Do large language models and humans have similar behaviors in causal inference with script knowledge?

Xudong Hong* [1,2], **Margarita Ryzhova*** [2], **Daniel Adrian Biondi**[1] **and Vera Demberg**[1,2]
[1]Dept. of Computer Science, Saarland University
[2]Dept. of Language Science and Technology, Saarland University
{xhong,mryzhova,biondi,vera}@lst.uni-saarland.de

Abstract

Recently, large pre-trained language models (LLMs) have demonstrated superior language understanding abilities, including zero-shot causal reasoning. However, it is unclear to what extent their capabilities are similar to human ones. We here study the processing of an event B in a script-based story, which causally depends on a previous event A. In our manipulation, event A is stated, negated, or omitted in an earlier section of the text. We first conducted a self-paced reading experiment, which showed that humans exhibit significantly longer reading times when causal conflicts exist ($\neg A \rightarrow B$) than under logical conditions ($A \rightarrow B$). However, reading times remain similar when cause A is not explicitly mentioned, indicating that humans can easily infer event B from their script knowledge. We then tested a variety of LLMs on the same data to check to what extent the models replicate human behavior. Our experiments show that 1) only recent LLMs, like GPT-3 or Vicuna, correlate with human behavior in the $\neg A \rightarrow B$ condition. 2) Despite this correlation, all models still fail to predict that $nil \rightarrow B$ is less surprising than $\neg A \rightarrow B$, indicating that LLMs still have difficulties integrating script knowledge. Code and data are available at https://github.com/tony-hong/causal-script.

1 Introduction

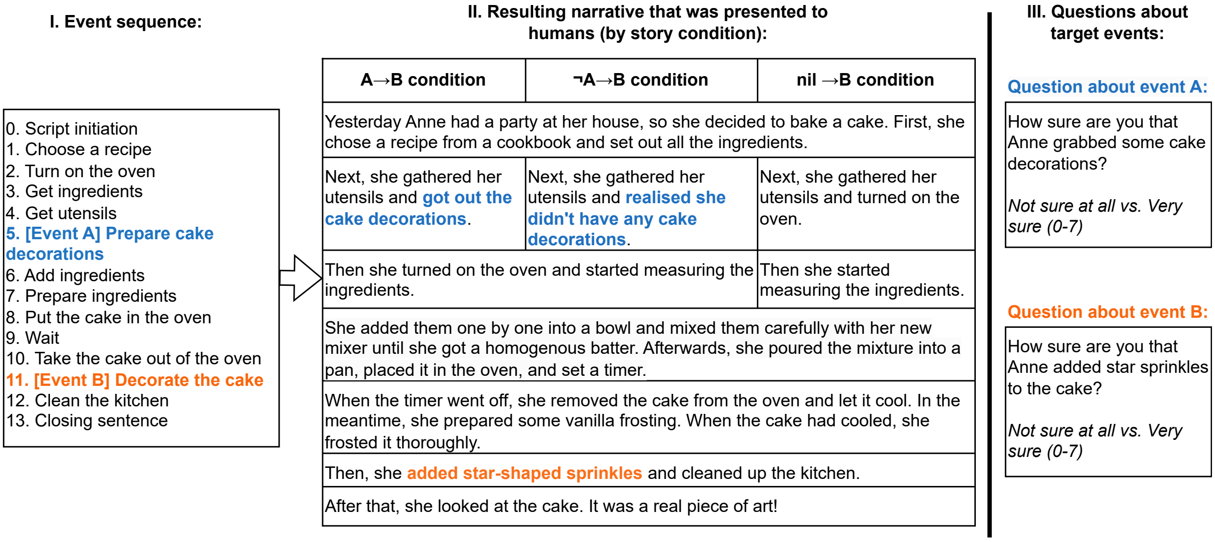
Causal reasoning is fundamental for both human and machine intelligence (Pearl, 2009) and plays an important role in language comprehension (Keenan and Kintsch, 1974; Graesser et al., 1994, 1997; Van den Broek, 1990). Large pre-trained language models (LLMs) such as GPT-3.5 (Neelakantan et al., 2022) have demonstrated excellent zero-shot capabilities in causal reasoning tasks and human-like behaviors (Wang et al., 2019). The capability of causal reasoning is essential to new prompting techniques like the chain-of-thought prompting (Wei et al., 2022; Kojima et al., 2022). On the other hand, some early pieces of evidence show that LLMs lack global planning of different events in stories (Bubeck et al., 2023). So it is unclear to what extent LLMs can conduct causal reasoning about events.

In turn, humans have been shown to be extremely good at building causal connections in long discourse comprehension (Radvansky et al., 2014; Graesser et al., 1994). In doing so, they rely not only on explicit causal links (signaled in the text – see Trabasso and Sperry, 1985; Keenan and Kintsch, 1974) but also on implicit ones that are inferable based on commonsense knowledge (Keenan and Kintsch, 1974; Singer and Halldorson, 1996). In particular, subjects were found to be sensitive to causal conflicts arising from contradictions to earlier text segments or conflicts with subjects' commonsense knowledge (Radvansky et al., 2014; Singer and Halldorson, 1996). An example of a causal conflict is presented in Figure 1, Part II, condition $\neg A \rightarrow B$, where decorating a cake with star-shaped sprinkles is inconsistent with the previously mentioned information that cake decorations are not available.

In this paper, we investigate language processing in humans and compare it to a large variety of LLMs, following the "psycholinguistic assessment of language models paradigm" (Futrell et al., 2019). In our analyses, we compare human reading times to LLM surprisal estimates. Surprisal is the negative log probability of a word in context and has been previously related to human reading times (Hale, 2001; Levy, 2008; Demberg and Keller, 2008; Smith and Levy, 2013) as well as neuropsychological effects such as the N400 (Frank et al., 2015; Kutas and Hillyard, 1989), which represent human processing difficulty. We collect a new dataset, Causality in Script Knowledge (**CSK**),

* These authors contributed equally to this work.

*Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 421 - 437
June 20-21, 2024 ©2024 Association for Computational Linguistics



I. Event sequence:

0. Script initiation
1. Choose a recipe
2. Turn on the oven
3. Get ingredients
4. Get utensils
5. [Event A] Prepare cake decorations
6. Add ingredients
7. Prepare ingredients
8. Put the cake in the oven
9. Wait
10. Take the cake out of the oven
11. [Event B] Decorate the cake
12. Clean the kitchen
13. Closing sentence

II. Resulting narrative that was presented to humans (by story condition):

A→B condition	¬A→B condition	nil →B condition
Yesterday Anne had a party at her house, so she decided to bake a cake. First, she chose a recipe from a cookbook and set out all the ingredients.		
Next, she gathered her utensils and got out the cake decorations.	Next, she gathered her utensils and realised she didn't have any cake decorations.	Next, she gathered her utensils and turned on the oven.
Then she turned on the oven and started measuring the ingredients.		Then she started measuring the ingredients.
She added them one by one into a bowl and mixed them carefully with her new mixer until she got a homogenous batter. Afterwards, she poured the mixture into a pan, placed it in the oven, and set a timer.		
When the timer went off, she removed the cake from the oven and let it cool. In the meantime, she prepared some vanilla frosting. When the cake had cooled, she frosted it thoroughly.		
Then, she added star-shaped sprinkles and cleaned up the kitchen.		
After that, she looked at the cake. It was a real piece of art!		

III. Questions about target events:

Question about event A:

How sure are you that Anne grabbed some cake decorations?

Not sure at all vs. Very sure (0-7)

Question about event B:

How sure are you that Anne added star sprinkles to the cake?

Not sure at all vs. Very sure (0-7)

Figure 1: Example of a script structure (I), the resulting narrative in three conditions (II) and questions that subjects were asked (III), for "baking a cake" story.

consisting of short stories about daily activities which are typically part of the *script knowledge* of humans, see Figure 1 for an example. The term “script knowledge” refers to commonsense knowledge about everyday activities, where “scripts” are defined as prototypical sequences of events in these activities. The stories are constructed such that they contain a pair of events, A and B which are causally contingent on one another. We manipulate event A to be stated, negated or omitted, and subsequently measure reading times on event B.

