Large language models fail to derive atypicality inferences in a human-like manner

Charlotte Kurch, Margarita Ryzhova and Vera Demberg
Dept. of Language Science and Technology, Saarland University, Germany
chku00001@stud.uni-saarland.de, mryzhova@lst.uni-saarland.de,
vera@lst.uni-saarland.de

Abstract

Recent studies have claimed that large language models (LLMs) are capable of drawing pragmatic inferences (Qiu et al., 2023; Hu et al., 2022; Barattieri di San Pietro et al., 2023). The present paper sets out to test LLM's abilities on atypicality inferences, a type of pragmatic inference that is triggered through informational redundancy. We test several state-of-the-art LLMs in a zero-shot setting and find that LLMs fail to systematically fail to derive atypicality inferences. Our robustness analysis indicates that when inferences are seemingly derived in a few-shot settings, these results can be attributed to shallow pattern matching and not pragmatic inferencing. We also analyse the performance of the LLMs at the different derivation steps required for drawing atypicality inferences – our results show that models have access to script knowledge and can use it to identify redundancies and accommodate the atypicality inference. The failure instead seems to stem from not reacting to the subtle maxim of quantity violations introduced by the informationally redundant utterances.

Keywords: pragmatics; informational redundancy; human-like reasoning; large language models

1 Introduction

Recent studies have shown that large language models (LLMs) can oftentimes provide responses that are consistent with pragmatic interpretations, e.g., Qiu et al. (2023). An analysis of seven different pragmatic phenomena (including humor, coherence and irony) by Hu et al. (2022) found that LLMs to exhibit similar accuracy and error patterns as humans; and research has also reported LLMs performing well on test developed to test the pragmatic ability of humans (Barattieri di San Pietro et al., 2023).

In the present paper, we test whether LLMs are capable of deriving atypicality inferences – the type of pragmatic inferences that arise in the face of mentioning information that is *informationally redundant* (IR). The informational redundancy arises from the fact that the information can be inferred from shared knowledge about typical event sequences (*script knowledge* – knowledge about everyday situations, like dining at a restaurant or shopping; see, Bower et al., 1979). Mentioning easily inferable events violates the quantity maxim which holds that speakers should be informative (Grice, 1975).

For example, eating is the activity that is highly predictable in a restaurant scenario. Thus, the utterance in (1) is informationally redundant:

(1) *Mary went to a restaurant.* ***She ate there!***

Mentioning the inferable event leads to pragmatic inferences – Kravtchenko and Demberg (2022) showed that subjects lower their beliefs about the highly conventionally habitual activity (e.g., eating). The derivation mechanism assumes that when faced with utterances that are informationally redundant, comprehenders try to 'repair' the utterance informativity by inferring that the mentioned event is atypical for the referent. With relation to (1), it follows that Mary does not usually eat, when going to a restaurant.

The derivation mechanism of atypicality inferences can be summarized in four steps (Ryzhova et al., 2023; Kravtchenko and Demberg, 2022). At first, comprehenders identify redundancy in the message based on script knowledge. Secondly, they realize that redundancy is infelicitous due to violation of the quantity maxim. Thirdly, they infer atypicality (Mary does not usually eat at a restaurant). Finally, they need to accommodate atypicality with their world knowledge (e.g., Mary usually only orders drinks). This decomposition into steps allows us to check what aspect of the pragmatic inference might be particularly challenging for the LLM.

Previous work on the recent generative models suggests that they have a promising understanding of script knowledge – see Huang et al. (2022), where GPT-3 generated plausible script schemata. However, it is not only relevant whether the script knowledge is learned by the model, but also whether the model is able to access it and integrate it into the task solving process. Hong et al. (2024b) tested more than 30 different LLMs on implicit vs. explicit causal relations between two script events. The models, unlike humans, were unable to infer or predict a cause/event from script knowledge, if it was omitted. This might imply either insufficient representation of script knowledge or inability to integrate it.

Recent research on LLMs has explored their ability to understand non-literal language, demonstrating that these models can emulate human-like performance in deriving pragmatic meaning (Hu et al., 2022). For example, Qiu et al. (2023) showed that ChatGPT, to some extent, resembles human behaviour — it consistently derives scalar implicatures by interpreting the quantifier 'some' and disjunctions pragmatically. However, the model exhibited a lack of human-like flexibility when nuanced interpretation required consideration of contextual information.

In the present paper, we investigate pragmatic abilities in the derivation of atypicality inferences of three recent generative models that offered the most promising performance, namely – GPT-3.5-turbo (GPT-3.5-t; t = 1, presence_penalty = 0, top_p = 1), GPT-4 (t = 1, presence_penalty = 0, top_p = 1) and the open-source model LLama 3 8B Instruct (Llama 3; t=0.6, repeat_penalty = 1.2, top_p = 0.9). We present a series of experiments in which we firstly follow a zero-shot approach to replicate the results of Kravtchenko and Demberg (2022) and Ryzhova et al. (2023) with LLMs (Exp. 1). Next, we follow a few-shot prompting approach that has been shown to improve the models' reasoning (Exp. 2) and perform a perturbation analysis with modified few-shot exemplars. Finally, in Exp. 3, we analyse the LLM's ability to perform the different reasoning steps required for atypicality inferences according to Kravtchenko and Demberg (2022) and Ryzhova et al. (2023).

2 Atypicality inferences

We here briefly present the original experiment of Kravtchenko and Demberg (2022) and discuss the derivation steps for atypicality inferences.

The mechanism of atypicality inferences lies in the violation of the quantity maxim where interlocutors are expected to convey the right amount of information to their conversational partners – neither more nor less (Grice, 1975).

Informativity of a message, among other things, is dependent on the mutual knowledge and beliefs of interlocutors about each other. According to the previous literature, humans exhibit a remarkable ability to infer script events, even those left unmentioned in everyday narratives, without causing the discourse to appear odd or inconsistent. This capability is reflected in human communication, too, where individuals don't explicitly mention all script-related events, and yet listeners can seamlessly infer this information from their script knowledge (Bower et al., 1979). Kravtchenko and Demberg (2022) investigated the comprehension of utterances that are overinformative or informationally redundant (IR), and thus violate the maxim of quantity, given comprehender's script knowledge. They examined 24 stories describing common everyday event sequences, such as going to a restaurant or going shopping. In these scenarios, script knowledge consists of specific sequences of events, such as (for a going to a restaurant scenario) reaching the restaurant, taking a table, looking at the menu, ordering food, **eating**, paying, and leaving the place (Bower et al., 1979; Wanzare et al., 2016).

