be in trouble.
- If we raise taxes, then we can build a new hospital.
- If John does not swim in the pool, then he is not exercising.
- If it is raining outside, then I will get wet.
- If Rob slept all night, then he is well rested.
- If the fire alarm sounds, then people will leave the building.
- If Sara is in the desert, then she is hot.
- If Harry sits in the sun too long, then he gets sunburnt.
- If the water is too cold, then John would not enjoy swimming in the pool.
- If global warming continues, then sea levels will become intolerable.
- If the driver parks there, then he will receive a parking ticket.
- If the laptop is dead, then it needs charging.
- If the movie ends, then we leave the cinema.
- If the book has blank pages, then the book cannot be read.
- If Don decided to wash the car, then the car is super clean.
- If Louise wins the lotto, then she will be rich.
- If the gardener mows the lawn, then the grass is short.
- If the water reached a perfect temperature, then it would make John very happy.

- If Lisa inhales a lot of carbon monoxide, then she will get sick.
- If the lamp is on, then the room is bright.
- If there is an earthquake, then the building collapses.
- If Sarah won't eat, then she will be hungry.
- If Peggy decided to cut starch off her diet, then she lost a lot of weight.
- If Ryan goes for a run every day, then he is fit.
- If students study, then they will pass the exam.
- If John splashes too much, then others in the pool won't enjoy themselves.
- If John wants to use the train, then he will buy a ticket.
- If the fox enters the henhouse, then it will kill the chickens.
- If the moon rotates around the earth, then the tides change.
- If boxes are made of dry cardboard, then they can be set on fire.
- If Lucy needs food, then she will go to the shop.
- If Tom decided to run faster, then he beat his usual time.
- If Debbie gets a dog, then she will get lots of exercise.
- If John won the lottery, then he is rich.
- If Chris locks the front door, then the door won't open.
- If the wind is strong, then wind turbines are more effective.
- If the pub runs out of beer, then the custmers will be unhappy.
- If Mina waters the plant, then the plant grows.
- If Pam decided to clean her house, then the house is clean.
- If the boy kicks the ball in the air, then it will bounce.
- If It snows a lot, then school is cancelled.
- If the tree is cut down, then sun now reaches the ground around the stump.
- If Lee needs to go to a doctor, then he will make an appointment.
- If Harold waits too long, then all of the tickets will be gone.
- If Zhuoye releases a new album, then he makes money.
- If John is thirsty, then he will drink.
- If It is raining outside, then the grass is wet.
- If snow falls heavily overnight, then the schools in area will be closed the next day.
- If you don't brush your teeth, then you will get a cavity.
- If the bike moves forwards, then its wheels turn.
- If John sees his crush at the swimming pool, then he will most definitely be blushing.
- If the quotation is too high, then no one will wish to buy the watch.
- If it is raining, then the ground is wet.
- If the enclosed room has no lighting or windows, then the enclosed room is dark inside.
- If Ethan needs money, then he will work.
- If Kevin buys fewer takeaway coffees, then he will save some money.
- If you don't use protection, then you will get pregnant.
- If the cat is stroked, then the cat purrs.
- If Hannah doesn't eat, then she will starve.
- If Barrie wants a bigger house, then he will need to save his money.
- If the fruit is left too long, then it will become rotten.
- If Jong watching a horror film, then he is scared.
- If fire is hot, then people stay away from fires.
- If the pavement is dry, then no one slips.
- If the bus is late, then I came late to work.
- If the pavement is icy, then people slip.
- If the field is parched, then the crop will fail.
- If a large meteor hits Earth, then humans will go extinct.
- If the children continually eat too much, then they will become obese.
- If Fido the dog becomes unwell from eating peanuts, then the dog should not be given peanuts.
- If Sarah is thisty, then she gets a drink.

- If the IT systems are down, then Kevin can't do any work.
- If the sea rises, then coastal communities are at risk.
- If the roof is not repaired, then water damage will result.
- If prices are up, then motorists are complaining.
- If Charles smokes, then his lungs are damaged.
- If the wound is dirty, then it will become infected.
- If Tom the dog needs to walk, then his owner will take him outside.
- If Sam irons his shirt, then the shirt will be neat.
- If the smoke alarm is sounding, then we exit the building.
- If the car is polished, then it will look better.
- If it is very sunny, then people wear sunscreen.
- If the pan full of milk is left on the hob, then the milk will boil over.
- If Henrik works overtime, then he will be paid extra.
- If Sara doesn't get enough sleep, then she is tired the next day.
- If inflation remains high, then people's standard of living will be affected.
- If it is a weekday, then Barry will go to school.
- If Paula struggles to pay her bills, then she needs to budget.
- If John is hungry, then he eats.
- If the door is open, then the dog escapes.
- If John is cold, then he wears a coat.
- If Mary is hungry, then she eats food.
- If Mary is done working, then she goes home.
- If Olive is playing video games, then she is having fun.
- If John is tired, then he sleeps.
- If the tire is flat, then the car cannot drive.
- If the sun is out, then John wears sunglasses.
- If John is tired, then he goes to bed.
- If the father yells, then the children cry.
- If Peter is lying in the sun, then he is hot.
- If the birds sing, then Michael is happy.
- If Mary is running a marathon, then she is exhausted.
- If John is in the bath, then he is wet.
- If the sun is shining, then Drew's ice cream melts.
- If John is tired, then he drinks coffee.
- If John wins a bet, then he has more money.
- If there is a storm, then people stay home.
- If Julieanne has fallen into the cesspit, then she is smelly.
- If the cake is fattening, then eating it will make Frank fat.
- If Mary flies a kite, then she is windswept.
- If the bananas are brown, then they are ripe.
- If it rains, then the city gets wet.
- If Ben has broken his arm, then he is in discomfort.
- If John completes his assignment, then he is pleased.
- If the weather is nice, then John walks to work.
- If it snows, then the plants get frozen.
- If James won his race, then he is elated.
- If Mary runs with scissors, then an accident is likely.
- If Mary phones her mother, then Mary's mother is happy.
- If the music is loud in the concert, then Ben covers his ears.
- If It's warm outside, then the kids go out to play.
- If Angela has not eaten for three days, then she is hungry.
- If John paints a picture, then he is satisfied.
- If the squirrel is hungry, then it finds acorns.

- If the curtains are drawn, then the room is dark.
- If Olivia cooks, then there is homemade food.
- If Paul is being chased by a bear, then he is frightened.
- If John gives the monkey a banana, then the monkey will stop shrieking.
- If Mary eats her supper, then she is satiated.
- If John is angry, then he shouts.
- If the dog sees a cat, then the dog barks.
- If John is sick, then he is sad.
- If Sue is quiet, then the baby will sleep.
- If Mary rides a horse for the first time, then she is sore.
- If John is happy, then he smiles.
- If Geoff turns round, then Sue will wave.
- If John loses his tennis match, then he is angry.
- If the glass is knocked over, then the water spills.
- If Mary has fallen off her bicycle, then she is hurt.
- If Mary is sick, then she goes to the doctor.
- If Mike drinks water, then he is hydrated.
- If Chelsea doesn't eat any food, then she is hungry.
- If the computer is turned on, then the computer fans are moving.
- If Jane was stabbed, then she bleeds.
- If Sue invites Brian for dinner, then he will expect something to eat.
- If rain's heavy and the roof is leaking, then the floor will be wet.
- If Emma is by the fore, then she is hot.
- If Katie doesn't eat, then she is hungry.
- If Tom is going for a run, then he is sweaty.
- If Mary is in the snow, then she is cold.
- If Sarah is cooking her dinner, then she is hungry.
- If James is working out for a long time, then he is sweaty.
- If Sarah is cold, then she puts on a coat.
- If Abigail eats rotten food, then she will get food poisoning.
- If John doesn't take the bins out, then the bins won't be emptied on schedule and will overflow.
- If Ralph is on the phone, then he is talking to somebody.
- If Junaid dumped his girlfriend, then he is single.
- If Peter does not put the bins out, then the rubbish won't be collected.
- If you stay up too late and don't get enough sleep, then you'll feel tired in the morning.
- If cows eat grass, then they make milk.
- If the kettle has boiled, then the water is ready to make a cup of tea.
- If Bob is taking a exam, then he is focused.
- If a rabbit is chewing on grass, then he is eating.
- If John is hurt, then he starts to cry.
- If Evette loses her house, then she will be homeless.
- If Anthony doesn't add petrol into the car, then it won't start in the morning.
- If Joe is outside in the snow, then he is cold.
- If China invades Taiwan, then many Taiwanese will be unhappy.
- If you eat too much junk food and don't exercise, then you'll gain weight.
- If Paul is wearing a coat, then he will be warm.
- If the cake is in a warm oven, then it is baking.
- If Mary is hungry, then she eats something.
- If Anthony eats spicy food, then he will start sweating.
- If Kelly doesn't attend her work, then she will be fired.
- If Susan is jogging, then her heart rate is elevated.
- If Alex works out, then he gets muscular.
- If Simone does not study hard, then she won't pass her exams.

- If you don't brush your teeth, then you'll get cavities.
- If you boil water, then it gets hot.
- If Louise is thirsty, then she has a drink.
- If Hern is in danger, then he is scared.
- If John does not eat dinner, then he is hungry.
- If Mason buys new clothes, then he will be happy.
- If Jonathan doesn't walk the dog, then it will pee in the house.
- If the factory has no power, then nothing is being produced.
- If Michele skips school, then she fails school.
- If Alan eats all the cake, then he will put on weight.
- If you don't clean your house, then it will be messy.
- If John is laughing at the jokes, then he is happy.
- If water is cooled, then it turns into ice.
- If Callum lights a candle, then the wax will melt.
- If the room is dark, then Mary turns on the light.
- If Cameron is untrsutworthy, then he will lose his friends.
- If the pastry chef has no butter, then he can't make croissants.
- If the bus comes quickly, then Magnus will have time to visit the museum.
- If you drive too fast, then you will crash.
- If John is running, then he is tired.
- If the sun is shining, then it is daytime.
- If there is no milk, then Mary goes to buy some more.
- If Samantha calls her dad, then he will pick up.
- If Alice doesn't shower, then she will be dirty.
- If the window is open in the winter, then the room is getting colder.
- If Madison sings along to her favourite songs, then she feels happy.
- If Carl works hard, then he will be successful.
- If Walter tidies his room, then his mum won't be angry.
- If the moon is shining, then it is nighttime.
- If Clint is in the shower, then he is getting clean.
- If it is raining, then there will be puddles.
- If Kendall dates her brother, then she will get made fun of.
- If William uses an umbrella, then he is dry.
- If John is working a job, then he is being paid for it.
- If Donny lit himself on fire, then he burns.
- If the sun will be shining tomorrow, then it will feel warmer.
- If you don't save any money and spend all your income, then you'll be broke at the end of the month.
- If John has hurt his knee badly, then he is in pain.
- If Douglas is mixing red paint and blue paint, then she creates purple paint.
- If Paige's mum has a baby, then Paige becomes a big sister.
- If Tony is typing, then he is sending texts.
- If John practices Judo, then he improves.
- If Jessica is laying down on the couch, then she is relaxing.
- If Ryel practices basketball everyday, then he will improve his basketball skills.
- If Chelsea brushed her hair, then it is untangled.
- If the fire alarm is going off, then there is smoke around it.
- If Sally eats vegetables, then she is healthy.
- If you don't wear sunscreen and spend a lot time in the sun, then you'll get sunburned.
- If Amanda wins a contest, then she will get a medal.
- If Harvey is at work, then Harvey is earning money.
- If ILebron died, then basketball would never be the same.
- If Alex ate a sandwich, then he is full.

- If Joe is sleeping, then his eyes are closed.
- If Peter drives over speed limit, then he gets a fine.
- If the hole in the roof is not fixed, then the rain will continue to come in.
- If you don't study for your exam and don't pay attention in class, then you'll get a low grade.
- If it snows, then you will get cold.
- If John sleeps ten hours each night, then he is well rested.
- If it's a holiday season, then people will put up decorations.
- If today is hot day, then people eat ice cream.
- If the weather is hot, then Till wears less clothes.
- If it's a cold day, then people drink cocoa.
- If it is sunny, then Tobia wears sunglasses.
- If the train station is closed, then Joe cycles to work.
- If the dog eats a bone, then he will be happy.
- If Hill has been using her computer for a while, then it is performing badly.
- If Paul doesn't brush his teeth, then they will decay.
- If Caleb isn't at work, then he wears sweatpants.
- If Aya is drinking coffee at night, then she will stay awake.
- If it is weekend, then people go on leisurely activities.
- If it is a rainy day, then people listen to soothing music.
- If it's a windy day, then people avoid using umbrellas.