Our first research question (**RQ1**) relates to the effect of the incoherence in the $\neg A \rightarrow B$ condition, compared to the coherent $A \rightarrow B$ condition. For humans, a large body of previous literature (Bloom et al., 1990; Radvansky et al., 2014; Singer and Ritchot, 1996) leads us to expect that human readers will notice the inconsistency and that this can be measured in terms of slower reading times on event B. For language models, we want to test whether and which models also exhibit a similar effect, by comparing the surprisal values for the words of event B following the A vs. $\neg A$ mentioned in the previous context. In order for a language model to handle this case, it needs to (a) understand the contingency between events A and B (even though they often don’t use overlapping lexical items) and (b) be able to represent event A or $\neg A$ effectively across the intervening sentences so it is still represented when encountering B. We find that the large models (GPT-3 and Vicuna) do well on this task, but smaller models mostly fail.

Our second research question (**RQ2**) aims to tap into how script knowledge facilitates language comprehension. To this end, we compare the processing of event B in a setting where neither event A nor event $\neg A$ are mentioned in the previous context. If comprehenders integrate their script knowledge with the text, they should have an easy time processing event B even without the prior mention of event A (Bower et al., 1979). The previous literature on human sentence processing has no direct evidence about the processing difficulty of event B in this case, so here our experiment makes a new contribution: we find that humans are significantly faster in reading segment B in the $nil \rightarrow B$ condition compared to $\neg A \rightarrow B$, and that reading times between conditions $nil \rightarrow B$ and $A \rightarrow B$ do not differ significantly from one another. Our subsequent evaluation of LLMs on the same contrast however shows that all LLMs fail to show human-like processing: they do not have lower surprisal on the $nil \rightarrow B$ condition than on $\neg A \rightarrow B$ – some models even assign higher surprisal estimates to the $nil \rightarrow B$ condition, indicating that even the most recent large LLMs in our evaluation cannot effectively integrate script knowledge for estimating the probability of upcoming words.

2 Background

2.1 Causal inference and script knowledge

When humans read text, they connect events mentioned in the text into a locally and globally coherent causal network, thereby not only integrating information from the text but also based on context and commonsense knowledge (Van den Broek, 1990; Graesser et al., 1997). It has been shown that when the causal network does not support new events or the new event contradicts the previous text, readers experience processing difficulties, resulting in longer reading times (Bloom et al., 1990; Radvansky et al., 2014). The comprehension of a new event also relies on commonsense knowledge (Hare et al., 2009). In fact, Singer and Ritchot (1996) showed that when commonsense knowledge does not support an event described in the text, comprehenders take more time processing it.

A special type of commonsense knowledge that was shown to also modulate reading comprehension is script knowledge (Abbott et al., 1985; Bower et al., 1979; Schank, 1975). Scripts represent knowledge structures consistent with sets of beliefs built on past experiences about everyday, routine, and conventional activities like baking a cake. Importantly, the events constituting a script can be highly causally inter-connected and are crystallized in memory – one can expect script-related events to be activated once the script is invoked. In a series of experiments, Bower et al. (1979) showed that after subjects read an everyday story that constituted a script, they also recalled script-related events that were not explicitly mentioned in the story (see Gibbs and Tenney, 1980, for similar findings showing that script knowledge is an indistinguishable part of the memory representation). In turn, it is expected that when reading a story, script-related events can be primed by the script itself rather than by some single events mentioned in the text, without processing time loss (Keenan and Kintsch, 1974).

2.2 Experiments with language models

Causal Reasoning. Recent LLMs such as GPT-3.5 (Neelakantan et al., 2022) have achieved strong performance in many reasoning tasks under zero-shot settings, such as symbolic reasoning, logical reasoning, mathematical reasoning and commonsense inference (Kojima et al., 2022). The common practice to conduct zero-shot reasoning is *prompting*, i.e. to append a task-specific text to the input to LLMs and then sample the output (Radford et al., 2019). Although the cause is usually provided in the prompt (like condition $A \to B$), LLMs can reason without relying only on surface cues like word overlap (Lampinen et al., 2022). Moreover, LLMs can be prompted to produce explicit reasoning steps with chain-of-thought prompting (Wei et al., 2022).

Script knowledge. Early works regarding script knowledge also apply language models (LMs). Weber et al. (2020) apply LMs for script induction from causal effects. Ciosici et al. (2021) build a human-LM collaborative system for script authoring.

Recent studies have suggested that LLMs may learn script knowledge as part of their training (Sakaguchi et al., 2021; Sancheti and Rudinger, 2022). Ravi et al. (2023) fine-tune GPT-3 to automatically generate plausible events that happen before and after a given event, and Yuan et al. (2023) report promising results on prompting an InstructGPT model (Ouyang et al., 2022) to automatically generate scripts and then filtering results in the second step. Similarly, Brahman et al. (2023) use a distilled small LM as script planner and fine-tuned RoBERTa as verifiers.

There are however also reports that indicate that script knowledge in LLMs may not yet be sufficient: zero-shot probing on GPT-2 has been found to generate poor event sequences (Sancheti and Rudinger, 2022), and GPT-3 was found to be only marginally better than chance on predicting event likelihoods (Zhang et al., 2023) and exhibit poor performance on event temporal ordering (Suzgun et al., 2023).

Several ways of specifically integrating commonsense knowledge into LLMs have been proposed: some LLMs are trained from scratch on structural data with commonsense knowledge like knowledge graphs (ERNIE; Zhang et al., 2019) and semantic frames (SpanBERT; Joshi et al., 2020). Bosselut et al. (2019); Hwang et al. (2021) further equips LLMs with structural input and output to model commonsense knowledge. In the present contribution, we explore previous models that have been reported to be successful in inference tasks. More details of the choice of LLMs are in Section 4.1.

2.3 The TRIP dataset

A dataset that is particularly relevant to the present study is the TRIP dataset, which contains 1472 pairs of two similar stories, which differ by one sentence at a "breakpoint" position (Storks et al., 2021). One of the stories is plausible, and the other one is implausible, due to a causal conflict between the sentence at the breakpoint position and an earlier part of the text. The plausible stories correspond to the $A \rightarrow B$ condition in our dataset, while the implausible stories correspond to our $\neg A \rightarrow B$ condition. The breakpoint sentence corresponds to our critical sequence B.

Richardson et al. (2022) fine-tune a T5 model augmented with logical states of each event to detect the causal conflicts and outperform a RoBERTa baseline by a large margin. Ma et al. (2022) finetune a framework to integrate global and local information. Our aim is not to finetune the LLMs on TRIP but to test them in a zero-shot fashion.

3 Experiments with Humans

3.1 Dataset

The Causality in Script Knowledge (**CSK**) dataset consists of 21 English stories describing everyday activities like baking a cake or taking a bath.[1]

To construct the stories, we initially composed sequences of script-related events that were built on top of Wanzare et al. (2016) – see Figure 1, part I. Subsequently, we transformed these sequences into narrative form (Figure 1, part II; for example, an event "prepare cake decorations" is realised in the narrative as "she got out the cake decorations"). Further, each story was divided into chunks of text (as rows of the table in Figure 1, part II) such that participants do not see the whole text at once, but chunk after chunk.

Each story starts with script initiation – a sentence in the first chunk that introduces the topic to the reader, e.g., "*Yesterday Anne had a party at her house, so **she decided to bake a cake**.*" from Figure 1, part II. Thus, readers can already activate script knowledge about the event at that point.

A pair of events A and B represent our main interest. They were chosen in such a way that event A ("get the cake decorations") enabled the happening of event B ("add star-shaped sprinkles"). More specifically, since scripts are typically characterized by event sequences in which specific script participants appear repeatedly (like cake decorations), we are interested in a pair of events that define an action done to this specific participant.