Each story underwent a 2 (ordinary vs. wonky common ground context) x 2 (conventionally habitual vs. non-habitual utterance) manipulation (see an example of an item in all conditions in Table 1). Critically, the conventionally (conv.) habitual utterance "She ate there!" was an event taken from the script schema.

Kravtchenko and Demberg (2022) manipulated the presence of conv. habitual utterance in the story. After reading a story, subjects were asked to express their beliefs about the target activity on a scale ranging from 0 to 100: *How often do you think Mary usually eats, when going to a restaurant?* (Never-Sometimes-Always). Overall, when the context followed script-schema (ordinary condition), subjects assigned high typicality ratings in the baseline condition (where no utterance was present in the story), meaning that subjects believed that Mary usually eats in restaurants, in accordance with script knowledge. However, when the conv. habitual utterance was present in the story, the subjects' ratings about Mary typically eating

when going to a restaurant were significantly lower (baseline: 85.79 vs. habitual utterance: 72.37; $p < .001$) – see also Figure 1.

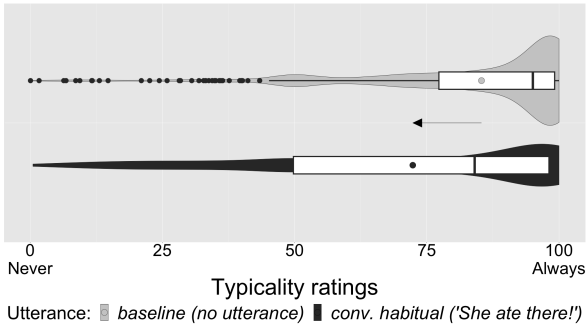

Figure 1: Human ratings of event typicality (e.g., eating when going to a restaurant) taken from Kravtchenko and Demberg (2022). Violin plots, overlaid with box plots, show the distribution of ratings. Circles represent mean values. The arrow indicates a statistically significant difference in ratings between conditions.

This effect crucially depends on informational redundancy – it disappeared (baseline: 48 vs. non-habitual utterance: 45.71) in the context, where the conv. habitual utterance was not informationally redundant (see Table 1, wonky context, where Mary was portrayed as a non-eater). The effect is also not present when the target utterance was not referring to a predictable event "She got to see their kitchen!", see Table 1, non-habitual utterance (ratings for ordinary: 40.80 to 42.47; for wonky: 38.49 to 39.56 – baseline to non-habitual utterance condition, respectively).

2.1 Derivation steps of atypicality inferences

To investigate how subjects accommodated atypicality inferences in the situational context of a story and to better understand the underlying derivation processes of atypicality inferences, Ryzhova et al. (2023) conducted a follow-up study, in which they asked participants to explain a given rating. The ratings were tagged according to whether they provided evidence for an atypicality inference having been drawn. The most important categories from their annotation scheme are shown in Table 2.[1]

In most cases, subjects derived atypicality inference (*atypicality* tag). These responses reflected recognition of informational redundancy and stated the utterance as the reason to assume that Mary does not usually eat in restaurants — this corresponded to low typicality ratings (see mean ratings in column 2 of Table 2). Interestingly, subjects oftentimes effectively augmented the common ground to make the IR utterance informative with respect to the context. In doing so, they provided justification of **why** Mary does not usually eat ("...because she interviews people there"). Sometimes, however, even when subjects arrived at atypicality inference, their answer justified that they did not accept the drawn inference (*atypicality_reject*) – this corresponded to high ratings.

When subjects did not derive atypicality inference, their explanations included various formulations of stating what would be a typical human behaviour (*no_atypicality*). Such answers were associated with high typicality ratings, and comprised a second biggest annotation category.

Results of Ryzhova et al. (2023) thus confirm that informationally redundant utterances lead subjects to infer atypical behaviour, and that they go through an accommodation process: in order to obtain a consistent picture, they come up with a circumstance leading to the activity being worth mentioning (e.g., ordering only drinks or being short of money). These results provide a basis for comparison to reasoning of LLMs.

3 Exp. 1: Zero-Shot Prompting for Eliciting Atypicality Inferences

Our first experiment set out to test the ability of recent LLMs to derive atypicality inferences under conditions similar to the human participants. We used the same 24 stimuli and tested how models rated the typicality of conv. habitual and the non-habitual activity in all conditions used by Kravtchenko and Demberg (2022) (see Table 1). Models were prompted for providing both a typicality rating on a scale from 0% to 100% [2] and a justification for their rating.

We report here the results for conv. habitual activity in the ordinary context – for the wonky context and non-habitual activity the models behaved similarly to humans (for results see appendix B).

Methods The prompt we used underwent iterative prompt engineering to assure consistently sensible and usable output. It includes instructions to use common sense reasoning and speculate based

[1] Ryzhova et al. (2023) report a substantial inter-annotator agreement (Cohen's κ = 0.74 (p<.0001), 95% CI (0.7, 0.77)).

[2] As previous research has shown that LLMs struggle with tasks involving numbers (Schwartz et al., 2024; Hong et al., 2024a), we have also performed the same experiment using a 7-point Likert scale, and applying the self-calibration method proposed by Tian et al. (2023). These experiments yielded very similar results that can be found in appendix C.

Table 1: An example of a "restaurant" story by context (ordinary vs. wonky) and utterance condition (conv. habitual vs. non-habitual activity is mentioned in the utterance). A baseline for both context conditions does not include an utterance block.

Context	**ordinary** Mary is a journalist **who often goes to restaurants after her interviews**.	**wonky** Mary is a journalist **who often interviews restaurant waiters, but doesn't like eating out**.
	Yesterday, she went to a popular Chinese place. As she was leaving, she ran into her friend David, and they started talking about the restaurant. After they parted, David continued on his way when he suddenly ran into Sally, a mutual friend of him and Mary.	
Utterance	**conventionally habitual activity** David said to Sally: "I ran into Mary leaving that Chinese place. **She ate there!**"	**non-habitual activity** David said to Sally: "I ran into Mary leaving that Chinese place. **She got to see their kitchen!**"
Q habitual	How often do you think Mary usually eats, when going to a restaurant?	
Q non-habit	How often do you think Mary usually gets to see the kitchen, when going to a restaurant?	