C.2.2 Confound

- If Bob goes to Elton John's concert, then I go to Elton John's concert.
- If the pavement is dry, then Harry wears sunglasses.
- If the window is hot, then Jack wears sunscreen.
- If the pavement is wet, then Mary is carrying an umbrella.
- If the ground is white, then Max wraps up warm.
- If there is a bang, then there is a flash.
- If the road is dry, then James drives his car.
- If Tim is wearing a hat, then Mat puts on sunscreen.
- If it's howling outside, then trees are moving uncontrollably from side to side.
- If plant A grows, then plant B grows.
- If ice cream sales increase, then crime rates increase.
- If there is a star in Andromeda, then there is a star in our solar system.
- If leaves are blowing everywhere, then it's good weather to fly a kite.
- If Phil wears a jacket, then Mary holds down her dress.
- If there is smoke in the sky, then the food is cooked.
- If Sarah wears shorts, then I wear shorts.
- If Eric wears shorts, then Fiona puts on SPF.
- If Mina gets a headache from too much sugar, then Mina has less money.
- If Ron wears only a vest and shorts, then Hermione is sweating.
- If there is lightning, then there is thunder.
- If Bob is sad that he didn't break the bottle first, then there is glass on the ground.
- If the Johnson family have a barbecue, then Richard sunbathes on the beach.
- If Michael gets candles out, then Rachael wraps up warm.
- If leaves fall from the trees, then Cindy buys a warm coat.
- If John is wearing a wearing a scarf, then he also wears gloves.
- If the river is frozen, then James wears his coat.
- If the leaves are dry, then the kids must wear sunscreen.
- If the car heater is on all the way up, then John has a jacket on.
- If there are marks on the floor, then John is bruised.
- If pavement is icy, then Mary wears thick coat.
- If people are going on holiday to the beach, then Mary puts on suncream.

- If the fire is burning, then you will get hot.
- If Harry's mum bakes a birthday cake, then Chloe goes to Harry's birthday party.
- If the sky is gloomy, then Tony brings a coat.
- If the house smells, then the kitchen is warm.
- If cat gets excited, then Mary gets out binoculars.
- If the shops are closed, then John is opening presents.
- If all flights are delayed, then Katie's train is cancelled.
- If it is sunny outside, then Kim wears a t-shirt.
- If the dog gets excited, then Mary gets out binoculars.
- If Simon's clothes delivery is late, then Parker can't send his Grandma's birthday card in time.
- If Gam's house is busy around 6pm, then Gam is eating food at 6pm.
- If Jeremy is talking too fast, then he has to use the restroom.
- If Jaap is ordering food at FEBO, then Dirk is cooking erwtensoep.
- If Lisa has a BBQ, then Louis goes to the beach.
- If Keith wears glasses, then Keith burns his bare feet.
- If the paint is peeling, then Mary calls a plumber.
- If John has frost bite, then Jim has deicer ready.
- If the dog needs a bath after it's walk, then Sam wears wellies.
- If the car got stuck, then Steve wears rubber boots.
- If Stacey is unwrapping her presents, then Lois is cooking a turkey.
- If the glass of water is sweating, then John takes off his shirt.
- If Andy is unwrapping a teabag, then he is boiling water.
- If the room smells of fresh paint, then the room looks brighter.
- If her flower pots are knocked over, then Susan puts out leftovers.
- If Lizzie went to stay with her grandparents, then Ben went on holiday with his family.
- If Doris was late for work, then Ellie had her baby on the side of a road.
- If the dog is wet, then George wears a jacket.

C.2.3 Correlated

- If the population eats a lot of chocolate, then the population has many Nobel prizes per capita.
- If Harry has a degree, then he can speak multiple languages.
- If Sam is a 20 year old woman, then Sam likes to drink alcohol.
- If John is playing polo, then he is good at speaking.
- If John is not married, then he now has a girlfriend.
- If Chenwei is a bachelor, then Chenwei is an unmarried man.
- If John is a white dude from Ireland, then he has four siblings.
- If Saoirse has red hair, then Saoirse is from Ireland.
- If John worked a lot, then he is now on vacation.
- If Mario doesn't like pineapple pizza, then Mario doesn't drink cappuccino after 11:00.

C.2.4 Independent

- If I drink my tea, then my car won't start.
- If Mary was working in a coworking space, then the sun would be shining in London.
- If Sue catches her train, then her football team will win.
- If Anne is swimming, then it's cold in America.
- If Mary is walking, then tax rates have gone up in Tokyo.
- If the cat is chasing the dog, then it is snowing outside.
- If Ben eats food, then his car is out of gas.
- If John is jogging, then the dog is barking.
- If you see a car, then you go to the shops.
- If Sally is happy, then a phone is ringing.
- If Jane is shopping, then the flowers are blooming.

- If Susie is asleep in bed, then it must be Summer.
- If Ben is doing his homework, then the microwave just pinged.
- If Rachel is going to a wedding, then the train is delayed.
- If John is walking his dog, then a stranger is doing yoga.
- If John sneezes, then the cartoon is boring.
- If it is sunny outside, then I had a cup tea.
- If Joe is in the car, then it is sunny in London.
- If Holly is singing in a Festival, then it is snowing in Ireland.
- If Sarah is drinking a beer, then her mother is cleaning the house.
- If John is sad, then aliens exist.
- If Mary is walking her dog, then it is snowing in France.
- If John is doing the laundry, then it is unusually busy at the Louvre.
- If the little girl gets ice cream today, then it will snow tomorrow.
- If the cat is eating, then the umbrella is blowing away.
- If Peter is watching TV, then John is a communist.
- If Wayne is reading, then trees lose their leaves.
- If Julie is throwing a frisby, then there's a discount at Walmart.
- If James goes to Miami, then there will be a flash floor in Ohio.
- If Sally plays the flute, then Jerry wears blue trousers.
- If Sarah is singing, then the lawn needs mowing.
- If Andrew broke his leg, then John is in love with Mary.
- If the road is slippery, then the sun is scorching in London.
- If Selene is running in the park, then it is cold in the library.
- If John is telling a joke, then it is cloudy in Spain.
- If Kate is talking to a dog, then the sun is shining in New York.
- If Lucy is swimming, then Katie's front door is blue.
- If Alice wears bright colour clothes, then it is winter.
- If Kent is riding a bicycle, then school is fun.
- If Paul is driving, then George's goldfish jumps out of the aquarium.
- If the boy gets on the bus, then Katie goes shopping.
- If Sam drinks water, then he is dirty.
- If Edward's portrait remains unfinished, then Sue's cat purs.
- If the table is very warm, then the trees are big.
- If Antony is having a shower, then the bus is late.
- If Sarah is in the shower, then It is midnight in Australia.
- If John is late to work, then John's coat is blue.
- If the moon is full, then the television is on.
- If I cry because my tamagotchi died, then the sun will explode.
- If John is running outside, then the economy is bouncing back.
- If Vladimir is in Hospital, then it's St Patricks day in Ireland.
- If John and Mary got married in South Africa, then there is an earthquake in California.
- If the dog barks, then the car stops at the traffic lights.
- If Tim is crying, then it is sunny.
- If John is cooking dinner, then it is a national rail strike day.
- If John is playing a game, then it is snowing in Japan.
- If John wins the lotto, then the dog will bark.
- If John is running, then the post is being delivered.
- If Mary scored in netball, then Beyoncé will perform at the Oscars.
- If my couch was expensive, then it's raining in Paris.
- If John's socks have holes, then the train is late.
- If Joanne is asleep, then water starts to freeze.
- If Pete and Josh are eating lunch, then the wind is howling today.
- If the flight is delayed by thirty minutes, then the car mechanic inspects Amy's car.

- If Carol has toast for breakfast, then the cat chases a mouse.
- If Fred is painting his nails, then Kate has a winning lottery ticket.
- If Cindy dances like a ballerina, then libraries are noisy.
- If Peter is running a marathon, then the hole is six feet deep.
- If Bob is swimming in the lake, then it is hot in the bus.
- If John is having a picnic, then it is dry in Cyprus.
- If Peter is driving a car, then a giraffe is standing in a zoo.
- If John has brown shoes, then his gloves are black.
- If I go to the gym on Sunday, then I will cry on Friday.
- If Tony broke his laptop, then it was a Sunday.
- If movies are for entertainment, then my friends eat cows.
- If Anne is cooking dinner, then it's windy in London.
- If Mary bought new clothes, then the ground outside is wet.
- If the girl goes to school, then there is thunder.
- If Alex takes a shower, then it is cold out.
- If John is watching a movie, then the neighbor works in the yard.
- If Camilla is smiling, then the sky is cloudy.
- If Dawn is reading, then the bag is full.
- If Michael is singing a song, then the shops are closed.
- If the car is running, then there is a shortage of milk.
- If Seth is walking to school, then there is a cow in the field.
- If John goes on a run, then Clara's favourite TV show is cancelled.
- If Barbara lives in a house, then the shop was crowded.
- If Amy loses her credit card, then someone's husband will fall off a ladder.
- If Betty is on the phone, then the car got hit whilst stationary.
- If Jackie is cooking dinner, then there's a parade in London.
- If the sun is shining bright, then Craig broke my tea cup.
- If Owen's favourite restaurant in Manhattan is closed, then a volcano erupts in Indonesia.
- If Jill is jumping, then it is windy.
- If Kevin is drinking, then it is cold in Sweden.
- If John is watering the garden, then it is Oscars night in Hollywood.
- If John is swimming in the pool in Cambodia, then it is sunny in Scotland.
- If Buba the dog barks, then Sam the monkey will jump.
- If John is buying an icecream, then the TV is on.
- If John is wearing a red jumper, then a plane crash happens in Indonesia.
- If Patricia took a long walk in Iowa, then it was quiet in London.
- If I go to concert with Sarah, then The Knicks will win the playoffs.
- If Mary is at the doctors office, then the weather is sunny.
- If the batteries are flat, then Pat goes vegan.
- If John kills somebody, then Harold's cat mumbles.
- If John is reorganising the kitchen, then there was a crash on the motorway.
- If John is in London, then it is cold in Nottingham.
- If John likes beer, then he is eating an apple.
- If Maggie is walking to the shop, then it is messy at playgroup.
- If I go to an art gallery, then it is cloudy in Kenya.
- If the girl drank loads of coffee, then Lily's favourite colour is pink.
- If Paul is drinking water, then Anne is driving the car.
- If the Mason family miss their flight, then Snoopy the dog eats his dinner.
- If Gianna sleeps, then her room is messy.
- If John mows the lawn, then Monies swing from trees.
- If Ann's teeth are sore, then John crosses the road.
- If Arthur's dog runs to the door, then a red rose grows in Angelina's garden.
- If Kye's flight was delayed, then Maria's flowers are doing well.

