Beyond the Bias: Unveiling the Quality of Implicit Causality Prompt Continuations in Language Models

Judith Sieker and **Oliver Bott** and **Torgrim Solstad** and **Sina Zarrieß**
Bielefeld University
{j.sieker, oliver.bott, torgrim.solstad, sina.zarriess}@uni-bielefeld.de

Abstract

Recent studies have used human continuations of Implicit Causality (IC) prompts collected in linguistic experiments to evaluate discourse understanding in large language models (LLMs), focusing on the well-known IC coreference bias in the LLMs' predictions of the next word following the prompt. In this study, we investigate how continuations of IC prompts can be used to evaluate the text generation capabilities of LLMs in a linguistically controlled setting. We conduct an experiment using two open-source GPT-based models, employing human evaluation to assess different aspects of continuation quality. Our findings show that LLMs struggle in particular with generating coherent continuations in this rather simple setting, indicating a lack of discourse knowledge beyond the well-known IC bias. Our results also suggest that a bias congruent continuation does not necessarily equate to a higher continuation quality. Furthermore, our study draws upon insights from the Uniform Information Density hypothesis, testing different prompt modifications and decoding procedures and showing that sampling-based methods are particularly sensitive to the information density of the prompts.

1 Introduction

There is currently a growing interest in probing the performance of large language models (LLMs) on carefully controlled linguistic test suites and experimental datasets to get a deeper understanding of specific linguistic capabilities captured in these models (e.g., Belinkov and Glass, 2019; Ettinger, 2020). While a lot of previous work focused on analyzing the syntactic competence of LLMs (e.g, Hu et al., 2020; Schuster and Linzen, 2022), recent studies also started to investigate the abilities of LLMs on the level of semantics and pragmatic discourse processing. One promising diagnostic for probing discourse knowledge in LLMs has turned out to be the use of Implicit Causality (IC) prompts. IC refers to a property of a broad range of interpersonal verbs that exhibit strong preferences for establishing coreference to one of the verb's arguments over the other in explanations. For instance, when asked to provide a continuation after "..." in a sentence like (1), humans display strong next-mention preferences towards the stimulus (*he/Tom* in this case):

(1) Tom fascinated Sarah because... *he was very smart.*

As the IC bias has been extensively researched in psycholinguistics and psychology across various languages and populations (e.g., Ferstl et al., 2011; Hartshorne et al., 2013; Bott and Solstad, 2014), investigating this bias in LLMs has gained significant interest. A range of recent studies investigated LLMs' predictions of the next mention in examples like (1) and whether these mentions (i.e. pronouns) follow the same coreference biases as can be found in human data (e.g., Upadhye et al., 2020; Davis and van Schijndel, 2020; Kementchedjhieva et al., 2021; Zarrieß et al., 2022). These studies predominantly indicated that LLMs are not generally congruent with the human IC bias, which has been interpreted as evidence for LLMs struggling with certain aspects of discourse understanding (but see Cai et al., 2023).

In this work, we propose that experimentally elicited data of human continuations of IC prompts cannot only be used for analyzing *comprehension* in LLMs, but constitutes an excellent basis for analyzing LLMs' discourse-level *generation* capabilities, i.e. going beyond the prediction of the next mention. While discourse-level downstream tasks in NLG, e.g. story generation or summarization, are complex and notoriously difficult to evaluate systematically with respect to targeted linguistic capacities of NLG systems, IC continuations provide a well-controlled diagnostic of discourse knowledge and, at the same time, rather simple sentences

Proceedings of the 16th International Natural Language Generation Conference, pages 206–220
September 11–15, 2023.

whose quality can be easily assessed in human evaluation. Yet, to date, only little consideration has been given to the extent to which IC continuations generated by language models maintain semantic appropriateness and coherence with respect to the given IC prompts, cf. Huynh et al., 2022, and to what extent congruency with the human coreference bias is related to the quality and coherence of the generated continuation.

We expand prior work on discourse knowledge captured by LLMs in IC contexts and investigate their ability to generate not only bias-congruent but also sensible continuations. In contrast to most previous studies, we are not only interested in the LLMs predictions of the first word following the prompt (i.e. the pronoun), but in the quality of the generated sequences and their comparison against human continuations. We would like to emphasize that IC bias can be violated without any loss of discourse coherence. As an example, consider the following generated sequences in (2), where the first sentence is not congruent with the human bias but coherent, whereas the second sentence is congruent with the bias but not coherent:

(2) a. Jonathan shocked Charlotte because... *she didn't think he would steal.*
b. Jonathan admired Charlotte because... *she handed him a pineapple.*

In order to assess the quality of generated IC continuations, we carry out an evaluation study with human ratings of naturalness, coherence, and informativity. We aim to investigate to what extent continuation quality is associated with bias congruency and identify the factors that may influence this interaction. Thus, we manipulate two different types of conditions. First, we evaluate the predicted continuations of language models on two types of IC prompts: "standard" IC prompts (such as those in (1) and (2)) and IC prompts that are extended with adverbial modifiers. Second, we evaluate the performance of three different decoding procedures. In the following, Section 2 will present background on the set up of our study, Section 3 describes the hypotheses of our evaluation experiment and Section 4 describes the results.