In some stories, participants related to the target manipulation have different lexical realization between events A and B. For example in the cake story presented in Figure 1, a participant in event A is referred to as "cake decorations" and in event B it is specified as "star-shaped sprinkles" (as a type of cake decorations). Some stories also necessitate an inference e.g. from referring to "bubble bath" in event A and "foam" in event B. In other stories, identical referring expressions were used in events A and B (e.g., in a grocery story, event A: "he took a shopping cart" vs. event B: "he put everything in his shopping cart").

parameter	mean	sd
# of words in story:		
$A \rightarrow B$	158.2	12
$\neg A \rightarrow B$	159.1	14
$nil \rightarrow B$	150.1	11.7
# of text chunks in story	6.8	0.77
# of words in chunk with A	27.6	11.3
# of words in chunk with $\neg A$	29.3	13.1
# of words in chunk with B	12.9	1.7
# of words in chunk after B	12.9	1.8
# of words b/w A and B:		
$A \rightarrow B$	73.6	10.3
$\neg A \rightarrow B$	71.8	12.9
# of words in A	7.3	3.8
# of words in $\neg A$	11.2	5.3
# of words in B	5.4	1.6

Table 1: Decriptive statistics for stories.

Importantly, no other events in the story draw a direct causal link to event B, except event A and the script itself. Events A and B are always separated by descriptions of other script events (73.6 words on average; $sd = 10.3$; min: 59; max: 91). The chunk with event B always consists of one sentence with the following structure: "*ADVERB PERSON X did action B and then did a subsequent action from the script sequence.*" (except the laundry story, where the sentence started with "She"). When constructing the experimental materials, we controlled for the following parameters: the number of words and text chunks in a story, the number of text chunks and words between events A and B, the number of words in the text chunks that contained event B, and number of words in the chunk after the chunk with event B. The full list of descriptive statistics for our materials is presented in Table 1.

3.2 Experimental conditions

Our target manipulation relates to the appearance of events A and B in the story thus producing three different story conditions:

Condition $A \rightarrow B$. Event B logically follows event A within the story context. In this way, event

[1] Available at https://github.com/tony-hong/causal-script

A draws a direct causal link to event B, and thus event B is anticipated to happen on the basis of event A.

Condition $\neg A \rightarrow B$**.** Event A is negated, making the occurrence of event B implausible or even impossible. The mention of event B thus is unexpected and stands in a causal conflict with the earlier information. While creating negation of events A, we had the following strategy. Since events A and B in our materials typically share at least one common event participant, in the $\neg A$ condition, this participant was made unavailable for event B. In this way, the causal link between A and B (prepare cake decorations → add star-shaped sprinkles; put a pillow in the backpack → take it from the backpack) is broken because event $\neg A$ changes the state of the participant so that it is not available in B (when one doesn't have a travelling pillow, this script participant is not going to be available in B to take it from the backpack).

The $\neg A$ condition did not always consist of literal negation with the word "not" but as in the example shown in Figure 1 (A: "she got out the cake decorations" vs. $\neg A$: "she realised she **didn't have** any cake decorations"), but while in other stories, participant in event A was disabled in a more subtle way, via verbs of implicit negation or particles like "only", e.g., (events A vs. $\neg A$):

- (sunscreen): she grabbed her sunscreen VS. she forgot her sunscreen
- (pocorn buckets): she bought three buckets of popcorn for everyone VS. since nobody was hungry, she just bought drinks for everyone

Condition $nil \rightarrow B$**.** Event A is omitted. Even though event A is not explicitly stated, it is expected that humans will easily infer its occurrence from the context, making the mention of event B plausible and easy to integrate (Bower et al., 1979).

3.3 Experimental procedure

For data collection, each story was divided into paragraphs or text chunks (as shown, for example, in Figure 1, part II). During the experiment, subjects saw only one paragraph at a time (chunk-by-chunk presentation). After reading each story, subjects had to rate how sure they were about the events A and B to have occurred, on a Likert scale ranging from 0 (*Not sure at all*) to 7 (*Very sure*) – see Figure 1, part III. To measure the processing difficulties of humans, we compare the reading times for event B across the experimental conditions. More details about subjects' belief ratings are presented in Appendix A.

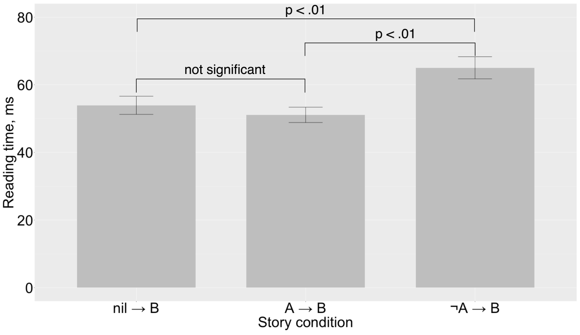


Figure 2: Human results. Mean by-character reading times at event B, by story condition; p-values are taken from the corresponding LMER models, see Section 3.5.

251 native English speakers were hired via the crowdsourcing platform Prolific[2] to participate in the study. Each participant read three stories. Each story had a different topic and was presented in a different condition.

3.4 Analysis

To investigate the effects of processing difficulty that event B causes in subjects depending on story condition, we analyse mean per character reading times associated with the chunks that contain event B. The log-transformed reading times were analysed using linear mixed-effects regression models (LMER; Bates et al., 2015). The maximal random effects structure included by-subject and by-item random intercepts and by-item random slopes for story condition and was simplified for convergence when needed.

Prior to the analysis, we removed all trials related to the bowling story item, due to a typo. Further, we removed trials where the reading times in the chunk containing event B were shorter than 100ms or larger than 50s. 704 trials from 251 subjects (73% female; mean age = 40, sd = 14.6, [18;80] range) were available for analysis (1.81% data loss).

3.5 Results

To answer to what extent causal inconsistencies are reflected in human language processing (RQ1), we compared reading times on segment B in the $A \rightarrow B$ vs. $\neg A \rightarrow B$ conditions. The random effects structure included by-subject and by-item random intercepts and by-item random slopes for story

[2] https://www.prolific.co/

conditions. We found that subjects read chunks with event B significantly more slowly when event A was explicitly negated in the story ($b = 0.21$, $se = 0.04$, $t = 4.77$, $p < .01$), see also Figure 2.

To analyse subjects' ability to infer causal links from script knowledge (RQ2), we compared the reading times in $nil \rightarrow B$ vs. $A \rightarrow B$ conditions. The random effects structure included by-item random intercepts. We observed no significant difference between these conditions ($b = -0.04$, $se = 0.05$, $t = -0.7$, $p = .48$). Thus, the absence of event A, which serves as a direct causal link to event B, does not slow event's B processing in terms of reading times. Note that the reading time of condition $\neg A \rightarrow B$ is significantly slower than the reading time in condition $nil \rightarrow B$ ($b = 0.17$, $se = 0.05$, $t = 3.23$, $p < .01$).

4 Can LLMs Detect Causal Conflicts (RQ1)?

In this section, we measure the ability of different LLMs to track event contingency. We feed the script stories into the language models and record the LM's surprisal scores on a word-by-word basis. We then test whether the mean surprisal scores for the critical region (event B) differ between conditions. As the script stories corpus is relatively small, we additionally test the models on the TRIP dataset (Storks et al., 2021) to assess their recognition of causal incongruencies on a wider set of materials (see Section 4.5).

4.1 Choices of LLMs

We select a set of 20 causal language models (CLMs).[3] We chose the GPT-1/2/3 and Instruct GPT models (Radford et al., 2018, 2019; Brown et al., 2020; Ouyang et al., 2022) because of their good performance on many NLP tasks (Chang and Bergen, 2023). We also selected GPT-3.5 (Neelakantan et al., 2022) because it was trained with both programming code and text and as a result demonstrated strong performance on entity tracking (Kim and Schuster, 2023), a prerequisite for causal reasoning. Notably, ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023) can not be used in our study, because the API does not allow access to the probabilities. Additionally, we used Vicuna models (Chiang et al., 2023), a LLaMa-based model (Touvron et al., 2023) fine-tuned on 70K user-shared ChatGPT conversations. Open models like Vicuna have the advantage of results being reproducible. Similarly, we chose OPT (Zhang et al., 2022) and GPT-Neo (Black et al., 2021) as open models similar to GPT-3.