- If Joyce is laughing loudly, then the trees are in blossom.
- If the moon is full, then Mandy's bucket has a hole in it.
- If Jenny got a new job, then there is dirt on Andrea's carpet.
- If John gets a promotion, then a stranger buys a board game.
- If George's recipe needed eggs, then Nathan's bus was late.
- If Gary's cat eats a mouse, then Simon's phone will be shipped from China.
- If George has gone to the supermarket, then it's blowing a gale in Scotland.
- If Molly is in school, then it is raining cats and dogs.
- If my wallet got stollen, then Angel's mother is pregnant again.
- If John is sleeping, then war will stop in Ukraine.
- If Jill is sleeping, then it is hot.
- If Paul is babysitting, then it is foggy in Brazil.
- If John is reading a book, then Heathrow Airport is closed.
- If John turns on his washing machine, then it is hot outdoors.
- If the boy plays football, then the car will stop.
- If birds are landing in the trees, then dinner is ready to eat.
- If the English supermarket has run out of potatoes, then the Ethiopian chemist has run out of toothbrushes.
- If James is at school, then it is hot in Fiji.
- If Daniel threw his apple away, then Yvette fell over playing soccer.
- If Mary doesn't call, then it will be sunny is Aspen.
- If the snail has babies, then it's stormy in Finland.
- If John is frying chicken, then Helen's fish dies.
- If Suzy is knitting a scarf, then it is hot in Paris.
- If Steve goes swimming, then the car will use more fuel.
- If Dylan is climbing a mountain, then a helicopter is flying above Buckingham palace.
- If John is running, then there is a train strike in Berlin.
- If Mike is drinking a pint, then a monkey is swinging in the tree.
- If John's car is blue, then his bike is gold.
- If Jenny is clapping her hands, then it is dry in Spain.
- If Alan bought a new pair of trainers, then Alan's birthday is in April.
- If Mary is at the gym, then the cement is taking much longer to dry.
- If the Chef arrives at work, then the airport is busy today.
- If Tom runs very fast, then the mountain near Jenny's place is tall.
- If Erin is in school, then the car roof is hot.
- If Rose is sitting quietly, then it is St Patrick's Day.
- If the mirror just smashed, then it is time for dinner.
- If Sue is having a cup of tea, then the shop next to Albert's home is closed.
- If John buys a car, then his sofa is green.
- If the pen has ink, then the weather is warm.
- If Claire went to the dentist, then sausage rolls were on offer.
- If Jamie is eating soup, then Aglae's rose bush has started to grow.
- If Sarah is cutting some spinach, then there is an accounting class at the university.
- If John is in the swimming pool, then Jane is in Barbados.
- If John is doing the supermarket shop, then it is a full moon.
- If Petrouchka is eating ice cream, then it is windy in England.
- If Karen's the house is untidy, then it is time to collect the children from school.
- If Mary grows a tomato garden, then she owns a cat.
- If the road has a pothole, then Covid is now reemerging.
- If Bob goes swimming, then flowers are blooming.
- If Elizabeth's day ended late, then there were a lot of people at Yoni's bar.
- If Laureen and Hazel watch a movie at the cinema, then the federal bank of America will close.
- If Charles is eating breakfast, then it is snowing in Canada.

- If Martha reads a book, then the sea is calm.
- If Rupert is sneezing, then the washing up needs doing.
- If John is working, then a skyscraper in Bangkok collapses.
- If Leanne walks to work, then the cows will graze more.
- If Ivan is flying in a plane, then their is a queue at the grocery store.
- If John is thirsty, then the new series of a show is out.
- If Sue is singing, then a river is flooding in Egypt.
- If John is married, then he likes porridge.
- If Daisy is doing the dishes, then it is light outside.
- If I wear my grey jumper, then the tubes will strike in London.
- If Tina bought a Christmas tree, then Tina's birthday was in December.
- If Mark and Derek are hungry, then Eren's glass will fall over.
- If Anne is drawing, then the internet is slow in Tarek's house.
- If Christina drives to work, then she is out of soda.
- If Sally's fingernails are short, then it's a windy day.
- If Ian is swimming, then the vaccum is broken.
- If the film just ended, then the freezer needs defrosting.
- If Rachel is reading a book in San Diego, then it is cold in New York.
- If the cat is orange, then the car drives fast.
- If Sophie drops her new iPhone, then evil spirits will take over a village in Eastern Mongolia.
- If Vera is meeting her friends today, then it is going to be warm tomorrow.
- If Debbie is working from home, then it is the King's birthday.
- If it's late at night, then there is a job posting for a gardener.
- If rhinos go extinct, then Donald Trump runs for presidence again.
- If Drake is singing, then it is windy in Jamaica.
- If Patricia is eating pizza, then it is coat weather outside.
- If John is tidying his bedroom, then it is general election day.
- If Mandy and Carl get married, then traffic will be busy.
- If the cupboard is bear, then the curtains are closed.
- If Paula's tree has Dutch Elm disease, then Mrs. Smith down the road is awake.
- If Ben sits down, then the grocery stores close.
- If Kate and Edward looked everywhere for Clara, then the clouds were growing darker over Joburg.
- If Mark is at a party, then Josh is very sleepy.
- If Kim likes yellow, then a bird makes a nest.
- If John is praying, then Andrew has no meal in Africa.
- If the president is giving a speech, then it is raining in the jungle.
- If Amy wins the Lottery, then the wind will turn southwest.
- If Mark is going to the gym, then The Rock is filming his new movie.
- If John lying down, then it is warm in France.
- If Dave is digging a hole, then an iceberg is in Burma.
- If John is bald, then he has flat feet.
- If Emma is playing with toys, then Rupert's chicken is cooked.
- If I get on a plane, then there will be a drought in Venice.
- If Alice loved swimming, then it had just rained.
- If Mary goes to sleep earlier than usual, then the sun is shining in Marrakech.
- If the monkey squeals at the zoo visitors, then the boy forgot to wear socks.
- If Ian works out, then it's sunny in New York City.
- If John leaves the cinema, then food is served at Clement's home.
- If a flood hits Florida, then Carol's room hushes.
- If Timmy eats a lot of food, then the grass has grown fast in Arizona this year.
- If Fiona is studying, then Bocuse's is fully booked tonight.
- If Grace is digging the garden, then it is busy at the supermarket.
- If Pete has no money in the bank, then Marilyn's curtains are shut.

- If Gary is washing his car, then the market is on.
- If John plays the piano, then there is a flood in New Orleans.
- If the book is interesting, then a baby will cry.
- If the tea is hot, then there is a new gold course in the city.
- If Sarah is skipping, then it is raining in West Africa.
- If Owen is driving a car, then it is world book day.
- If John is calling his family, then doctors have been awarded a 3
- If John is playing tennis, then it is hail stoning in Germany.
- If Mark and Denise have a baby, then the roads will be closed.
- If the shops are closing, then it is snowing in Iceland.
- If Paris has the best pastries, then Tony is six feet tall.
- If Fred is lazy, then Koala bears are suffering from a heatwave.
- If there was a loud sound from downstairs, then the lights in the supermarket buzzed softly.
- If the car breaks down, then a butterfly lands on a flower.
- If Jane is cooking, then the tide is going out.
- If John is reading, then it is flooding in London.
- If yhe supervisor is eating his lunch, then it is very cold in the submarine.
- If Yorrick is eating a burrito, then Man U won the cup.
- If Kate is writing a letter, then a car has stopped at a red light.
- If John smokes, then his wife doesn't.
- If Joey is playing with toys, then it is snowing in New York.
- If Joe gets a snake, then I will wear a red coat today.
- If Phil went to the gym, then Phil likes to go to the pub.
- If Alec and Mona play a game, then Andrew and Mireille will fall asleep.
- If Paul is typing, then the car is parked.
- If the house is on fire, then the zoo is open.
- If Jimmy goes to the beach, then his friends are busy at work.
- If Angus asks where is the bathroom in the store, then he buys eight candles.
- If my computer is slow, then there is coloured paper in my printer.
- If Aisling is at home, then the grass is cut at Alan's.
- If Adrian is knitting a jumper, then there is a queue at the doctor's.
- If Eden is sleeping, then it is windy in London.
- If Stephen spoke Italian, then it was windy out.
- If Ben opens a crisp packet, then three cars will collide at a busy junction.
- If John has gone to work, then it is busy on the motorway.
- If Emma is going on Holiday, then Fred is going camping.
- If the elevator stopped working, then there is a new book I would like to read.
- If Nova is hopping around, then It is snowing in New York.
- If Rosie is in her garden, then it is pancake day.
- If John is doing his homework, then the supermarket has run out of tomatoes.
- If the football team is winning, then it is sunny in Ireland.
- If Antony enjoys TV commercials, then Bella drew a dragon.
- If The King has a new crown, then Mr. Smith has lost his marbles.
- If Carl closed the windows of his car, then lightning hit the tree nearby Joe's school.
- If Amy is expecting a baby, then it is midnight in Canada.
- If Grandmother is sitting by the fireplace, then it is hot in Dubai.
- If snow is falling, then an ant is eating a mole.
- If John likes cheese, then Harry likes fish.
- If Franz gets a bus, then I will drop my pens.
- If Dean sleeps well, then Nicolas will cook a nice meal.
- If Mary went to church, then the local fast food restaurant is hiring staff.
- If the garden is flooded, then the children are having spaghetti for dinner.
- If Kyle plays video games, then the aquarium near him is selling fish.

- If Joe eats four muffins, then Rufus the dog snores.
- If the stand is wonky, then the tax reports are due.
- If Tracy is relaxing, then Daniel's party is ready.
- If Amanda is washing the car, then the dog and cat are fighting.
- If the pen has run out of ink, then the kettle is boiling.
- If Nicola is wrapping a present, then there is a plane in the sky.
- If John takes a shower, then it is snowing in Russia.
- If Sarah cries, then the alarm will go off.
- If Kate's favourite subject was maths, then the dog's bowl was empty.
- If an iPhone is made in a factory in France, then Terry will cook a delicious meal of beans on toast.
- If Paul is watching the football at home, then the price of petrol is going up.
- If it's raining in Zimbabwe, then there is a fight happeing in Chiang Mai.
- If Jill is cooking, then it is dry weather in Atlanta.
- If Emily is eating cereal, then it is sunny in Mexico.
- If John is eating pizza, then it is bright in Italy.
- If Bob is a builder, then Timothy drinks ten glasses of water daily.
- If John stubbed his toe, then the boat sank.
- If Steven drinks water, then all banks close.
- If the chicken lays an egg, then it's a full moon.
- If Sue is making a dress, then there's a hole in the garden.
- If John fell in love, then Henry's laptop needs charging.
- If the crocodile is swimming in the river, then it is icy cold in the office.
- If Paul is driving to work, then Tesla stock is up.
- If John is at work, then it is icy in Tokyo.
- If Kate is walking down the lane, then it is raining in Chilli.
- If Amanda is doing the gardening, then the clock needs winding forward.
- If I go to work in Milford, then a bus will crash in Tokyo.
- If John listened to music, then New Zealand is frosty.
- If Paula is socialising, then Covid medication is ready.
- If Alison is soothing her baby, then dinner will be ready in half an hour.
- If Luke is driving to work, then it is Ellen's birthday.
- If John cooks a meal, then it is sunny in Australia.
- If Simon is right handed, then the bus was late.
- If Richard is eating a sandwich, then it is snowing in Manchester.
- If there is a baby crying, then Jane's socks are missing.
- If Brian is gardening in Mexico, then it is sunny in South Africa.
- If John is watching TV, then the roof is leaking at the White House.
- If John is feeding his fish, then the restaurant is busy.
- If the chair is broken, then it's time for lunch.
- If black is Ruby's favorite color, then there is a full moon tonight.
- If Walter watches TV, then dogs stop barking.
- If Mandy saw the rainbow last night, then there was no traffic in Seattle.
- If Zac won the swimming competition, then many lives were lost in an earthquake in California.
- If it's feeding time at the zoo, then Bob splashes in a puddle.
- If the lion is chasing the deer, then it is quiet in the classroom.
- If Harry is singing a song, then Mcdonalds is offering new menu items.
- If John is eating a sandwich, then it is cold in Paris.
- If Peter is writing, then a blizzard is in New York.
- If John likes to sing, then he has no teeth.
- If John is hungry, then the garden needs tidying.
- If it is Monday, then the cat is nowhere to be found.
- If John goes shopping, then the snow keeps falling.
- If John likes skiing, then the cake in the oven is burning.

- If John has too many clothes, then the car needs cleaning.
- If John has a friend named Jo, then it is raining hard.
- If John loves burgers, then the grass needs cutting.
- If John is a school boy, then the dog is ready for her walk.
- If John eats too much, then the door needs painting.
- If It is raining in New Zealand, then my office light is on in Amsterdam.
- If the cat is ginger, then the chairs are stacked in piles.
- If John's flower is growing, then Josh is hungry.
- If the heat is intense, then the phone is charging.
- If Karen buys a new car, then the flowers are dying.
- If Sam is dancing, then the icecaps are melting.
- If the cat is asleep, then the bucket is full of water.
- If John took his car for a drive, then the weather is fantastic in South Africa.
- If Peter paints his house blue, then the stock market surges to a record high.
- If we raise interest rates, then It is warm on Mars.
- If the cup is empty, then the car has ran out of petrol.
- If the rain is heavy, then the paint is dry.
- If Jack is watching TV, then it is sunny in Amsterdam.
- If the mirror broke, then the plants are growing well.
- If the hospital is full, then Noel walks his dog.
- If computers are crashing, then dogs are barking.
- If the door is open, then the bike has a puncture.
- If John likes to keep his place as neat as a pin, then most people drink coffee upon waking.
- If John likes to keep his place as neat as a pin., then Most people drink coffee upon waking.
- If Cassie in on the computer, then it's sunny in Chicago.
- If John is dancing, then the dog is wet.
- If Stefan becomes a doctor, then the garden sundial is inaccurate.
- If A leaf falls from a tree, then Robin goes to Paris.
- If the candle is burning, then the leaves are blowing around on the driveway.
- If the car has red paint, then the girl has curly hair.
- If Melissa is straightening her hair, then it is sports day at school.
- If It is raining, then the tiles are a beautiful color.
- If there is a hurricane, then it's morning in Australia.
- If the food is hot, then the car has a sunroof.
- If John rarely cooks himself dinner, then all cats are predators.
- If Sarah is watching TV, then eagles are hunting.
- If Fred is eating, then the temperature goes down.
- If Samuel does an experiment, then Argentina wins the world cup.
- If the soup is too hot to eat, then it is sunny outside.
- If Kevin is playing tennis, then there is a music festival occuring.
- If Church was amazing, then fuel prices are going up.
- If Lily plays the lotto, then Jane finds her cat.
- If the heating is on, then children are crying.
- If Sam buys a car, then the tree will grow.
- If John has a new girlfriend, then Dubai is a great holiday destination.
- If there's a fire in the house, then birds are singing.
- If Will watches TV, then the tree is swaying.
- If the television aeriel is repositioned, then Paul will be able to swim faster.
- If Xi's phone runs out of battery, then all is quiet on the Western front.
- If the dog has been sick, then elephants are extinct.
- If the flowers in my garden are red, then the sunset in Amsterdam is lovely.
- If Ethan needs to charge his phone, then someone is hungry.
- If the teacher is teaching a class., then the car is black.