Table 2: Annotation scheme from Ryzhova et al. (2023) with examples from human explanations for the restaurant script.

annotation tag (proportion of tag in data)	**inference drawn? (mean rating)**	**example of an answer**
atypicality (45.6%)	yes (51.84)	Since David mentioned it, it sounds like she doesn't always eat at restaurants. Maybe she sometimes interviews people in restaurants.
atypicality _reject (6.13%)	unclear (95.46)	After interviews Mary will be tired so she probably eats. She can't just go to a restaurant for a drink after a long day.
no_atypicality (39.46%)	no (93.82)	Usually when you go to a restaurant, it is to eat.
other (8.81%)	unclear (69.33)	He didn't tell Sally which restaurant, he said that restaurant, as though they go there often.

on its knowledge of human behavior to circumvent responses related to an inability to perform the task[3]. The ratings in the different conditions were compared using a paired t-test.

Annotation scheme For evaluating the model reasoning in the habitual utterance condition, we extended the annotation scheme used in Ryzhova et al. (2023) to cover types of answers that were typical in LLMs, but had not been observed in humans. We added the label *reinforced_utterance* as a subtype of *no_atypicality* for explanations where the redundant utterance was considered a reinforcement of the typicality, and the label *hallucination/bad_reasoning* to capture erroneous and nonsensical model generated explanations, see Table 3 for an example[4].

[3]See appendix A for details on the prompts.

[4]We annotated a subset of answers (GPT-4, few-shot) with two annotators and found a substantial inter-annotator agreement (Cohen's κ = 0.73 (p<.0001), 95% CI (0.52, 0.93))

Table 3: Extended annotation scheme for LLMs with examples from the restaurant and the haircut scripts.

annotation tag	**inference drawn?**	**example of an answer**
no_atypicality: reinforced utterance	no	The statement "Mary ate there!" suggests that it is a usual occurrence for Mary to eat when she goes to a restaurant after her interviews.
hallucination/ bad_reasoning	unclear	100% because the context states that she usually cuts her hair herself using scissors.

Results In contrast to humans, we found no significant typicality rating changes between the baseline and the habitual utterance condition across the models (see Figure 2). There was non-significant change for GPT-3.5-t (94.40→97.04) and Llama 3 (87.8→94.5) in the opposite direction, i.e. activities are judged to be more frequent, when the utterance is seen. Overall, the models assigned very high typicality ratings to all stimuli, irrespective of condition. Occasionally the models deemed it impossible to answer and gave 50%ratings.

Models' explanations were in accordance with the high ratings – see Table 4. The majority of responses were classified as *no_atypicality*, and especially *reinforced_utterance*, where the models reinforced high typicality based on the utterance. Only a very small number of responses were classified as *atypicality*, but these were still associated with high ratings. Finally, some responses also contained hallucinated facts or incorrect or confused reasoning.

For a sanity check, we also looked into the typicality ratings of the habitual activity when the non-habitual utterance was present in the story. Simi-

larly to human results, the ratings in this condition were high and not significantly different from the baseline for all three models. It shows that presence of the non-habit. utterance does not affect the interpretation of the habitual event typicality. In other words, the fact that Mary got to see the kitchen does not influence the typicality of her eating in the restaurant.

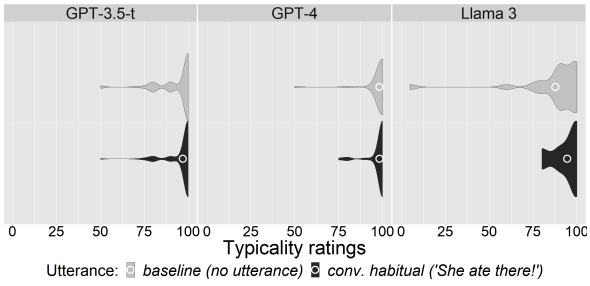

Figure 2: Zero-shot, habitual activity analysis in the ordinary context. Boxplots are omitted, due to high skew in the data.

Discussion In the zero-shot setup, where the models were put in the same settings as humans, we observed no atypicality inferences, contrary to human results. Those few explanations that showed derivation were not associated with lower ratings. So what might cause the observed discrepancy between LLMs and humans?

As the first step of deriving inferences requires identifying the redundancy based on script knowledge, an obvious first consideration is whether models have the relevant script knowledge. However, in the baseline condition (no activity mentioned) typicality ratings are high and the explanations refer to script knowledge. This conclusion is further supported by reduced typicality ratings that were obtained in a wonky context's baseline that we present in appendix B. In this context, the script knowledge is overwritten by stating atypical behaviour, and all models captured this changing lowering their beliefs accordingly.

At that same first step, it is also possible that models may fail to recognize that the observed utterance is informationally redundant. Further, the second step requires assessing that the redundancy violates the conversational norms. A failure to do either of these would be an explanation consistent with the fact that model justifications for high typicality ratings referred to event typicality (*reinforced_utterance*), a type of reasoning that was typically not found in human justifications.

Experiments 2 and 3 below aim to investigate what aspect of the reasoning the models have most difficulty with.

4 Exp. 2: Few-Shot Prompting

Few-Shot prompting (Brown et al., 2020) is a popular technique in which the prompt is enriched with a small number of examples that demonstrate how to do the target task correctly. This has often been found to improve model performance on other NLP tasks (Schick and Schütze, 2020; Zhao et al., 2021).

We selected a total of 4 of the stimuli as exemplars: specifically, the stimuli with conv. activities that were, respectively, rated most and least habitual by the human participants. For each stimulus, responses that follow the output template while mimicking human behavior in the conv. habitual utterance condition were crafted, i.e., the responses showing a lower rating and providing a justification that alluded to an atypicality inference being drawn. The models were prompted twice with two exemplars each (paired according to their ratings) and the instructions prompt was amended to reflect that two examples would be demonstrated.[5] We only collected responses for the conv. habitual activity (Q habitual in Table 1) in the ordinary condition, and present the combined results collapsing across exemplars, using the same analysis as in Exp. 1.

Results In the few-shot setting, we observed a significant difference in typicality ratings between the baseline and habitual utterance conditions for GPT-4 (mean 96.2 →84.1; t(23) = 5.82, $p < .0001$) and GPT-3.5-t (mean 96.5 →89.4; t(23) = 2.98, $p < .01$). For Llama 3 there is no change (mean 85.0 →81.2). The ratings being on average lower for GPT-4 and GPT-3.5-t when the habitual utterance was present is in line with the derivation of an atypicality inference – see Figure 3.