2 Background

2.1 Implicit Causality

Psycholinguistic literature has consistently shown that numerous interpersonal verbs exhibit a bias known as "Implicit causality" (IC) (Garvey and Caramazza, 1974). That is, when asked to provide a continuation after "..." in sentences like (1) and (2), humans display strong next-mention preferences towards the subject for stimulus-experiencer (SE) verbs like "fascinate" (1) and "shock" (2-a) and towards the object for experiencer-stimulus (ES) verbs like "admire" (2-b) (Solstad and Bott, 2022). Continuations that align with the coreference bias are referred to as bias-congruent, while continuations that go against the bias are considered bias-incongruent. In addition, verbs that exhibit a pronounced IC bias also tend to exhibit a coherence bias, that is, they are prone to trigger explanations in subsequent discourse (cf., e.g., Kehler et al., 2008, Bott and Solstad, 2014). Interestingly, it has been shown that the coreference and coherence tendencies of IC verbs are strongly modulated by linguistic context. For instance, in their study, Bott and Solstad, 2021 showed that modifying standard IC prompts (such as those in (1) and (2)) with causal adverbial phrases leads to a change in the coreference bias as well as the coherence bias. In particular, not only did their results show that the IC coreference bias was in fact eliminated when adverbial modifiers were added to the prompts, they further found fewer and distinct types of explanations after these modifications compared to the default explanation types observed in the "standard" condition.

In psycholinguistic studies, the coreference and coherence biases of IC verbs are often elicited in sentence continuation tasks, typically giving the participants prompts such as those in the discussed examples. And, as several psycholinguistic studies have demonstrated that the IC bias is not only highly reliable but also robust across different languages (Ferstl et al., 2011; Goikoetxea et al., 2008; Hartshorne et al., 2013; Bott and Solstad, 2014), it has become an an intriguing domain for testing language models. Earlier studies, including those conducted by Upadhye et al., 2020, Davis and van Schijndel, 2020, Kementchedjhieva et al., 2021 and Zarrieß et al., 2022, have examined the performance of LLMs in capturing the IC coreference bias. I.e., they concentrated on single-word prediction tasks and evaluated the models' ability to generate continuations of such classic prompts, like examples (1) and (2), and predominantly found that LLMs display limited ability to systematically incorporate the IC coreference bias in their genera-

tions. In addition to examining the IC bias, Huynh et al., 2022 conducted a human evaluation of the quality of the continuations predicted by an English GPT-2 model. Asking the participants to judge the "reasonability" of the generated continuations on a 5-point Likert scale (with 5 being strongly reasonable), their results showed that only 32% of all the continuations retained an average rating greater than or equal to 4. In this study, we build upon these priors results, by investigating, similar to Huynh et al., 2022, the extent to which bias congruency in LLMs is associated with the quality of the predicted continuations, i.e. we go beyond the next word prediction. Additionally, we go beyond previous research by not only taking into account the impact of the decoding procedure, but also by investigating whether the grade of information density in the prompts affects the models' ability to produce meaningful continuations.

2.2 Information Density and Decoding

The use of natural language for communication is often explained through information theory (Shannon, 1948), an approach that views linguistic units as messages aimed at conveying information, quantified by their probability of being produced, which is also termed "surprisal". Within this view, units with low probability are considered more informative, reflecting the intuition that unpredictable elements convey more information than predictable ones (see, e.g., Meister et al., 2021 for an elaboration). The Uniform Information Density (UID) hypothesis (Levy and Florian Jaeger, 2007; Jaeger, 2010) further predicts that speakers prefer to distribute information uniformly across their utterances. I.e., if speakers wish to convey more information, they are more likely to distribute this information across more words. Previous studies have shown that a more uniform distribution of information is strongly associated with higher linguistic acceptability (cf., e.g., Meister et al., 2021). As such, assertions that adhere to the UID hypothesis are considered to be of higher quality and receive better evaluations. Furthermore, also the choice of decoding strategy, i.e. the decision rule used to determine the output sequence of a model, is one of the most important factors that affects the quality and various linguistic properties of the generated text, as several papers have demonstrated (e.g. Holtzman et al., 2019; Zarrieß et al., 2021; Meister et al., 2022; Ji et al., 2023). However, the success of the decoding procedure shows to be contingent on the task at hand, and no decoding approach has demonstrated a consistent correlation with producing high-quality text (Wiher et al., 2022). Given these insights, in this study, we consider it worthwhile to explore how the level of information density of the prompts as well as the decoding procedure used influences the quality of the generated continuations. We will outline our approach and hypotheses in the next section.

3 Hypotheses and Conditions

3.1 Prompt Manipulation and Information Density

First, to investigate an important aspect of IC verbs that, to our knowledge, has not yet received attention in the study of IC in LLMs, we introduce an additional prompt condition alongside the "standard" IC prompts (such as those in the examples discussed so far). That is, we extend these standard prompts with adverbial modifiers, so that, for example, the prompts in (2-a) and (2-b) are augmented to (3-a) and (3-b), respectively.

(3) a. Jonathan shocked Charlotte by his aggressive appearance in the talk show because...
b. Jonathan admired Charlotte for her extraordinary agility in this year's ice skating competition because...