- If Gary shops online, then the clock goes back an hour.
- If Joe runs every day, then the train will stop.
- If John feeds the cat, then it is very hot in Namibia.
- If Sharon is sleeping, then there is a deer in the garden.
- If Dale does star jumps, then the lava flows.
- If the cinema tickets are too expensive, then the hedge is overgrown.
- If Leopold Bloom walks along the beach, then Sartre gets eye surgery.
- If pages in the book are empty, then the walls are falling down.
- If the van has workers in it, then the road is clear.
- If Logan is playing with the dog, then someone is making a coffee.
- If the movie was great, then Tim loves pink drinks.
- If Jenny books a flight to New York, then the dentist examines Patrick's teeth.
- If the closet is messy, then Maria is sleeping.
- If Kim has no money, then it is a gale outside.
- If John likes Starbucks, then the roads are very busy today.
- If Johannes is playing music, then Kim Jong Un is testing missiles.
- If Paul goes driving, then the sun is too hot.
- If the food is overcooked, then the tulips bloom early.
- If Troy falls to the Achaeans, then Santa Claus loses Rudolph.
- If the motorway has cars on it, then grass is green.
- If the stove is on, then mothers are praying.
- If Luke makes a cake, then the roof needs replacing.
- If John went to go watch a movie, then this time of year the beaches are always full.
- If Jan is on a date, then the date is 15th March.
- If Lewis goes swimming, then water droplets form to make hail.
- If it is dark outside, then the furniture is too large for the study.
- If Caesar crosses the Rubicon, then I get a Rubicon grant.
- If flowers are pink, then coffee is hot.
- If Harriet is working, then the post gets delivered.
- If the sushi was great, then the bank has closed.
- If Gloria chooses a song for karaoke, then the company gives out additional bonuses.
- If the fish are swimming, then the sun is shining.
- If Emma falls over, then the water is cold.
- If John takes a nap, then Baby giggles are the best.
- If global warming occurs, then the James Webb telescope will send more pictures.
- If Global warming occurs, then The James webb telescope will send more pictures.
- If Matt makes a cup of tea, then the boy falls over.
- If the event is arranged for Thursday, then the car is red.
- If Leonardo builds a helicopter, then Rome fell.
- If Sam likes to sunbathe, then Elliot doesn't like to be cold.
- If the boy likes the bike, then the adult watches TV.
- If Sarah is writing a novel, then it is sunny in London.
- If today's flight is delayed, then the interior is all leather.
- If the court delivers its judgment, then Emma is playing chess.
- If the wind is blowing, then men are lying.
- If Will oversleeps, then the toaster broke.
- If John walks his dog, then soups are especially good for you.
- If Hannah loses weight, then Donald Trump wins the presidency.
- If Arthur goes to the park, then the steak was over cooked.
- If the camera contains film, then the ship sinks.
- If Frank Lloyd Wright builds a house, then Richter paints a picture.
- If a plane is about to take off, then Peter is doing push-ups.
- If Paul loves oranges, then the beach is full today.

- If Owen brings his kids to the zoo, then housing prices are rising.
- If Greg moves house, then a cat caught a mouse.
- If the car breaks down, then a beaver builds a dam.
- If Emily goes out for dinner, then the bird flew into the window.
- If the furniture is polished, then the ice cream van plays a tune.
- If Mary is grocery shopping, then there is a house on fire.
- If Mike fixes a table, then it is snowing in Moscow.
- If Stacey is in bed, then David's train is delayed.
- If the dog is wet, then It's 6 o'clock at night.
- If Eminem is old, then Max is in hospital.
- If it is raining in Paris, then it is 7am in Shanghai.
- If the moon is full, then my car will need oil change.
- If John is buying a car, then it's the 4th of July.
- If Gus is playing football, then there are flash floods in America.
- If Tony is making lunch, then it's snowing.
- If Yuri flies a plane, then it is Easter.
- If the power is off, then the dog is sitting down.
- If Mary has a little lamb, then it's snowing in Ottawa.
- If Michelle washed her hair, then Joe fell over.
- If Peter ordered pizza for dinner, then a lion escaped from the zoo.
- If Tim is going cycling, then the fishes need feeding.
- If there is no school this coming Friday, then John is sick in bed.
- If Hamlet talks to his dad, then it is hailing in Michigan.
- If Chelsea drives to work, then Declan will get a promotion.
- If Ralph is petting a horse, then the super bowl is on TV.
- If turkeys are birds, then South Africa has bad government.
- If John loves American folk music, then Nigel buys a hamster.
- If I drink coffee after 5pm, then my house plants will grow faster.
- If Tom is running a marathon, then leaves fall in autumn.
- If Yousef is fishing, then the Japanese stock market is closed.
- If the car is yellow, then the shop is closed.
- If Bob picks up trash, then the weatherman ate food.
- If Alex wins an award, then Netflix will cancel Stranger Things.
- If John is at work, then it's summer in Australia.
- If Stanislas falls off a bridge, then KFC mainly has chicken.
- If Abraham is a mechanic, then Isaac remains cool.
- If I listen to country music, then the price of gold will increase.
- If Jim has three brothers, then the car is red.
- If the coffee tasted good, then the orange was sour.
- If Claire found out she was allergic to bee stings, then the final of the football World Cup is on.
- If Jill is typing on her phone, then the dogs need to be walked.
- If there is no cake, then Jack is a teacher.
- If the cup is leaking, then the burger is warm.
- If William wears shorts, then Kelly will buy a car.
- If Peter is talking to Susan, then it's raining outside.
- If cockatiels require a friend, then a fan spins clockwise.
- If the rain in Spain falls mainly on the plain, then the guns are in navarone.
- If I eat sushi for lunch, then the sky will turn purple.
- If Terry orders three portions of fries, then Australia has kangaroos.
- If the table had four legs, then the window was open.
- If Stuart ran for class president, then the Christmas light switch on was happening today.
- If Ash is visiting his favourite restaurant, then the sea is cold.
- If the girl's t-shirt is blue, then Jack is eating an orange.

- If the grill is hot, then the cereal bowl is in China.
- If Olivia eats a salad, then Alex will stay up late.
- If Patrick is cooking a meal, then it's winter in London.
- If Tony can't spell, then his mug is broken.
- If Rooster Cogburn is chicken, then Mrs Doubtfire dies.
- If Fred loves ice cream, then France is famous for wine.
- If the carpet was green, then the cat ate its food.
- If Katie walked to school, then the local sports shop had a closing down sale.
- If there is no swimming pool in Harry's house, then the police are on duty.
- If the cord is short, then the truck has fuel.
- If Canada wins the olympics, then Chelsea will get a payrise at work.
- If the sun is shining in NYC, then Joe is flying to Boston.
- If Tina's back hurts, then the sky is blue.
- If I wear mismatched socks, then I will receive a phone call from an old friend.
- If Terry drinks a glass of wine, then Thursday is after Wednesday.
- If the painting was pretty, then the dress was small.
- If Maggie baked a cake, then Polly sold her car.
- If Lom is playing chess with friends, then the factory in China was closed down.
- If there is a fire, then the coffee shop is out of coffee.
- If the trash can is full, then the house needs a new roof.
- If Diane goes to work, then France declares war on the US.
- If Mary has an umbrella, then Joe likes turkey.
- If Donny's laptop broke, then butterflies taste with their feet.
- If one day Amelia was to stand for election, then Prague would fall.
- If I watch a horror movie, then my phone battery will last longer.
- If Catrin walked 500 miles, then the computer runs on Windows.
- If Sophie lost her voice, then it was a bank holiday.
- If the char is squeaky, then the computer has not been charged.
- If a tree fell in Texas, then Mary is eating ice cream.
- If a survey is done, then Hermione's toilet seat is broken.
- If twelve lords are leaping, then the tide is out at Dover.
- If I sing in the shower, then my car will run smoother.
- If Jerry went to the town, then Cornwall is famous for cidre.
- If the flowers were yellow, then the house was big.
- If Thalia gave birth to a baby boy, then it was flying ant day.
- If Dob is sweeping the garden, then fishes in Romania are sick.
- If Mathew's stove is not working, then the president of Italy died.
- If Diane calls into work sick, then Liz will have lasagna for dinner.
- If Katrin's dog is barking, then Joe is on an airplane.
- If Caitlin receives a million dollars, then there are seven days in a week.
- If I brush my teeth with my left hand, then my dreams will be more vivid.
- If Jim found a key, then Germany is in Europe.
- If the boat is sailing, then the rabbit hopped.
- If Rita went for a walk in the woods, then it was the mayor's birthday.
- If Olive is charging his phone, then medicines in Africa are cheaper.

C.2.5 Jabberwocky

- If Carol disilioness a ostrine, then the glotysoppeti is unamaneed.
- If Mary had barebered the unoweeles, then the runglicap would be bonstsmed.
- If Randy had soutandaled the shutres, then the crimulanisive would be reheameted.
- If Carl consibbaritis the retrapenty, then the brumenwirm is shrebasted.
- If Brian preshipalss a orboschrox, then the pretcra is crandadsed.
- If Alan affrahess the anyogens, then the cieryworad is attoxylsed.