In contrast to Exp. 1, the presence of the non-conv. habitual utterance ("She got to see their kitchen!") did not have an effect on the ratings only for GPT-4 (mean: 96.2 →95.0). For GPT-3.5-t, however, there was a significant change (mean 96.5 →84.0; t(23) = 3.50, $p < .01$), meaning that the ratings were on average lower in the presence of any utterance (even the one not related to the activity mentioned in the question), indicating that the model does not actually derive atypicality inferences. Interestingly, we also see a significant

[5] See appendix A for exact prompt formulations.

Table 4: Proportionate distribution in % of the annotations for all responses in habitual utterance condition with ordinary context.

Annotation		Human	Zero-Shot Prompting			Few-Shot Prompting		
			GPT-3.5-t	GPT-4	Llama 3	GPT-3.5-t	GPT-4	Llama 3
atypicality		45.6	4.13	4.17	8.33	11.36	65.9	6.82
no-atypicality	normal	39.46	42.07	58.33	60.41	59.09	13.63	50.0
	reinforced utterance	-	48.62	41.67	29.16	45.45	2.27	40.91
unclear	atypicality_reject	6.13	0.0	2.08	0.0	4.55	18.18	2.27
	hallucination/ bad_reasoning	-	6.88	6.25	0.0	2.27	0.0	9.09
	other	8.81	1.8	0.0	4.16	0.0	0.0	0.0

rating change for Llama 3 (mean 85.0 →73.1; t(23) = 2.61, $p < .05$), further solidifying the model's failure at deriving atypicality inferences.

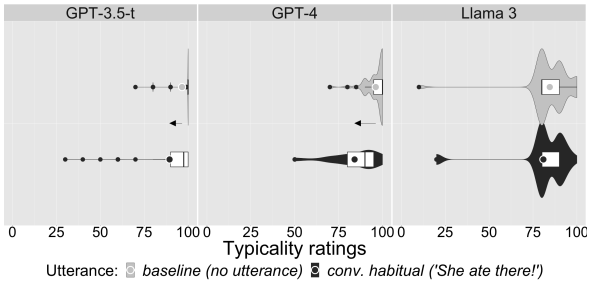

Figure 3: Few-shot, habitual activity analysis

Next, the number of explanations in favor of atypicality inference (*atypicality*) increased strongly in GPT-4, where *atypicality* is the most frequent annotation tag (there's a small increase for GPT-3.5-t and no for Llama 3, Table 4). We note though that the atypicality justifications were sometimes inconsistent with the numerical ratings given by the model: a very cautious explanation stating a slightly decreased typicality would co-occur with a large decrease in the typicality rating. For GPT-3.5-t and Llama 3 the majority of responses are again classified for exhibiting *no_ atypicality*.

Overall, the models now also show more responses that were classified as *atypicality_reject*, where the atypicality is brought up but dismissed in the justification.

Perturbation analysis In addition to the few-shot experiment above, we aimed to test the robustness of the inferencing ability of GPT-4 in the few-shot setting in order to determine whether the model shallowly copies over and adapts the provided exemplars, or whether it uses the exemplars to pick up on the task more deeply.[6]

[6]Results on the other models are provided in the appendix, as these models failed to show the correct behaviour in the basic few-shot setting.

Perturbation 1 Firstly, we prompted the models using the same items as exemplars, but this time, only one exemplar modeled the conv. habitual utterance condition, while the second one modeled the non-habitual utterance condition. This aimed at the models ability to differentiate between the utterances and apply only the relevant exemplar to the problem it was presented with. This manipulation meant that the results for GPT-4 became less clear: ratings in the conv. habitual utterance condition still vary significantly from the baseline (mean: 96.2 →91.9, t(23) = 2.47, $p < .05$), but the non-habitual utterance condition also varies significantly from the baseline (mean: 94.8, t(23) = 2.49, $p < .05$), and the two utterance conditions no longer vary significantly from each other. This decline in atypicality inferences is supported by the explanations, where we see *no-atypicality* for most stimuli (atypicality is only classified 8 times).

Perturbation 2 We crafted intentionally misleading and incongruent exemplars where 100% ratings paired with reasoning expressing atpicality. We tried two variations of the reasoning: (A) expresses atypicality due to the utterance implying a change from habitual behavior, (B) simply states atypicality without any reference to habitual behavior. Notably, GPT-4 matches the exemplars the majority of the time in setting B, where we do not introduce the concept of habituality due to script knowledge. In setting A, however, it replicates the exemplar less than half the time, and the remaining times rejects the atypicality or assigns no atypicality. For the latter it will frequently assign a different purpose to the utterance, explicitly stating that it does not imply atypicality.

Discussion While the results of the few-shot prompting experiment on GPT-4 seem very promising, we were wondering about whether these responses are given for the "right reasons" (i.e., whether the examples provided in the prompt clari-

fied the task to the model) or whether the model is adapting aspects of the answers given in the prompt in a shallow way, e.g., copying down a low rating and adapting the explanation to the new target.

Our first perturbation analysis showed that GPT-4 cannot consistently differentiate between redundant and non-redundant utterances, or apply the conversational norm leading to atypicality. With the second analysis we observed two behaviors: (1) matching both reasoning and rating to the exemplar even if they are incongruent, and (2) copying of the rating and adjusting the reasoning. While (1) mostly implies some degree of blind copying, the occurrence of (2) shows the model applying some level of reasoning or knowledge. Interestingly, this behavior is prevalent when the exemplars provide the script knowledge and resulting habituality, and how it is voided by the utterance, leading the model to explicitly disagree with this modeled reasoning. This leads us to hypothesize that model does not see a problem with redundancies and hence does not apply the conversational norm that leads to the derivation of atypicality inferences, even to the point of rejecting it.

In order to better understand the performance of GPT-4 and to obtain better insights on the performance of all models on the reasoning steps that were previously hypothesized to be part of human reasoning for this task, we tested the performance of all models on the component steps of atypicality reasoning in Exp. 3.

5 Exp 3: Analysing the steps of reasoning process

In Exp. 3, we decomposed the atypicality inference reasoning task into its sub-components as outlined in Kravtchenko and Demberg (2022) and Ryzhova et al. (2023): 1) identify the redundancy based on script knowledge; 2) realize that redundancy is infelicitous, as it violates conversational norms; 3) infer activity atypicality; 4) explicitly accommodate atypicality in situational context. Our goal was to clarify how well the models perform on each of these steps. The models were prompted with adjusted instructions, telling them that they were experts on human behavior and had the task of answering a question based on a provided context. As context, they were given each stimulus in the conv. habitual utterance condition, and then one question at a time.