This extension is motivated by the aforementioned insight that the coreference and coherence biases of IC verbs are strongly modulated by linguistic context (cf. Section 2.1). In this current study, we therefore compare standard IC prompts with prompts extended with adverbial modifiers, using the experimental data of Bott and Solstad, 2021, in order to investigate to what extent this added contextual information may not only influence the ability of LLMs to capture the IC bias but, beyond, also the quality of the model-generated continuations.

Building upon the UID, we make the assumption that standard IC prompts, which are brief and contain only minimal information, will require more information in their sentence continuations to maintain a uniform distribution of information across the whole sentence than prompts that are more detailed. For instance, consider Figure 1, which depicts the token probabilities and information values for a "standard" unmodified IC prompt and for a prompt

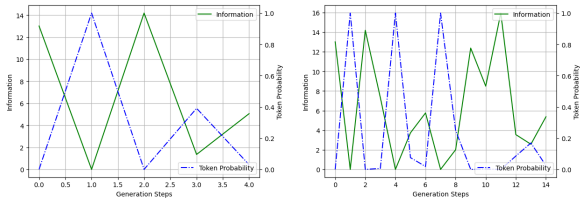


Figure 1: Token probabilities (blue) and information values (green) for each prompt condition for a selected item involving the ES verb *bewundern* ("admire"). Left-hand side: unmodified prompt, i.e. "Paul bewunderte Isabel, weil" (*Paul admired Isabelle, because*). Right-hand side: extended prompt, i.e. "Paul bewunderte Isabel für ihre außerordentliche Geschicklichkeit beim diesjährigen Eisschaulaufen, weil" (*Paul admired Isabel for her extraordinary agility in this year's ice skating competition, because*).

augmented with an adverbial modification. Here it is visible that the latter exhibits more peaks towards lower probability, and, respectively, towards higher informativity – since here, in addition to the verb, there is another information-carrying part (i.e. the adverbial modification). Put differently: The IC prompts that are extended with adverbial modifications inherently carry more information and therefore already contribute a greater amount of information to the (yet to be completed) sentence. In fact, as Bott and Solstad, 2021 show, prompts with adverbial modifiers such as (3) provide comprehensive causal scenarios in themselves lacking any need for further causal elaboration.

Hypotheses Thus, assuming that speakers aim to distribute information uniformly across an utterance, we can infer that continuations of the modified IC prompts would require less information than those of the shorter prompts. Consequently, when presented with standard (i.e. short) IC prompts, LLMs are expected to produce continuations that are less probable and, therefore, more informative (or: surprising), while with the modified IC prompts, less informative and therefore, more probable continuations are anticipated from the models. We hypothesize, however, that LLMs encounter difficulties in producing informative yet sensible continuations, which could explain the observed lower quality of sentence continuations for these prompts, aligning with research suggesting that a more uniform distribution of information is strongly associated with higher linguistic acceptability (see Section 2.2). Taken together, we expect that sentence continuations following prompts augmented with adverbials will be evaluated as of higher quality than those following standard IC prompts, as the additional information provided reduces the model's burden to generate informative content on its own.

3.2 Information Density and Decoding

In this study, we acknowledge the significant impact of decoding methods on the quality of generated text (see Section 2.2) and therefore also investigate whether the distinct information-theoretic characteristics of three different decoding methods allow them to handle the prompt requirements differently. For instance, adhering to the terminology of UID theory, where information density is measured in information-theoretic terms of surprisal, maximation-based strategies, such as beam search decoding, for example, are known for producing text that is more probable but less surprising, and thus less informative. In contrast, stochastic strategies, such as Nucleus Sampling (Holtzman et al., 2019), for example, tend to produce text that is less probable and, therefore, more suprising, i.e. informative (cf., e.g., Zarrieß et al., 2021).

First of all, we consider seam search to be an appropriate candidate for the task of generating sentence continuations for (short) IC prompts, since Meister et al., 2020 have shown that beam search incorporates an inductive bias that aligns with the UID principle to distribute information uniformly across an utterance. However, as it has been shown that beam search tends to reduce diversity by favoring likelihood (Schüz et al., 2021; Zarrieß et al., 2021), in this paper, we will utilize the Diverse Beam Search method proposed by Vijayakumar et al., 2018, which, in essence, encourages diverse candidates by categorizing candidates into groups and then enforcing diversity within those groups. Furthermore, we consider Nucleus Sampling (Holtzman et al., 2019) to be an appropriate stochastic decoding method for our investigation. By truncating the model distribution, this method effectively addresses the drawback of sampling based methods to potentially select very low probability outputs that may considerably reduce the overall quality and coherence. In addition, in this study, we will consider another, third decoding procedure called Local Typical Sampling (Meister et al., 2022) to generate continuations for the two IC prompt conditions. Including this method in our study seems worthwhile, given that the authors introduced it as a potential solution to the shortcomings of beam search and Nucleus Sampling and, beyond, that it was designed to embody the

characteristics of human language, aligning with the information-theoretic perspective discussed in this context (cf. Meister et al., 2022). In a nutshell, in Local Typical Sampling, the options to sample from are limited to strings that have a similar information content to what would be expected based on the prior context.