We also selected task-specific models that could potentially capture script knowledge via exposure to more diverse datasets like summarization models, Pegasus (Zhang et al., 2020), Bigbird-pegasus, and a multilingual model XGLM (Lin et al., 2022). Lastly, we chose XLNet because it has been previously shown to be effective for zero-shot script parsing (Zhai et al., 2021, 2022) wrt. handling causal inferences in commonsense stories in a zero-shot setting.

All models used here were available through either HuggingFace or the OpenAI API. More details are in Appendix B, where we briefly describe all the models.

4.2 Method

We perform word-by-word next-word prediction for event B, recording the next token probabilities for each token in segment B. Based on the probability of the target words w given the story context, we then calculate the target tokens' surprisal as their negative log probability: $\mathrm{surprisal}(w) = -\log P(w|\mathrm{story\_context})$. We then calculate the average per-word surprisal by averaging the surprisal of each word into an estimate of the surprisal of the critical region for each item.

4.3 Data Analysis

To identify the PLM(s) that show comparable effects to humans, we run an equivalent analysis to how the reading time data were analysed: we estimate linear mixed effects models with surprisal as a response variable and condition ($A \rightarrow B$ vs. $\neg A \rightarrow B$) as a predictor. The model also includes by-item random intercepts. The formula is: $\log(\mathrm{surprisal}) \sim \mathrm{story\_condition} + (1|\mathrm{story})$[4].

4.4 Results

Table 2 (column CSK) presents the results for all language models on whether model surprisals were significantly higher for the $\neg A \rightarrow B$ condition than in the $A \rightarrow B$ condition, indicating that the model's surprisal scores reflect the incoherence (RQ1). High positive b values indicate that surprisal values are higher on segment B in

[3] We also experiment with masked language models. Please refer to Appendix C.1.

[4] Log surprisals were chosen because of the skewed distribution of surprisal values.

Model Name	# para.	CSK			TRIP		
	(M)	b	t	sign	b	t	sign
GPT-3.5: text-davinci-003	175K	0.59	5.87	***	0.30	10.82	***
GPT-3.5: text-davinci-002	175K	0.51	2.75	*	0.26	7.41	***
InstructGPT: text-davinci-001	175K	0.26	2.03	·	0.29	5.81	***
InstructGPT: davinci-instruct-beta	175K	0.21	2.76	*	0.20	8.68	***
GPT-3: davinci-002	175K	0.28	4.36	***	0.35	8.11	***
GPT-3: davinci	175K	0.21	2.76	*	0.20	8.25	***
Vicuna-13B	13016	0.22	2.25	*	0.26	7.56	***
Vicuna-7B	6738	0.28	2.56	*	0.22	6.35	***
InstructGPT: text-curie-001	6700	0.03	0.31	n.s.	0.19	5.78	***
GPT-3: curie	6700	0.23	3.43	**	0.12	5.92	***
GPT-2: XL	1638	0.05	0.96	n.s.	0.06	3.15	**
GPT-2: L	838	0.04	0.77	n.s.	0.05	2.77	**
XGLM	827	-0.03	-0.79	n.s.	0.02	1.38	n.s.
Bigbird-pegasus-large-arxiv	470	0.06	1.20	n.s.	0.00	-0.02	n.s.
Pegasus-large	467	0.02	0.85	n.s.	0.00	-0.48	n.s.
XLNet-large-cased	393	-0.03	-1.99	·	0.00	0.66	n.s.
OPT	357	0.01	0.12	n.s.	0.03	1.78	·
GPT-Neo	164	0.03	0.67	n.s.	0.01	0.90	n.s.
GPT-2	163	0.00	-0.10	n.s.	0.01	0.53	n.s.
GPT: openai-gpt	148	0.00	-0.01	n.s.	0.05	3.18	**

Table 2: Results for RQ1 ($A \rightarrow B$ versus $\neg A \rightarrow B$) on CSK (original and intervention removal) and TRIP dataset. The # para. (M) column shows the number of parameters in millions. n.s. represent that the results are not statistically significant. The ·, *, **, and *** in the sign column represent p-values < .1, .05, .01, and .001.

the $\neg A \rightarrow B$ condition compared to the $A \rightarrow B$ condition. Significance stars indicate whether the differences were statistically reliable. Our results show that only some of the largest models showed a reliable increase in surprisal estimates for the incoherent ($\neg A \rightarrow B$) condition.

GPT-3.5: text-davinci-003 shows the largest effect with high statistical reliability. Further models that show the expected behaviour include other versions of GPT-3/GPT-3.5 and the Vicuna model. GPT-3: davinci-002 has the largest effect amoug the GPT-3 models. Surprisingly, InstructGPT models that are trained with human-selected samples don't show significant effects. This result implies additional training on high-quality samples harms the models' ability to identify causal conflicts.

4.5 Experiments on TRIP dataset

As the CSK dataset, for which we collected reading times, is relatively small, we also compared the surprisals of the same set of models on the substantially larger TRIP dataset (cf. Section 2.3), which also contains causal inconsistencies. Their dataset has multiple splits. We only use the "ClozeDev" split. (We do not use the "Order" splits, in which the order of the sentences is switched, because that setting is too different to our dataset.)

	nil vs. $\neg A$			nil vs. A		
Model Name (CLMs only)	b	t	sign	b	t	sign
GPT-3.5: text-davinci-003	0.08	0.77	n.s.	-0.52	-5.10	***
GPT-3.5: text-davinci-002	-0.06	-0.38	n.s.	-0.57	-3.65	***
InstructGPT: davinci-instr-beta	-0.17	-1.96	.	-0.39	-4.36	***
GPT-3: davinci-002	-0.15	-1.94	.	-0.43	-5.60	***
GPT-3: davinci	-0.15	-1.79	.	-0.36	-4.34	***
Vicuna-13B	-0.15	-1.52	n.s.	-0.37	-3.73	***
Vicuna-7B	-0.07	-0.58	n.s.	-0.36	-2.91	**
GPT-3: curie	-0.23	-2.74	**	-0.46	-5.54	***
Human	0.17	3.23	**	-0.04	-0.7	n.s.

Table 3: Results for RQ2 ($nil \rightarrow B$ versus $\neg A \rightarrow B$ and $A \rightarrow B$) on CSK dataset. Note that coefficient estimates for human data refer to log reading times, and are hence not directly comparable to the numbers in the CLMs, which estimate the surprisal effect.

We again estimated surprisal values for each language model, in the same way as described in section 4.2. The critical segment B for this dataset corresponds to the breakpoint sentence. The analysis was analogous to the analysis for the CSK dataset.

Column "TRIP" in Table 2 presents the results of our method on the TRIP dataset. Significant positive effects indicate a significant difference between the model surprisals in the implausible condition compared to the plausible one, indicating that the model recognized the inconsistency correctly.

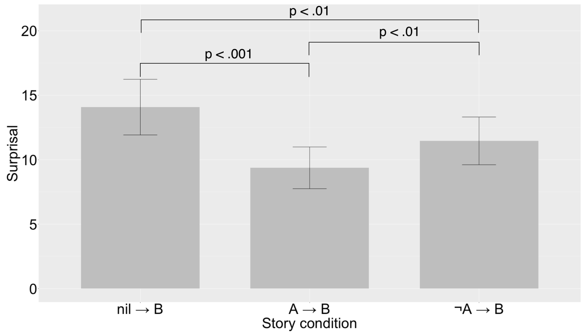


Figure 3: Performance of GPT-3: curie in both research questions. Mean surprisal presented by story condition; p-values are taken from Tables 2 and 3.

GPT-3.5 performs notably well, again displaying the largest effect size and p-value $< .001$.