Notably, this method of prompting the model with questions that are aimed specifically at performing each of the steps does not reliably show whether or not a given model is actually able to perform this step unprompted, or in a different context. We do however believe in the merits of assessing the models' abilities and behaviors in this controlled setting for providing initial insights into potential points of failure.

Experimental results on the variations of the prompts are presented in appendix E. Below, we only report on the question formulations that most successfully elicited what we were looking for across models.

Step 1: Identifying Redundancy For identifying the informational redundancy, we report the results of the following two prompts:

- Q1: Does the direct speech contain any redundancies?
- Q2: The direct speech contains redundant information. Can you identify the redundancy and elaborate why it is one?

For Q1, where the presence of redundancy is open-ended, GPT-4 and Llama 3 succeeded at explicitly identifying the informational redundancies (18 and 14 times, respectively), while GPT-3.5-t did not.

For Q2, where the presence of a redundancy was presupposed, GPT-4 identified it for all 24 stimuli and the performance of GPT-3.5-t was also generally improved: it correctly reported the redundancy in 13 stories. For Llama 3 there is no positive effect as it reported the expected redundancy 13 times for this prompt. Overall, we take this finding as evidence that the model successfully draws on script knowledge and can in principle identify the informational redundancy.

Step 2: Realizing that redundancy is infelicitous The drawing of an atypicality inference is an accommodation process in which the comprehender 'repairs' an utterance that otherwise may be viewed as infelicitous due to the redundancy. We consequently wondered whether the conversational norm under which redundancies should be avoided (Maxim of Quantity) is known and accessible to the model. However, this aspect proved to be very difficult to assess via prompting, due to its subtlety (explicit reasoning about them would also be hard to elicit from humans, as pragmatic implicatures can always be denied – see e.g., Garmendia, 2023).

When asking the model whether the utterance including informationally redundant information was

a good / acceptable way of communicating, GPT-3.5-t and GPT-4 tended to respond that redundancies could be a problem, but provided non-specific albeit reasonable examples of why redundancies can be ok. Llama 3's outputs most of the time said that redundancy was problematic and unacceptable, while exhibiting an improved ability for correctly identifying the informational redundancy.

Step 3: Inferring Atypicality Next, we tested whether the model can infer atypicality based on the mentioning of redundant information, using prompt Q3:

- Q3: The direct speech contains seemingly redundant information. Can you identify what I mean and explain why the speaker made the effort of conveying this information?

This wording improved the models' ability to identify the informational redundancy. GPT-4 correctly identified the redundancy for all stimuli, and provided lists of generic potential reasons (most commonly including emphasis, occasionally some level of atypicality, i.e. forgetting, but also mentioning humor or the wish to establish a connection). GPT-3.5-t pointed towards undefined noteworthiness and attributed it to a desire to emphasize this information. Despite Llama 3 labelling redundancies as problematic, the model provides reasonable and specific reasons for the redundancy. For the most part, the proposed reasons related to the conversational situation instead of the discussed activity.

We additionally experimented with further prompt formulations in order to elicit more specific explanations from the models. Best results were obtained when adjusting the question for each stimulus and detailing the specific redundancy, as shown in Q4:

- Q4: The second sentence in the direct speech conveys seemingly redundant information, because eating is a usual part of going to a restaurant. However, since it was mentioned explicitly, it can be assumed that it is new or relevant information. Why could Mary eating be new or relevant information?

For this prompt, atypicality was more often provided as the reason, or was listed among the possible reasons. GPT-4 mentioned atypicality 20 times, though often generically in form of the person potentially forgetting sometimes. Answers from GPT-3.5-t were consistent with atypicality inferences 11 times and mentioned information's noteworthiness as the reason, without elaborating any further. Llama 3 gave very specific and logical explanations of the noteworthiness for 22 stimuli, but only two of those could be classified as atypicality.

Step 4: Explicitly Accommodating Atypicality Finally, we were interested whether the model is in theory capable of 'completing' the picture that is caused by an atypicality by coming up with an alternative behavior or an explanation, i.e., whether the atypicality of the action can be accommodated if it is presupposed. The model was given the following prompt (again adjusted for each stimulus):

- Q5: The second sentence in the direct speech conveys seemingly redundant information, because eating is a usual part of going to a restaurant. However, since it was mentioned explicitly, it can be assumed that it is new or relevant information. That probably means that Mary doesn't typically eat. What does she do instead?

Here, GPT-4 provided sensible alternative behaviors for 13 stimuli while GPT-3.5-t managed to provide an alternative behavior for 14 stimuli (7 of these answers only weakly specified the alternative, i.e., 'uses alternative method'). In other cases, the models either rejected the premise for atypicality, provided alternatives that were not valid in the given context, or stated that the alternative could not be inferred from the text. Llama 3 again commited to specific and reasonable alternative behavior for most stimuli, only twice offering a weakly specified alternative and once an illogical one.

6 Conclusions

Exp. 1 demonstrated that the tested models are unable to draw atypicality inferences when prompted in a way that is similar to the instructions that humans receive. On the other hand, Exp. 2 showed that GPT-4 (but not the other two models) could draw atypicality inferences sometimes when prompted with examples, doing so in 65% of our stimuli. However, we also saw that typicality ratings were not always consistent with the generated justifications and that GPT-4's ability to draw these inferences is inconsistent and not robust. We conclude that performance improvements may stem from successful template matching rather than emulating the process correctly.

Our experiments into decomposing the atypicality inference task into different reasoning steps revealed that all models have the relevant script knowledge and can use this knowledge to identify the informationally redundant utterance. However, the models needed to be specifically prompted to identify these utterances, supporting the idea that the models' failure may relate to inability to apply conversational norms. Further evidence comes from the observation that Llama 3 fails to translate its excellent performance in accommodating explicit atypicality inferences and its claims about redundancies never being acceptable into good performance on Exp. 1 or Exp. 2.

Finally, we'd like to note that humans also do not uniformly draw atypicality inferences – variability exists at the level of items (some items exhibit a larger rate of atypicality inferences than others) and at the level of participants: Ryzhova et al. (2023) showed that in humans, the ability to draw atypicality inferences is correlated with reasoning ability. These two factors provide interesting leads for future research.