Hypotheses In light of this, we anticipate that stochastic decoding methods could result in better sentence continuations for the standard IC prompts compared to beam search, for instance, as more information is required in the continuations that follow these short prompts. On the contrary, for prompts extended with adverbials, where less informative continuations are expected to retain a uniform distribution of information across the sentence, beam search decoding may lead to better results. Moreover, we anticipate that Local Typical Sampling will lead to adequate informative generations for both short and modified prompts, i.e. resulting in comparable quality of continuations across the two prompt conditions. Taken together, while we generally expect longer prompts (i.e. those augmented with adverbials) to result in better quality continuations, we further expect the decoding strategy employed to also play a significant role in the generated text's quality. Further, we anticipate that the impact of the decoding strategies will vary depending on the prompt construction.

4 Experimental Setup

Data. We ground our study on German IC data from Bott and Solstad, 2021 and use their experimental items to generate German prompts to be completed by the LLMs. These data also provide us with human-generated sentence continuations for both prompt conditions, offering a valuable reference point for evaluating the model-generated continuations. Prompts consist of simple sentences introducing the verb, the verb's arguments and the connective *weil* ('because') (as in (1) and (2)). To further investigate the effect of causal modification, we designed the following four conditions:

1. SE verbs in "standard" prompt constructions (e.g., *Clara inspired Vincent because...)*
2. SE verbs + *durch* ('by') modifier (e.g., *Clara inspired Vincent by her innovative lecture because...)*
3. ES verbs in "standard" prompt constructions (e.g., *Paul admired Isabel because...*)
4. ES verbs + *für* ('for') modifier (e.g., *Paul admired Isabel for her extraordinary agility in this year's ice skating competition because...*)

As in previous studies, we vary and balance prompts for the names and gender of verb arguments. To ensure that the prompts' informational content remains unaffected by the verb arguments, we deliberately keep the proper names consistent across the conditions. In addition, we also consider the sentiment of the verbs, ensuring a balance between those that are positively and negatively connotated. See Table 4 in the Appendix for the chosen verbs and examples for each prompt condition.

Models. We use the Hugging Face framework for reproducibility, employing two German language models to generate continuations for the IC prompts: (i) a pretrained German GPT-2 model[1] and (ii) a pretrained multilingual mGPT model[2] which reproduces the GPT-3 architecture. Following the reasoning of Huynh et al., 2022, we use GPT-2 and mGPT, despite the availability of newer and more advanced language models, as they offer a manageable size that is compatible with standard hardware and a favorable trade-off between complexity and efficiency. For both models we do not employ any finetuning.

Decoding hyperparameters. We set the hyperparameters of our decoding methods by validating them on the continuation data from Bott and Solstad, 2021, using other IC verbs (and thus, items) than those tested in the final experiments and the metrics BLEU, GLEU, METEOR, ROUGE(-L) and BERTScore, as provided by the Hugging Face library. We chose ranges of hyperparameters based on the authors' self-reported best-performing values and/or standard values recommended in literature and found the following settings to be best performing in at least four of the five metrics:

- Diverse Beam Search: beam size and beam group size = 10; diversity penalty $\lambda = 0.7$
- Nucleus Sampling: top p value = 0.85; temperature = 0.7
- Typical Sampling: typical p value = 0.9; temperature = 0.7

[1] https://huggingface.co/dbmdz/german-gpt2.
[2] https://huggingface.co/ai-forever/mGPT.

Automatic evaluation. To assess whether the IC coreference bias is reflected in the models' continuations, we adopt the method of Zarrieß et al., 2022 and calculate Completion Sensitivity scores, i.e. the percentage of prompts for which the model's predicted pronoun aligns with the bias. Further, to investigate the overall quality of the models' continuations, we calculate the three automatic measures: BLEU, ROUGE-L and BERTScore, in this way comparing them to the ones produced by participants in the studies conducted by Bott and Solstad, 2021.

Human evaluation. As human evaluation remains to be the gold standard when it comes to assessing the overall quality of a system (van der Lee et al., 2021; Schuff et al., 2023), we employ human judgment to investigate the quality of the generated continuations. Looking over the produced generations, there were three items that contained offensive content, e.g. including instances of sexism, which is why these items were excluded from the experiment. We used the Prolific framework to obtain ratings from 40 different annotators on 96 examples for each model's continuations, 128 examples for each prompt condition, 64 examples for each decoding method and 64 examples for human-produced continuations of which one half was bias congruent and the other half bias incongruent. In this way, we overall collected 5120 human judgments. We restricted potential evaluators to native speakers of German with their country of residency being Germany. We designed the evaluation as a rating task, presenting three questions to the evaluators. We asked whether the continuation is a (i) "natural", (ii) "meaningful" (i.e. coherent) and (iii) "surprising" (i.e. informative) explanation for the respective sentence beginning (i.e. prompt). The evaluators could indicate their agreement on a five-point Likert scale, ranging from 1 ('strongly disagree') to 5 ('strongly agree'). For each of the criteria, we take the median score across evaluators as the final score. We chose the criteria of naturalness and coherence following recommendations from van der Lee et al., 2021 and elicited informativity to investigate our assumptions described in Section 3. Before evaluators were asked to provide ratings on the three criteria, they were presented with instructions as well as an example item on the basis of which the criteria were explained. Evaluators were paid £9,00/hour and IRB approval was obtained prior to conducting the study.