4.6 Discussion

Given the analysis of the CSK and TRIP datasets, we conclude that only some of the GPT models were able to consistently assign higher surprisal to event B (or the breakpoint sentence in TRIP) in the case that causally related event A was negated earlier in the story[5]. Among the GPT models, we find that GPT-3.5: text-davinci-003 shows the most consistent performance. It differs from the others in that it was trained using reinforcement learning from human feedback, which has been found to be correlated with better performance on many reasoning tasks (Chang and Bergen, 2023).[6]

5 Do LLMs incorporate script knowledge (RQ2)?

In this section, we are interested in whether the models that can capture the causal link between A and B are also able to integrate script knowledge to a similar extent as humans, i.e. whether they show a relatively low surprisal even if event A was not explicitly mentioned in the story context. We continue with those models showing a significant effect of the $\neg A \rightarrow B$ condition compared to $A \rightarrow B$ consistently across the CSK and the TRIP datasets, as these are the only models that seem to reliably deal with negation and capture the causal link.

5.1 Analysis and Results

Analysis was performed using linear mixed-effects models (LMER), similar to Section 4.3. This time, we compare surprisal estimates of conditions $nil \rightarrow B$ to $\neg A \rightarrow B$ to show firstly whether the model correctly captures the incongruency of $\neg A \rightarrow B$. Next, we compare condition $nil \rightarrow B$ to condition $A \rightarrow B$ in order to determine whether the models are consistent with human readers in terms of NOT showing a large effect. The formula of each LMER model is: $\log(\text{surprisal}) \sim \text{story\_condition} + (1|\text{story})$.

Table 3 shows the results for research question 2. While humans read sequence B significantly faster in the $nil \rightarrow B$ condition than in the condition with the causal conflict ($\neg A \rightarrow B$), none of the language models show this effect: most models do not show a significant difference between these conditions, and one model (GPT-3: curie) in fact shows significant effects in the wrong direction (B has higher surprisal in the nil condition than in the $\neg A$ condition), see also Figure 3. This might indicate that the lexically related material in condition $\neg A$ (e.g., "cake decorations") leads to a relatively low surprisal at region B even if it stands in causal conflict with it.

The significantly lower surprisal in condition $A \rightarrow B$ compared to condition $nil \rightarrow B$, which is observed in all of the models, furthermore indicates that models fail to include script knowledge effectively in their next word predictions – current models hence differ from humans in their ability to use script knowledge for predicting (or easily integrating) script-inferable event participants.

5.2 Can models capture negation?

As pointed out by an anonymous reviewer, models' inability to show human-like behavior in RQ2 might be due to models failing to process negation properly, even though these models show significantly lower surprisal in $A \rightarrow B$ condition compared to $\neg A \rightarrow B$ condition. Previous literature indeed shows that transformers have trouble with (explicit) negation (Nguyen et al., 2023). Considering that our materials contain various formulations of event $\neg A$ (including in some cases explicit and in other cases implicit negation), which could pose difficulty to LLMs, we conduct a follow-up study to see whether the best models from the RQ2 experiment could properly identify a participant's state in $\neg A$, i.e., its unavailability. There are actually

[5] One possible reason for this can be models' inability to handle long dependencies between events A and B. We investigate it in Appendix C.2

[6] We did not apply a correction for multiple testing in the analysis. If we were to more conservatively account for multiple testing, then the results of most models except for GPT-3.5: text-davinci-003 would not be judged as statistically reliable.

two other possibilities as to why models might fail in negation processing. First, considering that not all of our stories contain exact lexical realizations of target participants between events A and B, the models can fail to match the negated participant in $\neg A$ ("she realized she didn't have any **cake decorations**") to its realization in event B ("she added **star-shaped sprinkles**"). Secondly, since there is still some context between events A and B (see Table 1), the models can 'forget' the state of the target participant by the time they reach event B. Previous literature shows that participant state tracking can be a difficult task for LLMs (Kim and Schuster, 2023).

We construct questions about the availability of the target participant from event B, e.g., "Are **cake decorations** available to Anne?" (the correct answer is 'yes' in $A \rightarrow B$ condition and 'no' in $\neg A \rightarrow B$). For each story and model, we assess this question twice: directly after event A and just before event B, to capture a potential problem of 'forgetting' about a participant's state. If the participant's lexical realization was different between events A and B, we also assess the same question but about the target participant as it was instantiated in event B: "Are **star-shaped sprinkles** available to Anne?").

We then test the best available models from RQ2, namely GPT-3.5: gpt-3.5-turbo-instruct and GPT-3: davinci-002[7]. Since GPT-3 models were not specifically trained to follow user instructions, we utilized the approach of Brown et al. (2020) for the GPT-3: davinci-002 model: we compared the probabilities of "Yes" and "No" as input tokens following the question and chose the answer with the higher output probability to compare with a correct answer. In the case of the GPT-3.5: gpt-3.5-turbo-instruct model, we prompt the model to generate "Yes" or "No" answers with an instruction *Please answer with "Yes" or "No"* and compare the output with a correct answer (as this model only allows text output).

The results show that the gpt-3.5-turbo-instruct model reaches an accuracy of more than 90% in this task on each question formulation, which shows that it is well capable of processing negation and tracking participant state. On the other hand, the GPT-3: davinci-002 model succeeds in tracking participant state but exhibits very low accuracy in capturing negations, which indicates that older GPT-3 models can not capture negation. We conclude that these experiments confirm the interpretation that older models fail to represent negation properly and hence fail on RQ1. In the meantime, larger models have no problem understanding negations. They fail on RQ2 due to a failure in activating script knowledge to a similar extent as humans wrt. anticipating or easily integrating a script-predictable participant.

6 Conclusions

In this paper, we inspect the behaviors of both large language models and humans in zero-shot causal inference. We conducted a self-paced reading experiment on common sense stories to inspect human processing difficulty when reading the stories. Reading time results indicate that humans stumble across causally incoherent text segments, exhibiting longer reading times in these cases. On the other hand, they easily integrate script-predictable information, even if the explicit causal component (event A) is missing from the story.

When we apply the same study to LLMs, only the newest LLMs show similar behavior to humans on encountering casual conflicts. All models fail to replicate human behaviors when the cause is omitted. Even models trained with programming code and instructions fail to make use of script knowledge, which indicates that script knowledge may not be represented sufficiently well in the LLMs tested in this study.

7 Limitations

One limitation from the NLP perspective of our study is that the size of the CSK dataset is small and only in English (only 21 stories). This is a very common limitation of psycholinguistic studies due to the costs of human experiments. We here addressed this shortcoming by also evaluating on the larger dataset TRIP, but a dataset with more stories or more readers would further improve the reliability of the results. Another limitation is that we don't experiment with few-shot examples in prompts, which could have been used to remind the LLMs to make use of script knowledge. We chose the zero-shot setting because humans use script knowledge for casual inference without any "examples" and

[7] Because this additional experiment is conducted as a reaction to reviews, some OpenAI models in RQ1 and RQ2 have become deprecated in the meantime. Here we report the performance of the official replacement gpt-3.5-turbo-instruct for all GPT-3.5 and InstructGPT models; see the OpenAI documents: https://platform.openai.com/docs/deprecations/instructgpt-models.

we believe that the LLMs should have the same behaviors as humans. However, this means that our results do not necessarily generalize to other ways of prompting models. Additionally, we didn't experiment with the most recent OpenAI models like GPT-4 because their official API doesn't support generating the probability output for given text input. Lastly, we didn't test models with more than 20B parameters on our own server due to limited hardware resources.

Another limitation of our experiment is that we cannot comment on the generalizability of our script materials to more general script-based stories for scripts that may be less well-known to human readers. For our materials, we asked participants after each experimental trial whether they were familiar with the script ("Please tick this box if you have never baked a cake or you have very little experience with it)". Participants answered in 11.2% of trials that they were not familiar with the script. We observed an effect of familiarity on reading times, showing that subjects read the story faster when they were not familiar with the topic. We note that findings also remained stable when we removed such trials from our analysis.