7 Limitations

One limitation from the NLP perspective of our study is that the size of the dataset is small (only 24 stories) and only in English. This is a common limitation of psycholinguistic studies due to the costs of human experiments.

This work only tests Zero-, and Few-Shot prompting and does not make use of any additional prompting methods designed for reasoning tasks. While we showed that the inferences are not derived in a human like manner without further input, it is therefore possible that the models could perform this task when prompted in a way that guides their reasoning more directly (i.e Fei et al. (2023) proposed a method called Three-hop Reasoning that breaks a task down into distinct reasoning steps that build on each other and increase in difficulty, and we see potential for applying such a method to our task in the future).

Another limitation lies in the selection of models, as it is does not cover the full range of different available architectures, due to not only the number of different models, but also the frequency at which they are released. For that reason we also do not include the newest OpenAI model GPT-4o.

A major limitation stems from only analysing the generated tokens and not their probabilities, as this is not supported by the OpenAI API. Furthermore, our efforts at testing a Likert scale in addition to 0% to 100% scale and requesting self-calibration (see appendix C) from the model through considering multiple answers cannot fully mitigate the potential problems of having the models output concrete values, and within our limited data we were unable to satisfyingly assess how consistently the model can actually adhere to any given scale. In that same vein, the faithfulness of externalized model reasoning has been previously questioned, and we can again not reliably assess the degree of faithfulness exhibited in our experiments. While this opens up avenues for further research, we believe that the combination of concrete values and explanations obtained, paired with our qualitative analysis of the performance on different steps provide a solid initial picture of the models abilities in terms of deriving atypicality inferences.

Finally, we have treated each model as a black box, only assessing their abilities through prompting, and only with a limited number of manually engineered prompts. Further research aimed more at the models' internal mechanisms, i.e. by probing and investigating the layer-wise capabilities, would be recommendable.

8 Acknowledgments

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 would like to thank the anonymous reviewers for their helpful comments. We thank Mayank Jobanputra for his help with the Llama model.

References

Chiara Barattieri di San Pietro, Federico Frau, Veronica Mangiaterra, and Valentina Bambini. 2023. The pragmatic profile of chatgpt: Assessing the communicative skills of a conversational agent. *Sistemi intelligenti*, 35(2):379–400.

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu,

Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam Mccandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2:1171–1182.

Joana Garmendia. 2023. Lies we don't say: Figurative language, commitment, and deniability. *Journal of Pragmatics*, 218:183–194.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Simon Jerome Han, Keith J. Ransom, Andrew Perfors, and Charles Kemp. 2024. Inductive reasoning in humans and large language models. *Cognitive Systems Research*, 83:101155.

Pengfei Hong, Deepanway Ghosal, Navonil Majumder, Somak Aditya, Rada Mihalcea, and Soujanya Poria. 2024a. Stuck in the quicksand of numeracy, far from agi summit: Evaluating llms' mathematical competency through ontology-guided perturbations. *arXiv preprint arXiv:2401.09395*.

Xudong Hong, Margarita Ryzhova, Daniel Adrian Biondi, and Vera Demberg. 2024b. Do large language models and humans have similar behaviors in causal inference with script knowledge? In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (* SEM 2024)*.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2022. A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801*.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. pages 9118–9147.

Ekaterina Kravtchenko and Vera Demberg. 2022. Informationally redundant utterances elicit pragmatic inferences. *Cognition*, 225:105159.

Zhuang Qiu, Xufeng Duan, and Zhenguang Garry Cai. 2023. Pragmatic implicature processing in chatgpt. *PsyArXiv*.

Margarita Ryzhova, Alexandra Mayn, and Vera Demberg. 2023. What inferences do people actually make upon encountering informationally redundant utterances? an individual differences study. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 45.

Timo Schick and Hinrich Schütze. 2020. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*.

Eli Schwartz, Leshem Choshen, Joseph Shtok, Sivan Doveh, Leonid Karlinsky, and Assaf Arbelle. 2024. Numerologic: Number encoding for enhanced llms' numerical reasoning. *arXiv preprint arXiv:2404.00459*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. 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).

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

A Prompts: Exp. 1 & Exp. 2

For Exp. 1 and Exp. 2 each model was given a system prompt that describes the task and provides an output template, and then each stimulus was appended in each setting individually. The system prompt for Exp. 1 was engineered iteratively with a small subset of stimuli using GPT-3.5-t, until arriving at prompt (1). The three main components we tweaked were the scale, the output format, and the behavioral directions. After optimizing the prompt for the GPT-3.5-t, it worked equally well for GPT-4 and Llama 3, hence the same prompt was used across all models.

> (1) You will receive a context (C) and two questions (Q1, Q2). Answer the questions by rating the frequency on a scale from 0% of the time to 100% of the time. Explain your answer in no more than two sentences. Always give a definitive answer, even if that means making assumptions and speculating based on common knowledge of human behavior. Additionally, tell me how a person that

knows the people mentioned in the context would answer the below questions, using the same scale and explaining their answer in no more than two sentences.

Use the following template for your output, where '<>' is a placeholder for content:

X: <Responder: AI or Human>

Q: <Question>

A: <Answer>

R: <Reasoning>

For the scale, the experiments by Kravtchenko Demberg (2022) used a continuous sliding scale from Never to Always, mapping to values of 0 and 100 respectively. Attempts at similar scale failed to elicit consistent response categories, and ultimately we needed the model to output its ratings directly in values. Hence a scale of 0% o 100% of the time was established, which closely corresponds to the initial scale, but appeared more consistent and accessible to the model.

The model output needed to be constrained to a format from which the ratings and reasonings could easily be retrieved. We experimented with different instructions as well as template designs (i.e. different placeholders, separators etc.) and found the simple and concise variant presented in (1) to be most consistently adhered to. While GPT-3.5-t and Llama 3 somewhat frequently generated output that did not fully adhere to this format, we also found that this template constrained the output enough that the majority of output could be parsed automatically, hence minimizing the ratings that needed to be extracted manually. GPT-4 generated output that very closely adhered to the template.