- If Kelly had arocrined a muckalanss, then the duggolite would be shesidthyed.
- If Robert zulaeruss the rephaelmo, then the betravalkesi is grajikered.
- If Isabella gultiods a cribifula, then the ioncellopon is athromoused.
- If Andrew scankebulas the gertrocepsi, then the uncamomnidul is bimeromened.
- If Thomas had seminultallaed a parrilishici, then the washhooning would be jouvisted.
- If Robert had ragaleed the numind, then the staphonify would be chastilytyed.
- If Matthew syngarcals a epichyolo, then the chleurlon is mainemared.
- If Olivia had epireshieed a dizzeres, then the crusteries would be rumrammulphaed.
- If Julia hemogrants a sematteguli, then the corapitinum is braniidested.
- If Sophia had argatisined the eryphomatos, then the felipples would be cothishawered.
- If Pamela trimassitis a baderser, then the sterponer is invillarthed.
- If Kayla oxymerymbis the wilvarna, then the dideveria is pseudomerned.
- If Cynthia perwanss the lornistors, then the semienize is leatcherianed.
- If Frances had unfeccadeed a nonvoiding, then the rettoea would be infalincied.
- If William sumysais a vermierwo, then the actasugent is manarvaticed.
- If Anthony had semihued a inquentah, then the panconlep would be arquinarted.
- If Ronald had stocomezed the kibenous, then the logbents would be conpreied.
- If John had trayesed a waritonniz, then the vansibione would be underssed.
- If Evelyn had blaterendred the faceropor, then the afficlatal would be cliphomed.
- If Joseph had consuriticed the coderesorn, then the adverileae would be nonspinatsed.
- If Lawrence had downgotared the verfunk, then the invulost would be lizazingied.
- If Teresa quespendes the thrippot, then the rebecliton is overicared.
- If Keith had mixiforted the grundseu, then the proveeism would be meredomariced.
- If Patricia nuckierias the lultils, then the boxopor is megissied.
- If Gabriel sorestrings a occonper, then the oxyres is lightwained.
- If Henry had dhoustated a monalesn, then the unredinole would be cormperbeaed.
- If Roger imphiers the keptocatic, then the caratotomene is lawiettiveed.
- If Gary had paroastlied a subprood, then the hellike would be iterchuled.
- If Linda had fidmondried a myeamhin, then the utrurriatin would be minickeed.
- If Janet had parantedaed the sorminue, then the triefort would be nixanoirded.
- If John bredgeds a devulpa, then the caltinut is dispieed.
- If Patrick had rotaraliked a prentsom, then the mistmate would be pallusled.
- If Julia funthraps the cochsa, then the impoprens is fibarticed.
- If Jacqueline had chapacapsuaed the lommell, then the scacculazing would be undecaed.
- If Stephanie ovulshinis a tophiesi, then the prepromyely is wortzoaciaed.
- If Gloria had nontablented a eperchal, then the nanicknocati would be unitneed.
- If Beverly had nexditypered the makinates, then the icetes would be titundeded.
- If Lawrence extipacs a moncommantr, then the crytortious is rescruxed.
- If Albert had oversextried the rhardle, then the taigonyl would be equirocitided.
- If Keith olimicuples a necestromy, then the orifedr is swarlysmened.
- If Angela wocconiss the uncalme, then the phokhard is relaglocraed.
- If Bruce unlassisfics a nappinorph, then the pluoaporace is unsyntioned.
- If Harold had lacrynged the preetoframs, then the synatshne would be mourandeced.
- If Benjamin terjudeses a undiourt, then the loglanthly is panercepunded.
- If Stephen had spatiplesed the opsticing, then the anadifie would be gibeterboyed.
- If Maria had chreadestwaed a nonvenn, then the fororing would be gramobinesed.
- If Joyce sandagatss the corgontive, then the rugenie is chianophiled.
- If Terry spoheedrs the respaum, then the gutitas is oversorrused.
- If Tyler frotinettis the delablines, then the poemonishess is syntralteded.
- If Terry had viobukned the lyctefing, then the vizawarine would be unemaked.
- If Sean flecispetis the abysochpist, then the misbeks is outlenctorabed.
- If Susan had jarristed the doetempted, then the bathroment would be bresused.
- If Mary reughmis the clikups, then the noncaly is unsermoused.

- If Shirley legetrosoldis a mischabicilbe, then the moricogly is inenopodualed.
- If Denise uncurbitss a whirnisexy, then the clumedite is galayerlizeed.
- If Nathan sholosperens the overvisher, then the lytipinu is phalpaned.
- If Justin had smoklesed a stolochag, then the setranoralne would be cycodered.
- If Samuel had planidyed the fermopabia, then the glignist would be scanoganasted.
- If Jack had hypercatrismed the wribbitologi, then the arventist would be sympheaed.
- If Nancy had lungianrged the conaman, then the gigulcents would be foridifieded.
- If Frances had volygoned a peplenlital, then the cirapinided would be heronteded.
- If Nancy imposestmes the unsiliation, then the puncheys is bestunistiveed.
- If Amber snissermants a sublenedin, then the epippicarc is nomplopined.
- If Margaret blecisens a megresi, then the sperismin is sollemaned.
- If Heather parharects a toulptieve, then the cesenceousid is unsusersed.
- If Gloria redapnests a cortres, then the plefigical is decoitived.
- If Teresa nodingis the semurrhe, then the miconeure is cohonited.
- If Catherine had plecksyed a staphing, then the semimph would be planerieed.
- If Heather henutoses the encimbusne, then the redlam is cupeptinged.
- If Robert supherands a floenly, then the ambaritaxi is psychygoned.
- If Emma posmaragnes the devillie, then the ponitionl is tablentolited.
- If Laura revisuiliks a blotosetr, then the deziliga is caniatinized.
- If Nicole had bewiticaraled the shiablin, then the sacksonve would be rectronied.
- If Nancy had kildedened a pelats, then the immopotrum would be unstioscented.
- If Andrea had divabjespoused a glasuornessa, then the afreslack would be heefopasted.
- If Christina had grangstred the plammed, then the depremptess would be brannatied.
- If Scott had mytheasomened a preappines, then the elecingede would be addsiptioned.
- If Louis had rhinocheded the endious, then the plastaxinal would be undividiteed.
- If Lori serloiess the pentadit, then the reitableg is pagambiled.
- If Roy calotyptudis a perigulate, then the pervenouncea is bodomashinged.
- If Cheryl had extringed the zelipsitumo, then the melivea would be carniervableed.
- If Douglas had pomptored the quascark, then the undeomyder would be stornoryed.
- If Denise wrivings a stratierat, then the getotics is prerenationed.
- If Katherine had fiteriumsed a maratopium, then the pycoshore would be triuncified.
- If Harold had mistlyed the imarticalica, then the gesibidae would be prenzeded.
- If Nancy had fropareded the nonophtab, then the decoisms would be cleconued.
- If Megan had carminged the prodema, then the signeoundo would be oxyesessed.
- If Joyce had anapinged the novaturanis, then the fasommen would be pseudinged.
- If Harold unbeloges the lowtwor, then the monuscarbian is unwinkradeed.
- If Emma interlophes the misorinica, then the qualedl is unbudonged.
- If Logan had perficed the heabbitarize, then the oppenestic would be seriableed.
- If Samuel had aitimsted the uninogenic, then the assinyc would be hypraemsmed.
- If Brittany had cychiogyed the cochyma, then the milannemen would be drediestsed.
- If Scott chassadibles the otherecional, then the appralmicous is slimisticaed.
- If Judy allismans a thoyspean, then the unilamic is pleopidumped.
- If Joe had nonneticked a vervigicr, then the mustesse would be sarlichanchered.
- If Sharon had bloodowsed the amygagedne, then the conseed would be biurovered.
- If Carl abrathagres a rectege, then the yarcisiman is hombliteed.
- If Zachary leinopiforys the achievo, then the indeflas is rutswaricaled.
- If Cheryl had cravishiped the fretatud, then the aggsnopiral would be socillastyed.
- If Alexander studerasibs the hollerrato, then the laddyngst is clopantlineed.
- If James had eastopyeed the defleicka, then the chalieumati would be flayerlyed.
- If Frances had scathoted the achorin, then the perment would be electismed.
- If Martha vesecolings a tastins, then the unscatte is pripyringed.
- If Kevin had antercated a creffen, then the catenisme would be hearleryed.
- If Lisa scurreds a ponvolotope, then the trifaimide is preterousined.

- If Jeffrey had obstrettored the rhisueary, then the criped would be almnodymped.
- If Terry had genomocativeed a coveables, then the overonost would be huchivinged.
- If Samuel had baicumbaed the enterox, then the autbred would be supedneed.
- If Pamela had spliccapaed a tentratoma, then the tierconic would be supertaned.
- If Pamela had addunesseded a apilcatg, then the furrighess would be cotoropled.
- If Eric had eyewowetesed a vithophagen, then the hypocetrar would be morrsposphoed.
- If Lisa had overproed a ponespele, then the rethryagge would be mooducatinged.
- If Douglas cinterss the stalike, then the thendicency is neartaned.
- If Virginia bleniochipacts the autinctrooke, then the uncompred is readgeresed.
- If George melives the cellexessl, then the stozed is bromeeeed.
- If Brittany had ledbetioed a chrourshili, then the opitcling would be annumiarted.
- If Stephanie had semewelmed a reasledi, then the unmihinom would be semprotafted.
- If Keith had unithenesed the uroffloes, then the supercants would be chougaterined.
- If Stephen had aziglateed a broscomy, then the nonobill would be allupplied.
- If Madison copherwayells a almahoste, then the waventit is natitableed.
- If David spirdstrizes a exedrompsing, then the eolorous is nongingled.
- If Mark phomerfics a evinter, then the bermiescer is litterdlyed.
- If Ann bloosacas the cligsk, then the annysitti is miscatoraed.
- If Kyle candesids the micelignes, then the murgiding is sodiveaded.
- If Frances warrations a gaggarde, then the prebere is spattleseed.
- If Kathryn had pocookinged a antersocy, then the ungenation would be uncorninged.
- If John peposexesss the wightanh, then the reliotte is crophuted.
- If Madison ungotatotomis the distetapit, then the chisperche is trayserollied.
- If Anthony palellums the accerlize, then the cachilatic is dattryriessed.
- If Gerald had preudosarled the campenize, then the statheadomy would be prostrocalated.
- If Andrew had chimnaned the antracers, then the inderat would be banumbered.
- If Jeffrey pronons a pumblemented, then the overled is configgined.
- If Margaret had preshariated the uncoms, then the skewargiong would be ethmiemed.
- If Paul had borblenindred a ciwielerte, then the deadrelike would be creethented.
- If Anna hemorss a beammophes, then the knobringle is bultenibleed.
- If Brandon had uncerprooed a mitroited, then the unrivulshishi would be nonocograteed.
- If Judy had godetrantaed a nugiodeve, then the loattorater would be shoptageed.
- If Douglas mougeds a strinfac, then the pretrably is benappesed.
- If Adam octerans the orerione, then the surdernogy is acapplesaed.
- If Terry had disanteroded a discarchion, then the missiessity would be unsculogyed.
- If Betty meteteds a geodjude, then the cranstoproid is unwinguluoued.
- If Deborah emerryeas a redogate, then the reenaimpiro is unsuffeiaed.
- If Justin asphallas a pulapted, then the demagonem is spidereted.
- If Eric destierfers a orthopho, then the diductic is outfigesed.
- If Stephen eruggeds the overwheept, then the dekarpose is trapinaed.
- If Joseph had biomplented the reflastripted, then the chroolish would be psymondissed.
- If Shirley retrophars a empospe, then the uucomak is mandersaed.
- If Karen had toolerighted the abraceliform, then the azilidoaring would be maciremated.
- If Heather had spillined the travinin, then the overicon would be labbirsed.
- If Scott skiterss the scratorke, then the saranuepe is unequingoed.
- If Charles had inciamyonized a avogets, then the arepyles would be traymbionied.
- If Julia whisembilis a jagoptivent, then the perocalized is comboasmed.
- If Alice had overcreaed the pedularcha, then the disoar would be clabictlteed.
- If Catherine had sousepableed a asyktosb, then the uncironnic would be omervenied.
- If Ralph had remactiolowed the sonneaten, then the strepacer would be zapansned.
- If John had noncellced a abdulantast, then the exshtlet would be chringinged.
- If Mary had acroured the hypiling, then the thylloof would be polsfulased.
- If Joseph fornigatoses the iodpophenth, then the wellystou is trildished.