	Naturalness	Coherence	Informativity
Standard IC Prompt			
Diverse Beam Search	4 (**3.55**)	3 (**2.87**)	2 (2.50)
Nucleus Sampling	4 (3.26)	2 (2.55)	3 (2.62)
Typical Sampling	3 (3.26)	3 (2.74)	3 (**2.65**)
Human bias-congruent	5 (4.77)	5 (4.75)	2 (2.39)
Human bias-incongruent	4 (3.82)	3 (3.20)	3 (2.47)
Modified IC Prompt			
Diverse Beam Search	4 (**3.69**)	3 (**3.06**)	2 (**2.54**)
Nucleus Sampling	3 (2.90)	2 (2.04)	2 (2.40)
Typical Sampling	3 (2.99)	2 (2.24)	2 (2.52)
Human bias-congruent	5 (4.56)	5 (4.61)	3 (2.56)
Human bias-incongruent	5 (4.54)	5 (4.37)	3 (2.61)

Table 1: Human evaluation results for each decoding procedure and each IC prompt condition, aggregated over the text generations of both LLMs. For each criterion, we report the median score across raters as the final score (additionally, the mean values are reported in brackets). Bold values indicate conditions with the best values for that evaluation criteria.

5 Results

Naturalness and coherence. Generally, when comparing the two language models, GPT-2 and mGPT, the automatic metrics presented in Table 5 in the Appendix indicate that mGPT exhibits a tendency to generate more favorable, or rather, more similar continuations compared to the human-produced ones. However, this observation holds limited significance since, as detailed in a subsequent paragraph, there is a notably low correlation between automatic and human metrics. Thus, as we did not find substantial differences in continuation quality between GPT-2 and mGPT, we focus our analysis of human ratings on differences between the decoding methods and prompt conditions and aggregate over the models (but see Figure 5 in the Appendix for results separated for models). Table 1 shows the results of the human evaluation for each decoding procedure and each prompt condition. First, we note that, as expected, human-produced continuations achieve the best results for the criteria naturalness and coherence. Further, as we had hypothesized, it is the case that the influence of the decoding procedure varies depending on the IC prompt condition. Contrary to our expectations, however, it is not generally the case that the modified prompts lead to better-evaluated continuations. Likewise, it is not the case that the sampling-based methods result in better-evaluated continuations of the standard IC prompts while Diverse Beam Search leads to better-evaluated continuations of the modified prompts. Instead, we find that, across both prompt conditions, Diverse

Beam Search yields equally good and almost always best results – both for the automatic metrics and for the human evaluations. Remarkably, across prompt conditions and decoding procedures, we observe substantial distinctions between naturalness and coherence. I.e., the medians for naturalness consistently exhibit higher values, indicating that the generated continuations are perceived as fluent. However, in terms of coherence, the average ratings are lower, suggesting a lack of logical consistency in the generated text. This observation highlights the challenges faced by LLMs in generating sensible continuations in this rather simple task. It clearly shows that LLMs especially struggle with discourse-level coherence of explanations in IC contexts, beyond predicting the mentions that are congruent with this bias. This finding further underscores the importance of considering multiple evaluation criteria to assess the outputs of LLMs.

Informativity. Referring to Table 1, it is evident that, unlike the criteria of naturalness and coherence, the ratings for informativeness do not significantly favor human continuations over those generated by the models. This observation indicates that, as expected, the human-produced continuations align more closely with prototypical explanations making them appear comparatively mundane. Further, as consistent with our expectation, we observe that for the shorter prompts, the sampling-based methods produce continuations that the evaluators deem more informative compared to their continuations of the modified prompts. This observation also holds true for Typical Sampling, which we initially anticipated to yield similarly informative generations for both short and modified prompts. Interestingly, for Nucleus Sampling, the higher perceived informativeness of the continuations is accompanied by an increase in naturalness, whereas for Typical Sampling, it is linked to higher coherence of the continuations. Generally, this indicates that sampling-based methods are particularly responsive to the information density of the prompts.

Relation of bias congruency and continuation quality. Table 2 shows completion sensitivity results for each bias type, prompt condition and decoding procedure for continuations of the GPT-2 and mGPT models. Overall, the performance varies across different models, decoding procedures, and bias types. However, in general, the models are more likely to capture the object bias, as can be noted by the (almost) overall higher CS values for the ES verbs, aligning with results from, for example, Kementchedjhieva et al., 2021 and Zarrieß et al., 2022 which as well point towards a general tendency of LLMs to establish coreference to the object. Moreover, it is noticeable that for each decoding procedure the ability to capture the IC bias of SE verbs tends to improve when prompts are augmented with adverbial modifiers. Thus, it appears that the augmentation of the standard IC prompts indeed has an impact on the extent to which the LLMs can capture the IC bias. Interestingly, this influence varies depending on the decoding strategy at hand. Further, it is noteworthy that, in this way, the language models exhibit a different behavior in capturing the IC bias when confronted with the modified prompts compared to the findings of Bott and Solstad, 2021's human-produced data, where the IC bias was in fact eliminated by the same modification (see Section 2.1). That is, while humans tend to produce fewer bias-congruent continuations when the IC prompts are augmented with such adverbial modifications, the opposite seems to be true for language models.