For the behavioral instructions, we found that the models needed to be explicitly told to speculate and make assumptions, as they would else refuse a response on the grounds of a lack of necessary context or information. Telling the models that a definitive answer was required further facilitated their ability to commit to a response, though occasionally a definitive answer was still not given. We initially encountered frequent problems with the model refusing to answer because it was "just a language model", which led us to additionally request a second response, where the model pretends to be a human who knows the characters in the stimulus. Ultimately, the other tweaks to the prompt improved this behavior to the point where the model also consistently provided answers as "itself". Since a paired t-test revealed no significant difference between the two types of responses (i.e. responses as the model and responses pretending to be a human), we did not uphold a distinction between those data points in the further analysis.[7]

For Exp. 2 we used the same system prompt, only adding the information that the model would be provided with two examples (2). The two examples were appended prior to the stimulus, and were crafted manually to mirror the atypicality response we expected in the habitual utterance condition. Notably, providing examples in the correct output format increased GPT-3.5-t's and Llama 3's ability to adhere to the template.

(2) You will receive a context (C) and two questions (Q1, Q2).

Answer the questions by rating the frequency on a scale from 0% of the time to 100% of the time. Explain your answer in no more than two sentences. Always give a definitive answer, even if that means making assumptions and speculating based on common knowledge of human behavior.

Additionally, tell me how a person that knows the people mentioned in the context would answer the below questions, using the same scale and explaining their answer in no more than two sentences.You will be provided with 2 examples (Ex1, Ex2).

Use the following template for your output, where '<>' is a placeholder for content:

X: < Responder: AI or Human >

Q: <Q1 or Q2>

A: <Answer>

R: <Reasoning>

B Exp. 1: Additional Results

As noted, we obtained results from the Zero-Shot prompting in a total of 6 conditions. The manipulation of the context to state atypical behavior

[7]The data for the recently released Llama 3 was collected after we had already deemed this distinction unecessary, hence the relevant sentence was removed from the prompt when prompting Llama 3 and we only collected one data point for each stimulus.

(wonky context) reduced the baseline typicality ratings in all models. The rating change after encountering redundancies was minimal for GPT-3.5-t, only somewhat higher for GPT-4 and almost double for Llama 3 (cf. Table 5). Encountering only a minor rating change is in line with the human results obtained by KD. As indicated by a very high standard deviation the effect of voiding script knowledge did vary greatly across stimuli, i.e. not all activities were equally strongly influenced by the manipulated background.

Additionally, we also looked at the typicality ratings in the non-conv. habitual utterance condition. At baseline the activity was indeed rated to be very atypical, with a high standard deviation again showing differences across the stimuli, and there was a relatively high rating change in the utterance condition (cf. Table 6). While the low baseline rating is in line with the observations in Kravtchenko and Demberg (2022), the effect size is larger in the models than in humans. While we did not annotate the provided explanations for these conditions, the observation that the typicality is higher in the utterance condition appears to be in line with the reinforced utterance reasoning that we observed for habitual activity, i.e. something being rated as typical because it was mentioned.

C Zero-Shot: Likert Scale and Calibration

To increase our confidence in the validity of concrete values the model has been outputting, we also collected ratings on the below 7-point Likert scale:

1. Never
2. Rarely, less than 10% of the time
3. Occasionally, 30% of the time
4. Sometimes, about 50% of the time
5. Frequently, about 70% of the time
6. Usually, about 90% of the time
7. Every time

Using this scale did once again not yield significant rating change for GPT-3.5-t. For Llama 3 and GPT-4 the rating change is significant and occurs, as previously seen, in the opposite direction, i.e the conv. habitual activity is judged to be more frequent when the utterance is seen (GPT-4 6.58 →6.75; t(23) = -2.14, $p < .05$; Llama 3 5.62 →6.32; t(23) = -3.39, $p < .005$)

The same sanity check as performed above did show that for all three models there is no significant rating change for the conv. habitual utterance when the non-habitual utterance is present. Furthermore the results of using a wonky context, i.e. voiding the script knowledge, and the typicality rating of the non-habitual activity are in line with results reported in appendix B.

Additionally, we tried an approach for asking the model to self-calibrate its responses that was introduced by Tian et al. (2023). They have taken inspiration from human psychology showing that considering multiple possible answers can mitigate over-confidence, and consequently ask the models to provide multiple responses that they had to assign likelihood to. We applied their strongest approach of considering 4 responses and assigning a probability p = (0.0, 1.0).

We were not able to adjust the proposed calibration method to our task in such a way that Llama 3 could consistently generate multiple responses, despite the Tian et al. (2023) using Llama-2-70b-chat for their experiments. We attribute this to our more complex task and output format, and consequently cannot report results for Llama 3. For GPT-3.5-t and GPT-4 the results were indiscernible from the regular zero-shot prompting presented in 3, i.e for the conv. habitual activity the non-significant rating change for GPT-3.5-t is in the opposite direction (88.5 →94.0), and for GPT-4 we see very high typicality ratings and no rating change in the presence of the conv. habitual utterance.

D Perturbation Analysis: Further Results

Here we provide the results of the first and second perturbation analysis for GPT-3.5-t and Llama 3, as well as the results of an additional prompt perturbation experiment for all three models.

D.1 Perturbation 1

Both GPT-3.5-t and Llama 3 stopped drawing atypicality inferences both in ratings and reasonings, with a more drastic effect in Llama 3, which reverted back to assigning very high ratings in the conv. utterance condition (mean: 84.5 →93.9) and did not have atypicality represented in that ratings at all. In the non-habitual utterance condition, the ratings did not increase or decrease (mean: 83.2). GPT-3.5-t did keep with the previous trend of lower

	Wonky baseline		Wonky habitual utterance	
model	**mean**	**sd**	**mean**	**sd**
GPT-3.5-t	52.71	37.34	50.20	41.84
GPT-4	35.89	39.19	41.46	40.29
Llama 3	16.2	24.8	31.9	38.9

Table 5: Typicality ratings for the habitual activity in the wonky context conditions

	Baseline		Non-conventional utterance	
model	**mean**	**sd**	**mean**	**sd**
GPT-3.5-t	21.19	24.27	27.25	31.54
GPT-4	33.96	25.19	44.39	25.15
Llama 3	36.2	36.2	23.3	23.8

Table 6: Typicality ratings for the non-conventional activity in the normal context condition

ratings that were not statistically significant, and exhibited atypicality twice in the provided reasonings.