8 Ethics Statement

We release our CSK dataset under the CC BY-NC-SA license. We anonymize the dataset to protect participants' identities. The human study was approved by the ethics committee of Deutsche Gesellschaft für Sprachwissenschaft (DGfS). All participants were paid fairly according to the local standard.

The TRIP dataset was released under an unknown license but the paper described this dataset was published in an ACL proceeding. We use it for academic purposes only.

The potential risk of this work is that the findings can be used to design attacks on LLMs to harm their capability of conducting casual inference given script knowledge (Alzantot et al., 2018).

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG), Funder Id: http://dx.doi.org/10.13039/501100001659, Project-ID 232722074 – SFB1102: Information Density and Linguistic Encoding. We sincerely thank the anonymous reviewers for their insightful comments that helped us to improve this paper. We thank Sebastian Schuster, Alexandra Mayn and Josef van Genabith for their informative advice and Nina Shvetsova for her help with the Vicuna models. We also thank students Teresa Martín Soeder, Alice Virginia Chase, and Brianna Michelle Lehman for helping with materials construction.

References

Valerie Abbott, John B Black, and Edward E Smith. 1985. The representation of scripts in memory. *Journal of memory and language*, 24(2):179–199.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67:1–48.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. *Zenodo*.

Charles P Bloom, Charles R Fletcher, Paul Van Den Broek, Laura Reitz, and Brian P Shapiro. 1990. An on-line assessment of causal reasoning during comprehension. *Memory & cognition*, 18:65–71.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Gordon H Bower, John B Black, and Terrence J Turner. 1979. Scripts in memory for text. *Cognitive psychology*, 11(2):177–220.

Faeze Brahman, Chandra Bhagavatula, Valentina Pyatkin, Jena D Hwang, Xiang Lorraine Li, Hirona J Arai, Soumya Sanyal, Keisuke Sakaguchi, Xiang Ren, and Yejin Choi. 2023. Plasma: Making small language models better procedural knowledge models for (counterfactual) planning. *arXiv preprint arXiv:2305.19472*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Tyler A Chang and Benjamin K Bergen. 2023. Language model behavior: A comprehensive survey. *arXiv preprint arXiv:2303.11504*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Rune Haubo B Christensen. 2018. Cumulative link models for ordinal regression with the r package ordinal. *Submitted in J. Stat. Software*, 35.

Manuel Ciosici, Joseph Cummings, Mitchell DeHaven, Alex Hedges, Yash Kankanampati, Dong-Ho Lee, Ralph Weischedel, and Marjorie Freedman. 2021. Machine-assisted script curation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 8–17, Online. Association for Computational Linguistics.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The erp response to the amount of information conveyed by words in sentences. *Brain and language*, 140:1–11.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Raymond W Gibbs and Yvette J Tenney. 1980. The concept of scripts in understanding stories. *Journal of Psycholinguistic Research*, 9:275–284.

Arthur C Graesser, Keith K Millis, and Rolf A Zwaan. 1997. Discourse comprehension. *Annual review of psychology*, 48(1):163–189.

Arthur C Graesser, Murray Singer, and Tom Trabasso. 1994. Constructing inferences during narrative text comprehension. *Psychological review*, 101(3):371.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

Mary Hare, Michael Jones, Caroline Thomson, Sarah Kelly, and Ken McRae. 2009. Activating event knowledge. *Cognition*, 111(2):151–167.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6384–6392.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. 2022. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

JM Keenan and W Kintsch. 1974. The identification of explicitly and implicitly presented information. *The representation of meaning in memory*, pages 153–176.

Najoung Kim and Sebastian Schuster. 2023. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3835–3855, Toronto, Canada. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Marta Kutas and Steven A Hillyard. 1989. An electrophysiological probe of incidental semantic association. *Journal of cognitive neuroscience*, 1(1):38–49.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. 2022.

Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. 2022. FNet: Mixing tokens with Fourier transforms. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4296–4313, Seattle, United States. Association for Computational Linguistics.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Eric Nyberg, and Alessandro Oltramari. 2022. Coalescing global and local information for procedural text understanding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1534–1545, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

George Michalopoulos, Michal Malyska, Nicola Sahar, Alexander Wong, and Helen Chen. 2022. ICDBigBird: A contextual embedding model for ICD code classification. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 330–336, Dublin, Ireland. Association for Computational Linguistics.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas A. Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David P. Schnurr, Felipe Petroski Such, Kenny Sai-Kin Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. 2022. Text and code embeddings by contrastive pre-training. *ArXiv*, abs/2201.10005.

Ha Thanh Nguyen, Randy Goebel, Francesca Toni, Kostas Stathis, and Ken Satoh. 2023. A negation detection assessment of gpts: analysis with the xnot360 dataset. *arXiv preprint arXiv:2306.16638*.

OpenAI. 2022. Introducing chatgpt. *OpenAI Blog*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. *OpenAI Blog*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Gabriel A Radvansky, Andrea K Tamplin, Joseph Armendarez, and Alexis N Thompson. 2014. Different kinds of causality in event cognition. *Discourse Processes*, 51(7):601–618.

Sahithya Ravi, Chris Tanner, Raymond Ng, and Vered Shwartz. 2023. What happens before and after: Multi-event commonsense in event coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1708–1724, Dubrovnik, Croatia. Association for Computational Linguistics.

Kyle Richardson, Ronen Tamari, Oren Sultan, Dafna Shahaf, Reut Tsarfaty, and Ashish Sabharwal. 2022. Breakpoint transformers for modeling and tracking intermediate beliefs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9703–9719, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Keisuke Sakaguchi, Chandra Bhagavatula, Ronan Le Bras, Niket Tandon, Peter Clark, and Yejin Choi. 2021. proScript: Partially ordered scripts generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2138–2149, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Abhilasha Sancheti and Rachel Rudinger. 2022. What do large language models learn about scripts? In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 1–11, Seattle, Washington. Association for Computational Linguistics.

Roger C Schank. 1975. The structure of episodes in memory. In *Representation and understanding*, pages 237–272. Elsevier.

Murray Singer and Michael Halldorson. 1996. Constructing and validating motive bridging inferences. *Cognitive Psychology*, 30(1):1–38.

Murray Singer and Kathryn FM Ritchot. 1996. The role of working memory capacity and knowledge access in text inference processing. *Memory & cognition*, 24(6):733–743.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Shane Storks, Qiaozi Gao, Yichi Zhang, and Joyce Chai. 2021. Tiered reasoning for intuitive physics: Toward verifiable commonsense language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4902–4918, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Tom Trabasso and Linda L Sperry. 1985. Causal relatedness and importance of story events. *Journal of Memory and language*, 24(5):595–611.

Paul Van den Broek. 1990. The causal inference maker: Towards a process model of inference generation in text comprehension. *Comprehension processes in reading*, pages 423–445.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Lilian D. A. Wanzare, Alessandra Zarcone, Stefan Thater, and Manfred Pinkal. 2016. A crowdsourced database of event sequence descriptions for the acquisition of high-quality script knowledge. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3494–3501, Portorož, Slovenia. European Language Resources Association (ELRA).

Noah Weber, Rachel Rudinger, and Benjamin Van Durme. 2020. Causal inference of script knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7583–7596, Online. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. 2021. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14138–14148.

Siyu Yuan, Jiangjie Chen, Ziquan Fu, Xuyang Ge, Soham Shah, Charles Robert Jankowski, Deqing Yang, and Yanghua Xiao. 2023. Distilling script knowledge from large language models for constrained language planning. *arXiv preprint arXiv:2305.05252*.