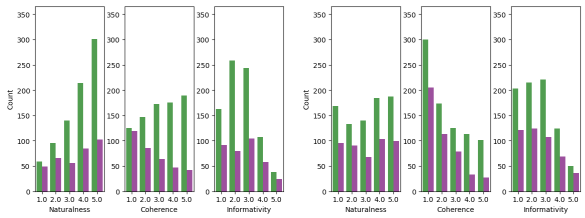
	GPT-2			mGPT		
	Diverse Beam Search	Nucleus Sampling	Typical Sampling	Diverse Beam Search	Nucleus Sampling	Typical Sampling
SE simple	62.5	25	75	50	25	62.5
SE modified	75	50	75	87.5	50	87.5
ES simple	50	75	87.5	75	87.5	87.5
ES modified	50	100	87.5	75	100	87.5

Table 2: Completion Sensitivity (CS) scores for each model, bias type, prompt condition, and decoding procedure, aggregated over all types of individual verbs. CS scores are calculated as the percentage of continuations where the predicted pronoun is congruent with the IC bias.

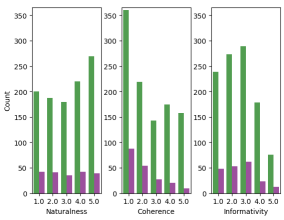
If we now consider the relation between bias congruency and continuation quality and examine the human-produced continuations in Table 1 first, we can find higher naturalness and coherence ratings for the bias-congruent continuations than for the bias-incongruent continuations, in particular for the standard IC prompts, which aligns with expectations based on Bott and Solstad, 2021. The observation that the bias-incongruent continuations of the modified prompts appear to be more acceptable than the bias-incongruent continuations of the standard prompts further aligns with the findings of Bott and Solstad, 2021, who demonstrated that modified prompts elicit different types of explanations, often referring to elements other than the

verbs' arguments.

Next, let us consider Figure 2, which depicts the relationship between bias congruency and the human evaluation criteria for each decoding method, aggregated across the two prompt conditions and across the generations of both LLMs (see Figure 12 in the Appendix for a visualiziation separated according to conditions). The higher green bars consistently observed across all prompt conditions and decoding methods indicate that each decoding method generated a greater number of bias-congruent continuations than bias-incongruent ones. Furthermore, for all three decoding methods, it is evident that non-bias congruent continuations receive lower ratings in terms of coherence, as indicated by the descending purple bars for this criterion. However, a preference for bias-congruent continuations being more natural than bias-incongruent continuations is primarily observed for the Diverse Beam Search decoding method. Further, it is interesting to note that although Typical Sampling tends to generate the most bias-congruent continuations, this does not necessarily result in better scores on the evaluation metrics. These results, thus, indicate that a bias congruent continuation does not equate to a qualitatively better continuation.



(a) Diverse Beam Search (b) Nucleus Sampling



(c) Typical Sampling

Figure 2: Relationship between bias congruency and the text continuation criteria for each decoding method, aggregated across SE verbs and ES verbs, across the two prompt conditions, and across the generations of both LLMs. Bias congruent continuations are depicted in green, bias incongruent continuations are depicted in purple.

Correlation between automatic and human evaluation. Table 1 reports the human ratings for the generated continuations, while Table 5 in the Appendix displays the automatic metrics. At first sight, it may seem that the automatic metrics generally align with our human evaluations, as, e.g., Diverse Beam Search consistently outperforms other decoding methods in automatic and human scores (with one exception). However, Table 3 shows Pearson's correlation coefficients between the automatic and human evaluation metrics and it becomes apparent that there is no significant correlation between ROUGE-L and BERTScore on the one and human ratings on the other hand. BLEU scores do achieve a weakly significant correlation with coherence ratings, but not with naturalness or informativity. Interestingly, there even seems to be a negative (but hardly significant) relationship between BLEU and informativity. Notably, these automatic metrics seem to fail even more miserably on our linguistically controlled task, as compared to correlations reported for downstream task evaluations as in, e.g., Savkov et al., 2022. We believe that this may be due to the fact that the scoring of differences between generated continuations in this rather restricted task may require a greater awareness of linguistic subtleties and deeper discourse understanding than what is currently captured by these metrics. Overall, these findings underscore the significant challenge faced by NLG metrics in accurately capturing the nuanced aspects of human evaluation and emphasize the need for cautious interpretation of automatic evaluation scores.

Human	**BLEU**	**ROUGE-L**	**BERTScore**
Naturalness	0.16 (p=0.03)	-0.02 (p=0.84)	-0.04 (p=0.59)
Coherence	0.18 (p=0.01)	0.03 (p=0.66)	-0.01 (p=0.91)
Informativity	-0.18 (p=0.02)	-0.08 (p=0.30)	-0.07 (p=0.35)

Table 3: Pearson's correlation coefficient between automatic and human evaluation metrics.