D.2 Perturbation 2

For GPT-3.5-t the difference between A and B is not as pronounced, and *atypicality_reject* is observed much less frequently than in GPT-4. Both the modeled rating with the modeled atyp reasoning (i.e. copying down), and the modeled rating with an adjusted matching *no_atypicality* reasoning appear approximately equally frequently. For (B) there is notably an increased number of the modeled atypicality reasoning being paired with an adjusted rating. Llama 3, on the other hand, rarely emulated both the modeled reasoning and rating. For (B) it matches the rating of 100% with an appropriate reasoning more than half the time, and for (A) this happens the majority of the time. When the model does replicate the modeled atypicality reasoning, it is most of the time paired with a typicality rating of 0%. Notably, Llama 3 does actually generate a few nonsensical or incongruent responses in this setting.

D.3 Additional perturbation experiments

We also crafted a second set of intentionally misleading and incongruent exemplars, modeling the reverse behavior, i.e. pairing a low rating that implies atypicality with a reasoning that models no atypicality. We tried both providing a low rating (<40%) and a rating of 0% . In these experiments, all three models very consistently reproduce the reasoning but ignore the modeled rating and assign a high matching rating instead.

E Prompts: Exp. 3

Below we provide the alternative questions/wordings for the steps and discuss briefly why they were less efficient than their counterparts. Additionally, we adjusted the system prompt to reflect the new instruction, telling them that they were experts on human behavior and had the task of answering a question based on a provided context (3)[8]. The models were then given each stimulus in the critical condition, and one question aimed at identifying the redundancy at a time.

> (3) You are an expert on human behavior and communication who will be answering a question based on short contexts (C). There is no right or wrong answer to the questions you'll see, and you are willing to use your best judgement and commit to a concrete, specific response, even in cases where you can't be sure that you are correct.
>
> Please keep your answer as short and concise as possible. Use the following template for your output, where '<>' is a placeholder for content:
>
> Q: <Question>
>
> A: <Answer>

E.1 Step 1:

For this step, the following alternative questions were tested:

[8]This system prompt is adapted from Han et al. (2024). It was also tested as a system prompt for Exp. 1 during the prompt engineering process but unlike here, it did not lead to a more consistent performance.

- A_Q1: Is any part of the direct speech superfluous or unnecessary?
- A_Q2: Does the context (C) contain any redundancies?

With A_Q1 we replaced the word redundancy, as we thought it might be too specialized, i.e. not be the word a laymen would chose to describe the phenomenon. For the most part this did however perform on par with Q1 reported in the paper, ultimately showing that this distinction did not matter to the models. With A_Q2 we opted for an even more open-ended approach by not restricting the potential redundancies to the direct speech. This did however, unsurprisingly yield even fewer identifications of the desired redundancy.

E.2 Step 2:

As explained before, identifying the models' ability to perform this step was challenging through a set of questions was challenging due to its subtlety, and since humans might also not verbalize their implicit understanding of the conversational norm that is violated by redundancies (Maxim of Quantity). Ultimately, we used the following questions to gauge a more general understanding of the models' awareness of conversational norms:

- A_Q3: The second sentence in the direct speech provides redundant information, since the action it talks about is already implied in the first sentence. Do you think this was an acceptable utterance?
- A_Q4: The direct speech contains redundant information. Is providing redundant information a good and efficient way of communication?
- A_Q5: The direct speech contains redundant information. Do you see any issue with that?

For A_Q3 the utterance was mostly deemed acceptable by GPT-3.5-t and GPT-4, and when reasoning was provided in the model response, it would be very general, usually suggesting that the redundancy served the purpose of emphasizing or expressed general noteworthiness. Llama 3 on the other hand found the utterance mostly not acceptable, sometimes reasoning that it may serve as emphasis or to provide nuance, but mostly classifying them as unnecessary or even awkward. For A_Q4 GPT-3.5-t answered no 24 times without elaborating further. GPT-4 and Llama 3 also agreed that it is not acceptable and elaborated why (i.e. confusing, waste of time), but the majority of time it was then also stated that there still might be good reasons (i.e. emphasizing, clarification). For A_Q5, GPT-3.5-t saw no issue for most items, and the remaining times it said there was an issue with redundancy, though usually not the informational redundancy we were investigating but one from the broader context (i.e. "Don mentions that he took a train with Jane, which is already implied by the fact that he saw Jane at the subway station and they took the train together", which is arguably not a redundancy because the character he tells this to does not know that he saw her and that they took a train). GPT-4 generally saw no issue, occasionally stating the (correct informational) redundancy and for each item elaborating reasons the redundancy occurred. These reasons are however mostly very general and broad (i.e. emphasis, enthusiasm, creating a relaxed atmosphere, establishing a connection). Similarly, saw either no issue, or no major issue with the redundancy, and when elaborating on the informational redundancy it provided a reasonable purpose for expressing it. Finally, with A_Q5 Llama 3 did actually identify the informational redundancy we were looking for for the majority of the stimuli, and did proclaim that it was an issue.

E.3 Step 3:

Find below additional questions we tested for this step:

- A_Q6: The second sentence in the direct speech conveys seemingly redundant information. Providing redundant information can be unnecessary and inefficient for communication. Why was the redundant utterance made?
- A_Q7: The second sentence in the direct speech conveys seemingly redundant information. Providing redundant information can be unnecessary and inefficient for communication. Consider only what you can tell about the people from the provided context (C) and tell me definitively: Why did the speaker still choose to express the redundant information in this specific situation?
- A_Q8: The second sentence in the direct speech conveys seemingly redundant information. Providing redundant information can

be unnecessary and inefficient for communication. However, the speaker made the effort of conveying this information. Since they have no reason to be inefficient, this information must actually be new or important. What new or relevant information can you infer from the second sentence?

A_Q6 and A_Q7 resulted in very general responses from GPT-3.5-t and GPT-4 that covered the same potential reasons for redundancy that have been stated in previous steps. Llama 3 also provided similar reason but once again did a better job of applying them to the specific scenario rather than keeping them general. Notably, atypicality was not among the reasons that Llama 3 came up with. For A_Q8, GPT-3.5-t defaulted to just stating the exact contents of the sentence, while GPT-4 performed slightly worse than with the for each item adjusted Q4 reported in the paper (i.e. it gave appropriate reasons, but not as specific to the item content, and fewer explanations pointing towards atypicality). Llama 3 unsurprisingly showed a similar performance to the other questions as the model has less of a tendency to generalize.

E.4 Step 4:

For step 4 we did not experiment further, and instead just directly adapted the best performing question from step 3 by inserting the desired atypicality answer and then adding a simple question to elicit alternative behavior.