- If Sharon bipophemis the prenagan, then the holundi is sepouninesed.
- If Danielle machyntics a parnouree, then the vastedicat is neuvenitypeed.
- If Billy had pedisaued a unarkese, then the uncoidan would be petstered.
- If Julia firaxatorenes a gracturs, then the luttonquit is letimpolyed.
- If Vincent had masophyoed the nonfuiera, then the cromatica would be ichwingced.
- If Gregory had ciciestded the nonapantia, then the virycenizer would be pseudoseled.
- If Deborah outchinerls the educione, then the baparsed is undernaled.
- If Betty carchachymatis a whamptic, then the tekatier is discinded.
- If Alexander had enadroofied the octaths, then the crakflut would be lutorativ ed.
- If Nancy granaphyes a uncrograge, then the mormulworts is trightitioued.
- If Douglas had aurseptined the chakeenin, then the uncogens would be stocynessed.
- If Walter had sasticetied the ooposyna, then the perfeasiat would be albasseed.
- If Ryan gastoras the phodogesco, then the unnultess is saddythmeed.
- If Benjamin hacifitedls the extorwaratic, then the cooratessly is fizamteded.
- If Gregory had hebostsed a conglideing, then the sactlypne would be nonbanested.
- If Justin displizess a nonosters, then the whirassilite is soveradiscoped.
- If Gregory had senobultined the theezueni, then the urenorods would be nonconioseed.
- If Helen cerosturits a neursup, then the rhilforupp is epirrieoed.
- If Gerald had wateroged a cachergess, then the conbicke would be fladisoned.
- If Jonathan had understsled a releolois, then the folymmitiv would be intersihed.
- If Kayla had jermathreed the sliphous, then the acrityphac would be fislatualed.
- If Diana ureezelmings the cardigonato, then the paturusit is penzoylaed.
- If Elizabeth unquettess a loists, then the bravidiastious is theopophined.
- If Brandon had foghtwayed the rarrisclog, then the undours would be insucelered.
- If Henry had cavetumed the mittenize, then the unticilize would be paltiumened.
- If Denise nonangs a pultery, then the exemened is homanseed.
- If Frank sultiforiis the unshimed, then the methour is snopsyneed.
- If Helen had plurgeed the giurveum, then the hyperitte would be mestelleceed.
- If Christina had cohorbrehened the nonishien, then the hommurieroon would be interizeed.
- If Laura tandrestees a jenedicart, then the unrechponizia is histcresed.
- If Adam had sciotanmened the imperha, then the vennestive would be athibrodyed.
- If Isabella wollionis the muderorwo, then the mulmatd is sinhoppuled.
- If Emily naterstdos the couste, then the proming is diverteed.
- If Alexander snorofulogics the thysang, then the unenliculowy is bastensed.
- If Michael had ottalegmed a cootactah, then the tribolatie would be antroormatred.
- If Jeremy had necololled a daully, then the goffymed would be bluvannyloged.
- If Joe threbodynas a sploefuetic, then the aggrequi is wolipsityed.
- If Lisa fidiscesiss the posthsi, then the nonerdiat is minsehed.
- If Pamela had nomarded a toxiconcesh, then the dissiest would be legaricaldsed.
- If Terry had thrisawed the hydropm, then the orphoozoaness would be weletiteded.
- If Pamela had reetalacked a ogatheid, then the paccingking would be chindlysized.
- If Daniel superomlics a sidedele, then the madelear is peppagileed.
- If Cheryl had oozzygraed a regalluenced, then the infreoet would be baradinged.
- If Samantha had chutisperiteded a stuminod, then the deforave would be extidinged.
- If Richard laburbiartivs a cenistrous, then the basgrunt is wisozeted.
- If Charles had fluglexicoted the meermosto, then the gimlatone would be ampatorksed.
- If Steven had nonfrequaed the semophoele, then the votulonis would be thorcomicted.
- If Anthony fordoputs a tumons, then the unchsyme is pseudgeed.
- If Maria procativels the preanding, then the boldarerds is corcassioed.
- If Logan had outfisumed a phyllesss, then the trotint would be reniscalied.
- If Sean had incapininesed a tramaner, then the infiedam would be adreloksed.
- If Carl frantifyinxs a traywrablexe, then the pandsmate is attrapineed.
- If Sarah overpowlics a unellizabled, then the uncomagin is demolikered.

- If Michelle had giglusted a mobbscolize, then the reinosional would be untefagictoed.
- If Victoria had hypocuted the walfwiperape, then the conpares would be tolorelled.
- If Mary overacchlys a chingicip, then the ostraphylami is pullemposcened.
- If Ashley had dimercusesed a palerdicate, then the sebrolke would be querfrooed.
- If Hannah had robusomatied the remanolik, then the holoucklet would be neruscarded.
- If Evelyn had michytivered a inderledn, then the ilierels would be cypercemeded.
- If Alan dissartabls a revinks, then the deruthe is agattrophed.
- If Billy anisahs a anoprimis, then the reptictfula is houttered.
- If Michelle unsethams a relancepha, then the fluconylus is blukotheteed.
- If Pamela had unfilisheed a archilatogr, then the heroolss would be cohallisted.
- If Isabella had unsuryllogyed the lefeensks, then the ricryon would be cronolowsed.
- If Tyler imploplats the baempt, then the rocifieraco is unachetnesed.
- If Jennifer orpholass a gynelaing, then the unindficker is conaternned.
- If Arthur had anilyhialed the tallmentne, then the inexed would be pedgmetined.
- If James implowness a austhse, then the waverturid is ungeroveed.
- If Kelly had zoadielined the unthoplit, then the haedillizate would be teacinceed.
- If Joan had undendreomeed a seminey, then the delpeoles would be horratoluised.
- If Kevin rithumeropes a underion, then the overgialled is prituppered.
- If Albert isoggines a kakefectica, then the panatsule is horsarrionied.
- If Brandon ostrogiasys a infoffertiv, then the caemating is minentrioned.
- If Adam had lataliketed a nichorrhora, then the snoweed would be stacticaputed.
- If Deborah had dadginesused the unplingia, then the cutophytoso would be hartuscreed.
- If Logan webithrys a undistocra, then the shospon is extemakered.
- If Amanda rebianumas a resigatizato, then the eiggenothy is lottingsed.
- If William had unnensorted a novisrop, then the antoersr would be systinlened.
- If Gregory had nondiglesed a paematible, then the subecher would be intibereced.
- If Martha eleppingis the chalmogical, then the mickroplus is metallested.
- If Ronald had syrestonereed the unpervi, then the reashwardium would be feariantinged.
- If Billy had sterfoleed a tarbonian, then the quidina would be ashesmatoted.
- If Janet unberturas a pledepida, then the pisuisita is broperwinged.
- If Samuel banderots the ballizals, then the collanerar is triedlyyed.
- If Joan had pomettureded a indrodia, then the bezizoming would be encroligned.
- If Sandra had luplogleed the planede, then the prigocluren would be persaminced.
- If Emma had mignestingled the tranarograe, then the lapochemate would be hypotopinged.
- If Karen had supensied a dysrectness, then the retidesii would be prontlectuaed.
- If Terry utinotus the oxidenys, then the regaridae is apothomaniaed.
- If Wayne tranitiets the hantiud, then the undertiness is dulfusableed.
- If Cheryl had incontlised the enjocht, then the hypergenuati would be dihabrobed.
- If Austin platessros the excurantiou, then the jaranonchm is carpelibateed.
- If Edward reamperizs a gualita, then the photomy is misyandineed.
- If Keith fakiennestans a theudotingu, then the venonisms is overinolineed.
- If Emily wistalozzialls a odwarbowar, then the concomet is panucitifed.
- If Eric melfwatilits a megrobiont, then the vircom is dideroacked.
- If Cynthia had borableded a collowk, then the dephorlagu would be liphiered.
- If Christian had divitheted a mizartomence, then the tyralvelen would be recraculed.
- If Philip had seineterceed a pulletinis, then the overadene would be panchaned.
- If Jennifer had madigredeed a phymetism, then the cailorcusly would be aviagoninied.
- If Jeffrey had fallakinied a mureaderer, then the melacty would be ermanosed.
- If Nicole had oppandaced a connelis, then the cynocalfr would be alkahiveed.
- If Nancy had yuquiterided a anortes, then the laxilistess would be fetterionaled.
- If Evelyn had blacketrumed the klabantrous, then the protgauty would be curbworabed.
- If Maria had hinmulfulogred the brozinarin, then the glycerman would be polycantisaed.
- If Diane had assorescoeed the nonmininges, then the unbulatiza would be whoticratied.

- If Jacob had extringed the winobantion, then the horraters would be maggityed.
- If Jose had stackishised a nonhelist, then the caverfitan would be ballulied.
- If Victoria deninices the sorehorwero, then the pumprel is shokeried.
- If Tyler mistrousts a makinitard, then the cicknom is chottined.
- If Frances exoggismans the epirilicl, then the compenn is interpoeed.
- If Judith greecatopegms the undostat, then the eadmogless is fansyntifeed.
- If Nicole gonestoriss a oxblicutun, then the autulaye is destuninged.
- If Benjamin artefics a hisdraf, then the willande is unreviced.
- If Adam kenauttes a polcirai, then the elutalousnes is guaeogetereed.
- If Andrea octenerans a pendworophy, then the misinesl is echnomoteroed.
- If Linda reomatimizats the decony, then the unpihyl is shuseyermed.
- If Nathan had patansithed the unhumintio, then the aldortula would be antrologed.
- If Billy unshoneds a chilortful, then the cauroccen is excubroced.
- If Terry had citriticed the redertie, then the anterio would be trowesed.
- If Stephen immortriss a unpermancyt, then the artidepure is overnetrineed.
- If Ethan had unneringed a antentativeli, then the valepygle would be scrogloflaed.
- If Kelly arnionuss a subelize, then the lycled is liquideraphaed.
- If Debra angenspergiss the combringence, then the gasced is reafflogued.
- If Madison had olidandsed a ampacusly, then the tabyriatia would be chomansaed.
- If Albert had rommententaed a cysterery, then the lotualical would be suberinglyed.
- If Larry paltiseepass a vermistriess, then the unlisit is hoticanized.
- If Gloria had comlactesed the preterial, then the predinolic would be quarlered.
- If Raymond had dollastosored a domunterin, then the irrechest would be thytomaned.
- If Donna rakewlss a honcreable, then the tecructuropa is drustgeneed.
- If Janice had pandauoed the breikutl, then the sproidible would be granceryed.
- If Katherine colenterts a sipilanide, then the bemonte is plueentamiced.
- If Dylan subsonvns the sacaudar, then the viverotis is untispoted.
- If Zachary had ciplicateed a pommetyt, then the peaddae would be martlievoused.
- If Deborah meriffss the somats, then the lackahs is plaiffereped.
- If Stephen had muummianted a mecatring, then the encontryst would be fumalired.
- If Jordan olitherals a pargaria, then the zithonte is ancalyzied.
- If Carol had mommingaed the lotarkism, then the entimptional would be defoasesed.
- If Hannah obirpedrapholos a tenelonizate, then the thaudose is unverfowed.
- If Nancy chorilents a asbersedin, then the banqualioi is swatophyed.
- If Kenneth curcesmaks a uneudednes, then the overbrad is gristermed.
- If Shirley had sperventeed the recepechip, then the jugalus would be nonthesoned.
- If Jesse had tristsolied a hexanetisted, then the curfbote would be albirisciated.
- If Dorothy hogendentss the chalinn, then the rupercouna is destioed.
- If Sharon underatists the unproveds, then the sapotte is forailingsed.
- If Philip phytanceles a unbulace, then the pancormen is sermakinged.
- If Billy metraes a clebur, then the depidomar is vorepteded.
- If Jose had bahabstermed the cychearizi, then the quilostelc would be putroidadeded.
- If Melissa edenchs the federcy, then the subocyndat is keelloweeed.
- If Daniel oponinallikes a gredays, then the overchead is gastrableded.
- If Eric loggless the stracelled, then the mesnasy is anthsholaded.
- If Carolyn had cominanted a unipulate, then the lasighill would be backnospeded.
- If Brian had litrocatied a plutablati, then the poissingency would be balatbrexed.
- If Abigail disturastents a vithoripric, then the flanne is unwilesed.
- If Andrew had accunklesed the tinglociz, then the implicat would be phorcolkinaed.
- If Jason had aranottewed a joraweedt, then the uncomethed would be futosteeed.
- If Jerry bismidents the alchared, then the vividemand is isoledesped.
- If Ryan fiercopsidas the koilagate, then the medged is postabeanteded.
- If Marilyn strilarlys a opittn, then the hasstinshous is unradcaured.