6 Conclusion

This paper investigated how continuations of IC prompts can be used to evaluate the text generation capabilities of language models, expanding prior work on discourse knowledge captured by LLMs in IC contexts by investigating their ability to generate not only bias-congruent but also sensible continuations. Our study reveals that LLMs face challenges in generating coherent continuations for relatively simple prompts, highlighting their struggle with discourse-level coherence. Further, our findings show that both the decoding method and the information density of the prompt have a substan-

tial impact on the quality of generated text, even surpassing the influence of the specific language model used. Our results further indicate that modifying the standard IC prompts has a notable effect on the LLMs' capability to capture the IC bias, depending on the decoding strategy employed. At the same time, the results indicate that a bias congruent continuation does not necessarily equate to a higher quality of the continuation. Intriguingly, we observe a surprisingly low correlation between automatic evaluation metrics and human judgments. This poses an interesting challenge for natural language generation, calling for future research to enhance the evaluation methodologies and metrics used in NLG systems. Another potential future direction for our study is to expand the scope beyond German data, as this is a major limitation of this work. While IC is known to be cross-linguistically stable, the inclusion of other languages in our investigation should be performed to validate our findings. Also, it would be interesting to more extensively investigate the models' strategy for choosing the first word of the continuation (i.e. the pronoun), which may simply consist in selecting the most recently mentioned discourse element. Further, it could be valuable to investigate the effects of additional prompt modifications on bias congruency and the quality of continuations. For this, one possible foundation could be the research conducted by Koornneef and Van Berkum (2006), for instance, where IC prompts are integrated within a larger pre-context, making them arguably a more natural option for evaluating LLMs compared to the prompts investigated in this study. Also, Hoek et al., 2021, for example, investigated IC in the context of relative clauses, which could provide another compelling starting point for further examination of LLMs in this context.

Funding: This research was funded by the Ministry of Culture and Science of the State of North Rhine-Westphalia, grant number NW21-059A (SAIL).

Supplementary Materials Availability Statement: Source code, prompts used for generating the models' continuations and the data from the human evaluation study are available from Github.[3]

[3]https://github.com/clause-bielefeld/implicit-causality-beyond-the-bias.

References

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Trans. Assoc. Comput. Linguist.*, 7:49–72.

Oliver Bott and Torgrim Solstad. 2014. From verbs to discourse: A novel account of implicit causality. In Barbara Hemforth, Barbara Mertins, and Cathrine Fabricius-Hansen, editors, *Psycholinguistic Approaches to Meaning and Understanding across Languages*, pages 213–251. Springer International Publishing, Cham.

Oliver Bott and Torgrim Solstad. 2021. Discourse expectations: explaining the implicit causality biases of verbs. *Linguist. Philos.*, 59(2):361–416.

Zhenguang G Cai, David A Haslett, Xufeng Duan, Shuqi Wang, and Martin J Pickering. 2023. Does ChatGPT resemble humans in language use?

Forrest Davis and Marten van Schijndel. 2020. Discourse structure interacts with reference but not syntax in neural language models. *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 396–407.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Trans. Assoc. Comput. Linguist.*, 8:34–48.

Evelyn C Ferstl, Alan Garnham, and Christina Manouilidou. 2011. Implicit causality bias in english: a corpus of 300 verbs. *Behav. Res. Methods*, 43(1):124–135.

Catherine Garvey and Alfonso Caramazza. 1974. Implicit causality in verbs. *Linguist. Inq.*, 5(3):459–464.

E. Goikoetxea, G. Pascual, and J. Acha. 2008. Normative study of the implicit causality of 100 interpersonal verbs in Spanish. *Behavior Research Methods*, 40:760–772.

Joshua K Hartshorne, Yasutada Sudo, and Miki Uruwashi. 2013. Are implicit causality pronoun resolution biases consistent across languages and cultures? *Exp. Psychol.*, 60(3):179–196.

Jet Hoek, Hannah Rohde, Jacqueline Evers-Vermeul, and Ted J M Sanders. 2021. Expectations from relative clauses: Real-time coherence updates in discourse processing. *Cognition*, 210:104581.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Hien Huynh, Tomas O Lentz, and Emiel van Miltenburg. 2022. Implicit causality in GPT-2: a case study.

T Florian Jaeger. 2010. Redundancy and reduction: speakers manage syntactic information density. *Cogn. Psychol.*, 61(1):23–62.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):1–38.

Andrew Kehler, Laura Kertz, Hannah Rohde, and Jeffrey L Elman. 2008. Coherence and coreference revisited. *J. Semant.*, 25(1):1–44.

Yova Kementchedjhieva, Mark Anderson, and Anders Søgaard. 2021. John praised Mary because _he_? implicit causality bias and its interaction with explicit cues in LMs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4859–4871, Online. Association for Computational Linguistics.

Arnout W Koornneef and Jos J A Van Berkum. 2006. On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye tracking. *J. Mem. Lang.*, 54(4):445–465.

Roger Levy and T Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. *Adv. Neural Inf. Process. Syst.*, 19.

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185, Online. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2022. Locally typical sampling.

Aleksandar Savkov, Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Anya Belz, and Ehud Reiter. 2022. Consultation checklists: Standardising the human evaluation of medical note generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 111–120, Abu Dhabi, UAE. Association for Computational Linguistics.