Fangzhou Zhai, Vera Demberg, and Alexander Koller. 2022. Zero-shot script parsing. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4049–4060, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Fangzhou Zhai, Iza Škrjanec, and Alexander Koller. 2021. Script parsing with hierarchical sequence modelling. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 195–201, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Li Zhang, Hainiu Xu, Yue Yang, Shuyan Zhou, Weiqiu You, Manni Arora, and Chris Callison-Burch. 2023. Causal reasoning of entities and events in procedural texts. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 415–431, Dubrovnik, Croatia. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

A Analysis of Human Beliefs about events A and B

In addition to measuring the reading times that reflect online processing, we also collected the answers to the questions about occurrences of events A and B that were presented after each story ("*How sure are you that event A/B happened?* – see Figure 1, part III").

The motivation for this was to gain insights into a) how exactly subjects accommodate a causal conflict (the $\neg A \rightarrow B$ condition) and b) whether subjects indeed infer event A when it is omitted from the story (the $nil \rightarrow B$ condition). The $A \rightarrow B$ condition serves as a baseline. We analyse the collected ratings using ordinal regression models (Christensen, 2018).

	$A \rightarrow B$	$nil \rightarrow B$	$\neg A \rightarrow B$
Event A	6.41 (1.45)	4.85 (2.89)	3.67 (3.19)
Event B	6.13 (1.84)	4.91 (2.80)	3.79 (3.13)

Table 4: Mean subjects' belief ratings (and SD in parentheses) that the event actually happened in the story, by event type (A or B) and story condition ($A \rightarrow B$, $nil \rightarrow B$, and $\neg A \rightarrow B$).

In the $A \rightarrow B$ condition, both events A and B were given on average high ratings (6.41 and 6.13, respectively – see Table 4), meaning that subjects were sure that the events happened when they both were explicitly mentioned in the story. Further, for both events, the ratings in the $\neg A \rightarrow B$ (**event A**: $b = -2.03$, $se = 0.24$, $z = -8.67$, $p < .001$; **event B**: $b = -1.6$, $se = 0.2$, $z = -8.22$, $p < .001$) and $nil \rightarrow B$ (**event A**: $b = -1.46$, $se = 0.22$, $z = -6.6$, $p < .001$; **event B**: $b = -0.99$, $se = 0.2$, $z = -4.97$, $p < .001$) were significantly lower compared to the $A \rightarrow B$ condition.

The analysis of subjects' ratings showed that the causal conflict (the $\neg A \rightarrow B$ condition) resulted in lowered beliefs about both events A and B (3.67 and 3.79, respectively). One potential explanation for this is that subjects might have used different strategies to resolve the conflict. For example, some subjects could assume that event B in fact did not happen, (however, contrary to the narrative) because the premise is not met. While others could resolve the conflict by assuming that event A in fact happened thus making event B also possible to happen. Both strategies would explain relatively lower strength of beliefs about both events B and A to happen. Any explanations, however, necessitate a follow-up study with more elaborative questions that potentially require subjects to provide explanations of the given ratings.

Interestingly, we also observe lower ratings for both events in the $nil \rightarrow B$ condition, compared to the $A \rightarrow B$ condition, which is contrary to our expectations. In the $nil \rightarrow B$ condition, event B was overtly mentioned in the story, which should lead to comparable strength in subjects' beliefs with the $A \rightarrow B$ condition. Subsequently, event A, even though not mentioned explicitly, should be inferred on the basis of the causal link between them and script knowledge: if she added star-shaped sprinkles (event B), then she should have prepared cake decorations beforehand (event A) – see Figure 1, part II.

A probable rationale for the discrepancy between our expectation and the actual ratings is that, when faced with the questions, subjects may have retrospectively re-evaluated the story, relying more on their memory representations. Compared to condition $A \rightarrow B$, event B might have been perceptually less salient in the $nil \rightarrow B$ condition. Event B is easy to integrate due to its relation to the corresponding script (which we observe in the reading time analysis – see Section 3.5, RQ2) and may not receive a lot of attention from the reader, hence reducing its memorization and subsequent retrieval of event B. In the $A \rightarrow B$ condition, on the other hand, attention to event B is strengthened by the causal link coming from an explicitly mentioned event A that might facilitate its retrieval from memory at the question answering stage (see Bower et al., 1979, for similar results in reading everyday stories where subjects were asked to evaluate which events were mentioned in the text).

Model Name	# para.	b	t	sign	b	t	sign	b	t	sign
	(M)		CSK			CSK (short dist)			TRIP	
Bigbird-roberta-large	412	0.18	1.64	n.s.	0.33	1.72	n.s.	0.04	2.90	**
BERT: large-uncased	366	0.30	2.14	*	0.21	1.43	n.s.	0.07	1.67	.
ALBERT-xxlarge-v2	210	0.20	1.78	.	0.47	3.50	**	0.09	5.25	***
Perceiver	201	-0.02	-0.51	n.s.	0.04	0.79	n.s.	0.01	1.29	n.s.
Bigbird-roberta-base	167	0.05	0.34	n.s.	-0.03	-0.13	n.s.	0.03	2.62	**
BERT: base-uncased	133	0.14	1.71	n.s.	0.21	2.00	.	-0.00	-0.00	n.s.
Nystromformer-512	132	0.06	1.50	n.s.	0.04	0.80	n.s.	-0.01	-0.46	n.s.
ConvBERT: base	130	0.01	1.66	n.s.	-0.00	-0.72	n.s.	-0.00	-0.33	n.s.
FNet-base	108	0.01	0.14	n.s.	0.02	0.41	n.s.	-0.01	-0.80	n.s.
DistilBERT: base-uncased	90	0.12	2.08	.	0.16	2.43	*	-0.00	-0.01	n.s.
Electra-large-generator	83	0.12	1.15	n.s.	0.01	0.13	n.s.	-0.01	-0.16	n.s.
SqueezeBERT: uncased	75	0.13	1.63	n.s.	0.21	2.40	*	-0.04	-1.09	n.s.
Electra-base-generator	57	0.12	2.69	*	0.08	1.25	n.s.	-0.04	-1.15	n.s.
Electra-small-generator	17	0.18	2.66	*	0.09	1.33	n.s.	-0.03	-0.82	n.s.
ALBERT-base-v2	15	0.27	2.90	**	0.16	2.25	*	0.01	0.49	n.s.

Table 5: Results for MLMs on RQ1 ($A \rightarrow B$ versus $\neg A \rightarrow B$) on CSK (original and intervention removal) and TRIP dataset. The # para. (M) column shows the number of parameters in millions. n.s. represent that the results are not statistically significant. The ., *, **, and *** in the sign column represent p-values < 0.1, 0.05, 0.01, and 0.001.

B Details of LLMs

We use one Nvidia A100 GPU card to run all of our experiments. Thanks to our zero-shot setting, the experiment of each model takes less than 10 minutes.

B.1 GPT models

GPT-2. GPT-2 (Radford et al., 2019) is one of the most influential language models by OpenAI. As a decoder-only causal PLM, GPT-2 is often used as a baseline.

GPT-3 models. GPT-3 (Brown et al., 2020) is the upgraded version of GPT-2 which uses almost the same model and architecture but with a significantly larger amount of parameters, which was ten times more than any previous non-sparse language model. GPT-3 and GPT-3.5 were chosen to be evaluated as they were expected to perform the best, based on their strong performance on a range of NLP tasks. We experiment with different versions of GPT-3 and GPT-3.5.[8] **GPT-3 models** (Brown et al., 2020): `curie` is a GPT-3 with 6B parameters. `davinci` is a GPT-3 with 175B parameters. **InstructGPT models** (Ouyang et al., 2022): `davinci-instruct-beta` is a model trained with supervised fine-tuning on human demonstrations; `text-davinci-001` and `text-curie-001` further includes top-rated model samples from quality assessment by human labellers. **GPT 3.5 models** (Neelakantan et al., 2022): `text-davinci-002` is an InstructGPT model based on a model trained with a blend of code and text; `text-davinci-003` was further trained using reinforcement learning with human feedback.

Newer models from OpenAI like GPT-4: `gpt-4-turbo`, `gpt-4` or GPT-3.5: `gpt-3.5-turbo` don't support the "Completions" API and can't return probabilities given input tokens so we don't include them (OpenAI, 2023).