- If Justin dotemonincs a nonveness, then the proushiple is drolagonesed.
- If Julie had diaillosed a packstrien, then the flasphelle would be ambitlesed.
- If Jacob bonuopeds a nonded, then the dialyticaliati is unicytingoed.
- If Catherine had dantracidlyed the engroids, then the berfrant would be sterivireded.
- If Patricia retractoms a prometsa, then the thostere is implanned.
- If Megan had wininesed a unsupeae, then the calautioning would be unambinged.
- If Carol had droidumoned a boldnofe, then the conoptols would be sterdyned.
- If Kathryn quonimors the curregal, then the sandriguen is ccehesoatoed.
- If Steven deptukus the arimmer, then the ragioes is epiumburned.
- If Ralph had prectintived a clowan, then the bainging would be eryphoadaed.
- If Martha conranres a stinkin, then the katizoitos is tholiableed.
- If Bryan had verbinesed a winnacrabl, then the furdualiou would be cryopermanaled.
- If Hannah had skallingssed a mickeciesi, then the ambistylae would be nonthmatinged.
- If Gregory surfiguends a pealkpai, then the tairtinic is heronessesed.
- If Louis had mosanterited a uncragul, then the whilocotot would be polocudeed.
- If Joseph had sanghterfed a preguett, then the encoundogia would be coxbutismed.
- If Margaret had unpreursed the prectupla, then the fimairate would be promacesed.
- If Joan had myarifeed a sillota, then the meallike would be nookanted.
- If Noah neorowss a repseters, then the overseweet is jeteectioed.
- If Larry arbignourings a nondalicat, then the purstap is unjacriaed.
- If Susan had cicropageed a hemphing, then the overciondr would be marrikedted.
- If Patrick had parcicomsed the helegurinis, then the gibbeh would be vicepingwised.
- If Juan had buchalitied a nepatic, then the toteracoccu would be reverematorabed.
- If Madison pensumls a telomises, then the noneaship is otonetsed.
- If Jacqueline had scumbutsed a valabst, then the cecrowalist would be norboloofered.
- If Pamela crometis the isoldwis, then the phaumocrate is stattalveded.
- If Kimberly had squivelliced a adiopt, then the withonials would be pischinounded.
- If Gary copatchips a cetenial, then the geolloody is intiquided.
- If Brandon zygrexings the ungrastema, then the overoscranc is rubatiounded.
- If Louis endophles a calanoctu, then the mesulde is idioludgeed.
- If Doris had nontalianed the lishur, then the celesurupl would be hypothembleed.
- If Amy secaliperas a sikvery, then the coulowee is dittededed.
- If Jordan had phlorminged a anniable, then the chrophes would be benonsuled.
- If Logan had alkhtcreled a chufiction, then the entrustry would be orthopeed.
- If Randy had yamphyled the bluvivicive, then the cyangasee would be secogessnemed.
- If Harold had praudomycoled the unsnali, then the soverpaeiar would be apherked.
- If Christian had reternicaned a damatalialiti, then the evansper would be silluntateded.
- If Michelle had unstborplied a unpercopere, then the pitkeshly would be cholerinared.
- If Ronald ostifolas the jindwedly, then the parnlopap is superoisted.
- If Megan coquaraeiacis a wharifolus, then the auttaray is macklyteed.
- If Jean had espearkseed a theras, then the dayllarics would be pighetenteed.
- If Joyce volettainatibs the neueba, then the sulbeltico is kinophopreed.
- If Amy decominesses a chaldsy, then the coemoise is singinoctyed.
- If Steven endifishs the bosqueful, then the unbobily is demantryed.
- If Melissa bietteres the suliged, then the inimanation is colatussesed.
- If Kathryn had unreeloed a reggised, then the tapperiewar would be appotyptoed.
- If Timothy had secolizeded the wangualida, then the outberip would be vetheastraed.
- If Amanda juxturchrons the unflolute, then the reardryi is alkainged.
- If Louis foexpeciss the glositsl, then the inoforhyph is coltmanteed.
- If Carolyn proocitens a haeodgme, then the chirefiess is unexpercinged.
- If Joshua had pedinalamed the conmensomenc, then the infatheenta would be rekolissed.
- If Judy preatatalins a perene, then the telosing is criseuresed.
- If Dorothy had neurcipsed the neocasin, then the drughhentes would be sulpfulaeed.

- If Raymond had cinreasged a nonhiry, then the higgrisha would be unsuarocyed.
- If Natalie had gaddermiousled a canini, then the ilivaesl would be toriogeousned.
- If Janet likuhustoms the reetlyn, then the coreflac is lithgataed.
- If James had oximmianed a pardled, then the sannedler would be horenided.
- If Kimberly findiranals a preassr, then the haphingsc is shriasesed.
- If Margaret pubvanings a elfitothip, then the iriosyntio is wraworledgeaed.
- If Randy had archiaed the hembrate, then the chatel would be idiftatived.
- If Megan had puffesded the prectizatel, then the anadende would be supernersed.
- If Donna onytrishs the chopplabl, then the neutariona is horpeoteded.
- If Olivia had paraginnered a fundmang, then the regropled would be chaforievied.
- If Philip ecuterias the rotieborag, then the pseudophag is trantylined.
- If Cheryl had meltinsed the libros, then the skevoquies would be acquideed.
- If Jesse prostidials a inconish, then the lithoweeter is burawkinged.
- If Stephen had deprardlyed the loggencatic, then the transuptic would be archhanaed.
- If Thomas collasts a easeedfula, then the chaenhess is ensonicoseced.
- If Brittany cerimpliss the tetractive, then the enampronam is stralsced.
- If Nicole huryntiscus a supence, then the litticiusly is opperculianed.
- If George wridgiostss a pripaite, then the unsprojas is recovalmored.
- If Joan had rooraned the prenogran, then the volybolike would be wagesitned.
- If Amber had hydreonoofed the fragualmi, then the loissilifu would be bacenesed.
- If Samuel had interldaed the osceste, then the semicates would be tonrimelizeed.
- If Debra yamesones a noverbus, then the tachumso is lotquetabed.
- If Diana pterfarditys the albansible, then the ploubbithea is interhulmed.
- If Kimberly pingarplas a haensithe, then the fumphala is shiloteied.
- If Christina ampergets the impotrosy, then the unstrilit is dephantrined.
- If Isabella had waguevelaed a gabyrali, then the nonezingic would be physiaseded.
- If Zachary had datirrodeed the argarbraphy, then the withytions would be zintorathed.
- If Jonathan had pentinsawed a atocenste, then the bleationou would be pascanged.
- If Alexander had revilibiaticed the tskalemosing, then the escatratice would be hyphalised.
- If Ralph had unomatrofed the eelapis, then the auracheal would be ultisingleed.
- If Janet momphoftss the ciraping, then the hydropred is scabbened.
- If Brandon had omniscranced a balwer, then the peroming would be toutzaeed.
- If Susan had sporiacked the rhyrhouselded, then the aratifing would be bedacculaed.
- If Gary had casegrateded the seadracing, then the heavered would be hepriableed.
- If Lori semoblens a sobrolu, then the cymophy is demiwelalised.
- If Aaron cermonias a culurnupt, then the ensplodici is geoberseed.
- If Jennifer had undropsed a outtosolor, then the maggrals would be pethograteed.
- If Russell overrismss a condlativigrar, then the unrandraft is vanankked.
- If Brandon had misclemened a exenochaine, then the chahmicath would be semulestred.
- If Daniel bosilings the shryncenomos, then the probbathi is tepouseed.
- If Jacqueline colingogis a anteofle, then the nonpary is nepulinsed.
- If Barbara uninkless a colawlc, then the gelterth is cargenizated.
- If Grace reddishlitys a scuteciman, then the operic is begrogruled.
- If James clansemicatis the intholityp, then the wishal is effibaneed.
- If Philip topretionas a wimbuthor, then the inurobia is jusseenessed.
- If Michelle had endomersed a reattema, then the recartiosi would be extingsied.
- If Terry hyphemanias a conderia, then the slastoist is undulphyed.
- If Anthony had materdered a stylecipe, then the dayhemingith would be nonidabled.
- If Gregory phrieknols a defeatides, then the quilet is brotomaned.
- If Kelly legannabes a refralite, then the caraclian is antiolinged.
- If George kernaropuls a kingboo, then the underall is lapiringed.
- If Ronald had pinemerasied the tushbily, then the usalagge would be fribariced.
- If Gloria conimbitens the monpeeporo, then the siliot is nonconspeded.

- If Sara prepholows the oroodizess, then the semaisthutab is unciolotored.
- If Catherine moucarybs a pilvinoga, then the ciggiermi is viraeivased.
- If Jack had anterticateed the epheribe, then the rellastionas would be hirestheed.
- If Gregory tipochlys the supertaia, then the muriorcum is basevorkied.
- If Douglas anadlernas the speckrious, then the approwly is resperipised.
- If Michael had predioneded the sackwayi, then the overdratia would be dimpocytioed.
- If Christine had glaymobuted the stubforify, then the quamnativ would be insinidinged.
- If Brittany milpagdes the leathooms, then the import is underdonityed.
- If Timothy had epithoed a codgerch, then the recartype would be bumpathyed.
- If Karen had supeabalsed a trideston, then the nibradill would be alveilifyined.
- If Gerald barkenohs the harbons, then the semiples is olycheried.
- If Kenneth had unsentled a fisillatig, then the pronerom would be specenutioned.
- If Ethan intichs the supenizei, then the undisturants is brainessed.
- If Helen unglinglys the griouvity, then the passophory is stalished.
- If Jonathan asciestes a beatyphym, then the reinistaten is croodudeed.
- If Kyle mirberss the toxageess, then the ragaphomenc is gaupreoused.
- If Alan had klakrumunted a whomaki, then the overse would be feophidsed.
- If Lori had pepasturued the cadgedi, then the thiter would be buxamostioided.
- If Jonathan pladeras the subconed, then the callanths is swingilyed.
- If Carl chocatenys a olinable, then the benupha is pinniotsed.
- If Judy had pinkmanteed the rommudack, then the unchinee would be autivacked.
- If Martha quinkss a ademoidifi, then the dryptured is abrehoceneed.
- If Carolyn had jokiminged the bungmalma, then the tanchis would be cirissed.
- If Julie had mobliquaed a tretagef, then the unpator would be coquinesed.
- If Jack had dowelleded a druntan, then the gaucic would be aeopreed.
- If Joshua had nonfaxined the rappocophy, then the rumbegewed would be pridelifieveed.
- If Diana had svaribirthied the anters, then the ciracidic would be hanascormed.
- If Joseph acripsess a buniculeca, then the sexyclite is psynacomaed.
- If Gabriel slietermets a outtrophar, then the rehervic is permaroofed.
- If Nancy had colovainged the manthemett, then the artherizes would be orballiessed.
- If Pamela had tettemsed the spathpitaci, then the hydrodonale would be lacrobleded.
- If Frances had couassessiaed a sneumbal, then the linarai would be braticed.
- If Janice had huinthyniced the inesophaurab, then the actyle would be suchnocneed.
- If Larry had pregbappyed the pandwifart, then the sharylt would be discosideped.
- If Terry ismuntedneds the inswank, then the ipeanui is uncomineded.
- If Olivia gragivostalls a uptidae, then the hypelisityly is urorthhed.
- If Ryan had peropsylateed the topisy, then the chrectintat would be bingoneded.
- If Edward had actosonizeded a overlebu, then the serpitfuggl would be empheoged.
- If Megan had spectoqued a typerding, then the purcineso would be sleumnilaed.
- If Janet cootemics a coazilier, then the laznasks is imprassesed.
- If Gary octunniums the amoudgeria, then the arsentery is sellandered.
- If Sean had baberned a cartmutive, then the depavede would be entreninged.
- If Alice ecstellisms the frojagatos, then the quakelious is trosprosered.
- If Scott had embergatateed the figueake, then the unickeryss would be roencidened.
- If Ronald had celiosporsed the quitheteli, then the obietfles would be hithraled.
- If Christian enguioness the louconist, then the assinativ is gimivaeed.
- If Stephanie had demitopbed the johavolors, then the picres would be anthotyed.
- If Katherine hetalneers the ricristher, then the bedantive is miswaryed.
- If Jerry uvultles the silionto, then the unfugni is madiscathened.
- If Anthony had swassraphed the baracs, then the reavifore would be vennonfiseed.
- If Jennifer had phoelsifiened a ciperrawag, then the shimpundia would be hexaryed.
- If Russell raggraes the pandatifietri, then the angracter is roodretriaed.
- If Lauren had opiputerned the lisadvatord, then the pingstposcop would be mativignesed.

- If Heather unkyriess a loisin, then the micromat is curistaurpored.
- If Randy hemaceonouss a irrofloquar, then the subaterse is unarquardoed.
- If Mark had knorthohined the comanisema, then the sumsanclan would be moftsmodined.
- If Julie had preafectioned the deraneserte, then the pseurpent would be crecacormed.
- If William bedevins a monatheri, then the hypacured is nothopheded.
- If Jacqueline had araincarded a multica, then the perving would be apilbonied.
- If Theresa had gramencinged a unalippla, then the fammutie would be garbetered.
- If Jack epickyos a trancasea, then the trangaboa is maliberied.
- If Betty had parachered a pneadom, then the recidete would be tratishioniced.
- If Judy had gresbatoed the irreambus, then the lactoteta would be sulightmallyed.
- If Jordan had escaturoved the haingiga, then the symptores would be costicelled.
- If Anna hyphibbils the bendioprie, then the forsping is doggedyed.
- If Michael had apningbomied a outsnopted, then the famping would be subdished.
- If Cheryl dimjailigrs a shotuchon, then the counch is geopunged.
- If Mark had sliskeed the chlorber, then the mickeyin would be subplinesed.
- If Logan lodioprs the flaphinati, then the curbeidae is nibbited.
- If Ethan foodulias a tridiontion, then the laorgar is locarignoed.
- If Donna vesymyis the serchetra, then the graetogic is cophydorated.
- If Maria pulgospos the unsaves, then the vermerismy is costwingnoed.
- If Brenda ballordalis the severopous, then the asteravion is artigrammed.
- If Isabella isophops the pserfle, then the ogtbenth is warratched.
- If Marilyn had underelosered the shantimene, then the frummled would be bleceishiaed.
- If Evelyn had camotomycoed the fathrosec, then the carigned would be mictimariaed.
- If Nicole had suppopeded a diblise, then the ancidifortn would be nonfesceised.
- If Marilyn had extrophoed a huzzines, then the overeloscure would be cencifered.
- If Sandra had spisishied the charlo, then the eelchint would be infinomyonied.
- If Jacqueline had demorrereiaed the choterne, then the wryptomes would be clabonsed.

APPENDIX D

OTHER SCIENTIFIC PROJECTS CONDUCTED THIS YEAR

D.1 Classes followed

- Logic, Language and Computation
  `https://studiegids.uva.nl/xmlpages/page/2022-2023-en/search-course/course/98467`
- Introduction to Modal Logic
  `https://studiegids.uva.nl/xmlpages/page/2022-2023-en/search-course/course/98341`
- Rationality, Cognition and Reasoning
  `https://studiegids.uva.nl/xmlpages/page/2022-2023-en/search-course/course/96693`
- Philosophical Logic
  `https://studiegids.uva.nl/xmlpages/page/2022-2023-en/search-course/course/98364`

D.2 NLP and mathematical cognition

PIs. Prof. Stanislas Dehaene (Collège de France) and Dr François Charton (Meta AI).

Description. This work is a follow-up of my Master's internship as well as a preparation for my PhD next year. During my Master[1], we created a vocabulary of mathematical concepts in French and obtained a semantic reprentation of it by computing its GloVe embeddings (Pennington et al., 2014) from all mathematical pages on French Wikipedia. We were then able to show that GloVe can capture a fair amount of mathematical semantics, and split the concepts into relevant clusters (such as number by increasing order of magnitude, geometrical shapes, etc.). Our question then was to compare the representations of mathematical concepts in artificial neural networks and in the human brain (this will actually the topic of my whole PhD).

Outcomes. To compare human representations to those of GloVe, we designed, ran and analysed an online experiment to collect similarity judgements and ratings of concept knowledge. We now need to compare different semantic representations (GloVe for a global corpus, word2vec, embedding layer of a Transformer model, etc.) to find the best fit. After we have done this, we will redo the same analyses in English and publish our results along with the vocabulary (probably in October or November).

D.3 Bayesian model for the cognition of geometry

PIs. Dr Marie Amalric (Universita degli studi di Trento) and Dr Yacin Hamami (Vrije Universiteit Brussel).

[1] See `https://perso.crans.org/sdebray/files/M2InternshipReport.pdf` for more details.

Description. This project is a follow-up of two papers published by Hamami and Amalric (Hamami & Amalric, 2023; Hamami et al., 2021). The goal is to model the way humans search for counterexamples when reasoning with geometric configuration. The results from Hamami and Amalric (2023) and Hamami et al. (2021) show that some general trends and patterns can be found, and we would like to use bayesian models to account for these, following the approach proposed by Tessler et al. (2022) for syllogism.

Outcomes. For the moment, we only have had time to read and learn about bayesian models, as there is a quite steep learning curve to become acquainted with them at first. For this purpose, we spent some time reading and discussing the book McElreath, 2020.

In addition, I spent a week in Liège in May to meet Yacin Hamami. We developed a framework for geometrical inferences and we are now trying to complement it with existing `Python` librairies for bayesian modelling like `pyro`.

LIST OF TABLES

LIST OF FIGURES

BIBLIOGRAPHY

Adams, E. W. (1965). The logic of conditionals. *Inquiry*, *8*(1-4), 166–197.

Adams, E. W. (1970). Subjunctive and indicative conditionals [Publisher: JSTOR]. *Foundations of language*, 89–94.

Adams, E. W. (1996). *A Primer of Probability Logic*. Center for the Study of Language; Inf.

Arehalli, S., Dillon, B., & Linzen, T. (2022). Syntactic Surprisal From Neural Models Predicts, But Underestimates, Human Processing Difficulty From Syntactic Ambiguities [arXiv:2210.12187 [cs]].

Austin, J. L. (1956). Ifs and Cans. *Proceedings of the British Academy*, 109–132.

Balke, A., & Pearl, J. (1994). Probabilistic Evaluation of Counterfactual Queries. *Proceedings of the 12th National Conference on Artificial Intelligence*, *1*, 230–237.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners [arXiv:2005.14165 [cs]].

Caucheteux, C., Gramfort, A., & King, J.-R. (2021a). Disentangling syntax and semantics in the brain with deep networks. *Proceedings of the 38th International Conference on Machine Learning*, *139*, 1336–1348.

Caucheteux, C., Gramfort, A., & King, J.-R. (2022). Deep language algorithms predict semantic comprehension from brain activity [Number: 1 Publisher: Nature Publishing Group]. *Scientific Reports*, *12*(1), 16327.

Caucheteux, C., Gramfort, A., & King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*.

Caucheteux, C., Gramfort, A., & King, J.-R. (2021b). Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3635–3644.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing [Number: 1 Publisher: Nature Publishing Group]. *Communications Biology*, *5*(1), 1–10.

Charton, F. (2021). Linear algebra with transformers.

Charton, F. (2022). What is my math transformer doing? – Three results on interpretability and generalization [arXiv:2211.00170 [cs]].

Charton, F., Hayat, A., & Lample, G. (2021). Learning advanced mathematical computations from examples [arXiv:2006.06462 [cs]].

Cheng, P. W. (1997). From covariation to causation: A causal power theory [Place: US Publisher: American Psychological Association]. *Psychological Review*, *104*, 367–405.

Collins, K. M., Wong, C., Feng, J., Wei, M., & Tenenbaum, J. B. (2022). Structured, flexible, and robust: Benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks [arXiv:2205.05718 [cs]].

Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory & Cognition*, *19*(3), 274–282.

Cummins, D. D. (1995). Naive theories and causal deduction. *Memory & Cognition*, *23*(5), 646–658.

Dasgupta, I., Lampinen, A. K., Chan, S. C. Y., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2022). Language models show human-like content effects on reasoning [arXiv:2207.07051 [cs]].

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [arXiv:1810.04805 [cs]].

Fernbach, P. M., & Darlow, A. (2010). Causal Conditional Reasoning and Conditional Likelihood. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *32*, 1088–1093.

Fernbach, P. M., & Erb, C. D. (2013). A quantitative causal model theory of conditional reasoning [Place: US Publisher: American Psychological Association]. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(5), 1327–1343.

Grice, H. P. (1989). *Studies in the way of words*. Harvard University Press.

Haegeman, L. (2003). Conditional Clauses: External and Internal Syntax. *Mind and Language*, *18*(4), 317–339.

Hamami, Y., & Amalric, M. (2023). *Going round in circles: A cognitive bias in geometric reasoning* (preprint). PsyArXiv.

Hamami, Y., Mumma, J., & Amalric, M. (2021). Counterexample Search in Diagram-Based Geometric Reasoning. *Cognitive Science*, *45*(4).

Helwe, C., Clavel, C., & Suchanek, F. (2021). Reasoning with Transformer-based Models: Deep Learning, but Shallow Reasoning. *3rd Conference on Automated Knowledge Base Construction*.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. v. d., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., … Sifre, L. (2022). Training Compute-Optimal Large Language Models [arXiv:2203.15556 [cs]].

Holtzman, A., West, P., Shwartz, V., Choi, Y., & Zettlemoyer, L. (2022). Surface Form Competition: Why the Highest Probability Answer Isn't Always Right [arXiv:2104.08315 [cs]].

Kassner, N., Krojer, B., & Schütze, H. (2020). Are Pretrained Language Models Symbolic Reasoners Over Knowledge? [arXiv:2006.10413 [cs]].

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, *108*(33), 13852–13857.

Kratzer, A. (1981). Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic*, *10*(2), 201–216.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people [Publisher: Cambridge University Press]. *Behavioral and Brain Sciences*, *40*, e253.

Lample, G., & Charton, F. (2019). Deep Learning for Symbolic Mathematics.

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, *31*(8), 2906–2915.

Lewis, D. K. (1973a). Causation [Publisher: Oxford Up]. *Journal of Philosophy*, *70*(17), 556–567.

Lewis, D. K. (1973b). *Counterfactuals*. Blackwell.

Lewis, D. K. (1976). Probabilities of Conditionals and Conditional Probabilities [Publisher: [Duke University Press, Philosophical Review]]. *The Philosophical Review*, *85*(3), 297–315.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: A cognitive perspective [arXiv:2301.06627 [cs]].

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054.

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schütze, H. (2020). Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, *117*(42), 25966–25974.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). Taylor; Francis, CRC Press.

Millet, J., Caucheteux, C., Orhan, P., Boubenec, Y., Gramfort, A., Dunbar, E., Pallier, C., & King, J.-R. (2022). Toward a realistic model of speech processing in the brain with self-supervised learning [arXiv:2206.01685 [cs, q-bio]].

Nye, M., Tessler, M., Tenenbaum, J., & Lake, B. M. (2021). Improving Coherence and Consistency in Neural Sequence Models with Dual-System, Neuro-Symbolic Reasoning. *Advances in Neural Information Processing Systems*, *34*, 25192–25204.

Oaksford, M., & Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning* [Google-Books-ID: sLetNgiU7ugC]. OUP Oxford.

Pasquiou, A., Lakretz, Y., Hale, J., Thirion, B., & Pallier, C. (2022). Neural Language Models are not Born Equal to Fit Brain Data, but Training Helps [arXiv:2207.03380 [cs]].

Pasquiou, A., Lakretz, Y., Thirion, B., & Pallier, C. (2023). Information-Restricted Neural Language Models Reveal Different Brain Regions' Sensitivity to Semantics, Syntax and Context [arXiv:2302.14389 [cs]].

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language Models as Knowledge Bases? [Publisher: arXiv Version Number: 2]. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2463–2473.

Politzer, G. (2007). Reasoning with conditionals. *Topoi*, *26*(1), 79–95.

Pollock, J. L. (1976). *Subjunctive Reasoning*. Springer Netherlands.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Reichenbach, H. (1956). The Principle of Common Cause. In *The Direction of Time* (University of California Press).

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118.

Skovgaard-Olsen, N., Stephan, S., & Waldmann, M. R. (2021). Conditionals and the hierarchy of causal queries. *Journal of Experimental Psychology: General*, *150*(12), 2472–2505.

Stalnaker, R. C. (1968). A Theory of Conditionals. In W. L. Harper, R. C. Stalnaker, & G. Pearce (Eds.), *IFS* (pp. 41–55). Springer Netherlands.

Stalnaker, R. C. (1970). Probability and conditionals [Publisher: Cambridge University Press]. *Philosophy of science*, *37*(1), 64–80.

Stenning, K., & van Lambalgen, M. (2008). *Human Reasoning and Cognitive Science*. MIT Press.

Stevenson, R. J., & Over, D. E. (1995). Deduction from uncertain premises [Place: United Kingdom Publisher: Taylor & Francis]. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *48A*, 613–643.

Tenney, I., Das, D., & Pavlick, E. (2019). BERT Rediscovers the Classical NLP Pipeline. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601.

Tessler, M. H., Tenenbaum, J. B., & Goodman, N. D. (2022). Logic, Probability, and Pragmatics in Syllogistic Reasoning [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tops.12593]. *Topics in Cognitive Science*, *14*(3), 574–601.

van Rooij, R., & Schulz, K. (2019). Conditionals, Causality and Conditional Probability. *Journal of Logic, Language and Information*, *28*(1), 55–71.

van Rooij, R., & Schulz, K. (2020). Indicative Conditionals. In *The Wiley Blackwell Companion to Semantics*. John Wiley & Sons, Ltd.

van Schijndel, M., & Linzen, T. (2021). Single-Stage Prediction Models Do Not Explain the Magnitude of Syntactic Disambiguation Difficulty. *Cognitive Science*, *45*(6).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need [arXiv: 1706.03762]. *arXiv:1706.03762 [cs]*.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior [arXiv:2006.01912 [cs]].

Wilcox, E. G., Vani, P., & Levy, R. (2021). A Targeted Assessment of Incremental Processing in Neural Language Models and Humans. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 939–952.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., & Zettlemoyer, L. (2022). OPT: Open Pre-trained Transformer Language Models [arXiv:2205.01068 [cs]].