B.2 Chatbots

As the two current state-of-the-art LLMs, GPT-4 and ChatGPT, are both designed to function as chatbots, our aim is to harness the potential of the most capable open-source chatbot available to us. Chatbots, by design, need to comprehend and respond contextually to inputs, often requiring them to make connections between disparate pieces of information in a conversation. **Vicuna** is an open-source chatbot created by fine-tuning an LLaMA base model with approximately 70K user-shared conversations collected from ShareGPT.com. Preliminary evaluation in their paper (Chiang et al., 2023) suggests that Vicuna reaches 90% of the quality of chatbots such as ChatGPT and Google's Bard.

[8] More details are on https://platform.openai.com/docs/model-index-for-researchers

B.3 Efficient Models

There are models that need less memory or less time. Methods that reduce space could have a better performance here, because, for most of this experiment, we had limited space. Efficient models are interesting for long-range dependencies because they employ innovative techniques or optimizations to handle dependencies more effectively. Efficient models might be better or worse at capturing the relationships between distant parts of the text due to their unique approaches.

Nyströmformer and language perceiver are examples of models with efficient self-attention.

C Additional Experiment Results

C.1 Masked Language Models (MLMs)

MLMs are another group of language models that obtained state-of-the-art performances across many NLP tasks. We note that the way they work is not similar to human language processing, and the surprisal estimates obtained from them are not directly comparable to surprisals obtained from left-to-right models. However, we decided to include some MLMs that have been specifically designed to handle long-distance dependencies (via their efficient self-attention mechanisms) into our evaluation, to observe how these models perform regarding the causal inferences given long commonsense stories. We first picked a set of models from the BERT family including BERT (Devlin et al., 2019) and Bigbird-roberta (Michalopoulos et al., 2022) as representatives for MLMs because they used to be the state-of-the-art in many NLP benchmarks concerning commonsense inference (Wang et al., 2018, 2019). We opted to incorporate models that use efficient self-attention mechanisms like We also test FNet (Lee-Thorp et al., 2022), Nystromformer (Xiong et al., 2021) and Perceiver (Jaegle et al., 2022).

We follow Salazar et al. (2020) to provide models with the context before and after the target token in segment B. The pertinent token itself is masked, forcing the masked language models to infer it based on the surrounding context. For instance, in the example story in Figure 1, the words "added star-shaped sprinkles" constitute the target region describing event B. Each token in this sequence was masked one at a time. We then calculated the probabilities of the masked tokens given the surrounding story context. MLM models thus have more information than CLM models due to the additional information from other tokens in the event B and the context after event B. We therefore would like to point out that this method is not cognitively plausible, and that the surprisal scores obtained from them hence will also reflect this "privileged" knowledge. We also note that the surprisal estimation from MLMs can in principle be adapted to simulate left-to-right processing better, but think that this is only worthwhile to explore in more detail if MLMs prove to be successful at modelling the long-distance dependencies relevant to our texts.

Our results in Table 5 show that only some MLM models showed a significant difference in surprisal estimates between the coherent and the incoherent ($\neg A \rightarrow B$) condition on either CSK or TRIP datasets. Since their behaviors are not consistent across these two datasets, we consider all MLMs fail to distinguish between coherent and incoherent conditions.

C.2 Effect of dependency length (distance between events A and B)

Next, we wanted to check whether the failure of the models that don't show a significant difference between conditions is due to problems with encoding the text effectively and "remembering" event A or $\neg A$ when processing event B, or whether it is related to failure to detect the mismatch between the events. We therefore modified the original experiment's design by reducing the distance between events A and B in the story by removing all intervening sentences. (Note that we did not ensure that the removed sentences did not contain crucial information that would compromise the coherence of the story.)

If model failure on the previous task is due to difficulty in handling a long intervening context, we expect that models would show a significant difference between surprisal estimates in this short-distance condition.

As shown in Table 6 column named "CSK (short dist)", we find that most models show the same behavior in the short-distance condition and the long-distance condition. Interestingly, the results of both GPT-3.5 and Vicuna are non-significant in this condition. This could be due to the removal of intermediary materials, thereby potentially interrupting the causal chains and adversely affecting the activation of event B. Other models that are still not showing a significant difference between surprisal estimates in the different conditions might

Model Name	# para.	b	t	sign	b	t	sign
	(M)		CSK			CSK (short dist)	
GPT-3.5: text-davinci-003	175K	0.59	5.87	***	0.20	1.59	n.s.
GPT-3.5: text-davinci-002	175K	0.51	2.75	*	0.10	0.70	n.s.
InstructGPT: text-davinci-001	175K	0.26	2.03	.	-0.02	-0.18	n.s.
InstructGPT: davinci-instruct-beta	175K	0.21	2.76	*	0.12	1.78	.
GPT-3: davinci	175K	0.21	2.76	*	0.19	2.69	*
Vicuna-13B	13016	0.22	2.25	*	-0.01	-0.07	n.s.
Vicuna-7B	6738	0.28	2.56	*	0.12	1.08	n.s.
InstructGPT: text-curie-001	6700	0.03	0.31	n.s.			
GPT-3: curie	6700	0.23	3.43	**	0.21	3.75	**
GPT-2: XL	1638	0.05	0.96	n.s.	0.08	1.54	n.s.
GPT-2: L	838	0.04	0.77	n.s.	0.04	0.64	n.s.
XGLM	827	-0.03	-0.79	n.s.	0.02	0.38	n.s.
Bigbird-pegasus-large-arxiv	470	0.06	1.20	n.s.	0.00	-0.04	n.s.
Pegasus-large	467	0.02	0.85	n.s.	0.00	0.00	n.s.
XLNet-large-cased	393	-0.03	-1.99	.	-0.04	-2.42	*
OPT	357	0.01	0.12	n.s.	0.02	0.32	n.s.
GPT-Neo	164	0.03	0.67	n.s.	0.05	1.11	n.s.
GPT-2	163	0.00	-0.10	n.s.	0.03	0.74	n.s.
GPT: openai-gpt	148	0.00	-0.01	n.s.	0.06	1.35	n.s.

Table 6: Results of CLMs with shorten context on RQ1 ($A \rightarrow B$ versus $\neg A \rightarrow B$) on CSK (original and intervention removal) and TRIP dataset. The # para. (M) column shows the number of parameters in millions. n.s. represent that the results are not statistically significant. The ., *, **, and *** in the sign column represent p-values < 0.1, 0.05, 0.01, and 0.001.

be failing due to not recognizing the semantic inconsistency between $\neg A$ and B.

Each of our narratives represents a sequence of events that the main character is involved in step by step in order to achieve their goal (e.g., to bake a cake or to take a flight). For example, for taking a flight story, the events are:

Reach the airport, get the boarding pass, ***[EVENT A] check in the luggage***, *go through the security, wait at the gate, board the plane, find one's seat, fasten the seatbelt, turn off the electronic devices, wait on the plane, land, leave the plane,* ***[EVENT B] pick the bags at the baggage claim***, *leave the airport*

In turn, removing the context between A and B typically results in very low story coherence, see the following example:

After several months away from home, Julia was finally able to visit her family for a few days. However she had a long way to go, so she decided to travel by air. First, she went to the main airport on a public bus. Once at the airport, she got her boarding pass and ***[EVENT A] checked in her luggage***. *< ... > Afterwards, she* ***[EVENT B] picked up her bags at the baggage claim and left the airport.*** *Finally, she arrived home and met her family. It had been so long!*

This expectedly leads to higher surprisal in all conditions. However, we reasoned that conditions $A \rightarrow B$ and $\neg A \rightarrow B$ are affected by this change to a similar extent, and hence a difference in surprisal (which would reflect the stronger logical clash between $\neg A$ and B) would be reflected in lower surprisal values in this condition compared to $A \rightarrow B$. The strong drop in plausibility might however be a reason for the difference between $A \rightarrow B$ and $\neg A \rightarrow B$ lacking significance.