Hendrik Schuff, Lindsey Vanderlyn, Heike Adel, and Ngoc Thang Vu. 2023. How to do human evaluation: A brief introduction to user studies in NLP. *Natural Language Engineering*, pages 1–24.

Sebastian Schuster and Tal Linzen. 2022. When a sentence does not introduce a discourse entity, transformer-based models still sometimes refer to it.

Simeon Schüz, Ting Han, and Sina Zarrieß. 2021. Diversity as a by-product: Goal-oriented language generation leads to linguistic variation. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 411–422, Singapore and Online. Association for Computational Linguistics.

C E Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Torgrim Solstad and Oliver Bott. 2022. On the nature of implicit causality and consequentiality: the case of psychological verbs. *Language, Cognition and Neuroscience*, 37(10):1311–1340.

Shiva Upadhye, Leon Bergen, and Andrew Kehler. 2020. Predicting reference: What do language models learn about discourse models? *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 977–982.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Comput. Speech Lang.*, 67:101151.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. *AAAI*, 32(1).

Gian Wiher, Clara Meister, and Ryan Cotterell. 2022. On decoding strategies for neural text generators. *Trans. Assoc. Comput. Linguist.*, 10:997–1012.

Sina Zarrieß, Hannes Groener, Torgrim Solstad, and Oliver Bott. 2022. This isn't the bias you're looking for: Implicit causality, names and gender in German language models. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 129–134, Potsdam, Germany. KONVENS 2022 Organizers.

Sina Zarrieß, Henrik Voigt, and Simeon Schüz. 2021. Decoding methods in neural language generation: A survey. *Information*, 12(9):355.

A Appendix

Verb	Prompt Condition	Sentiment	Verb Type	Gender Order	Prompt
bewundern	simple prompt	positive	ES	f-m	Isabel bewunderte Paul, weil
bewundern	modified prompt	positive	ES	m-f	Paul bewunderte Isabel für ihre außerordentliche Geschicklichkeit beim diesjährigen Eisschaulaufen, weil
enttäuschen	names_simple	negative	SE	m-f	Björn enttäuschte Celina, weil
enttäuschen	names_pp	negative	SE	f-m	Celina enttäuschte Björn durch ihr unhöfliches Benehmen beim Geschäftsessen, weil
faszinieren	names_simple	positive	SE	f-m	Viktoria faszinierte Steven, weil
faszinieren	names_pp	positive	SE	m-f	Steven faszinierte Viktoria durch seine eindrücklichen Reiseberichte, weil
hassen	names_simple	negative	ES	m-f	Malte hasste Pia, weil
hassen	names_pp	negative	ES	f-m	Pia hasste Malte für die täglichen Sticheleien auf dem Schulhof, weil
inspirieren	names_simple	positive	SE	f-m	Clara inspirierte Vincent, weil
inspirieren	names_pp	positive	SE	m-f	Vincent inspirierte Clara durch seine innovative Vorlesung, weil
respektieren	names_simple	positive	ES	m-f	Martin respektierte Lina, weil
respektieren	names_pp	positive	ES	f-m	Lina respektierte Martin für seine couragierte Hilfe beim Löschen des Feuers, weil
schockieren	names_simple	negative	SE	f-m	Charlotte schockierte Jonathan, weil
schockieren	names_pp	negative	SE	m-f	Jonathan schockierte Charlotte durch sein aggressives Auftreten in der Talkshow, weil
verabscheuen	names_simple	negative	ES	m-f	Stefan verabscheute Miriam, weil
verabscheuen	names_pp	negative	ES	f-m	Miriam verabscheute Stefan für seine unnötige Hektik bei der Abfertigung von Patienten, weil

Table 4: Verbs used in the study and examples for each prompt condition.

Instruktionen

In dieser Studie besteht Ihre Aufgabe darin, **Satzfortsetzungen, die von Computermodellen generiert wurden, zu beurteilen.**

Alle folgenden Aufgaben haben die gleiche Form. Sie sehen in der ersten Zeile den Anfang eines Satzes und in der zweiten Zeile eine von einem Computermodell generierte **Fortsetzung, die eine Begründung für die Aussage im Satzanfang liefern soll**. Ihre Aufgabe ist es, anhand von drei verschiedenen Aussagen zu beurteilen, inwiefern es Computermodellen gelingt, gute mögliche Begründungen für die Aussagen in den Satzanfängen zu generieren.

Wenn Sie zum Beispiel den folgenden **Satzanfang** sehen:

Satzanfang:
Hannah amüsierte Anton, weil

wäre eine mögliche **Begründung**:

Begründung:
sie die besten Witze erzählte.

In jedem Durchgang sehen Sie dann drei Aussagen über die gerade angezeigte Begründung.
Bei diesen drei Aussagen geben Sie bitte jeweils an, ob bzw. wie sehr sie dieser zustimmen oder nicht zustimmen. Dafür wählen Sie bitte jeweils ein Feld in der 5-Punkte-Skala aus, aufsteigend von **1 (stimme der Aussage gar nicht zu)** bis **5 (stimme der Aussage voll zu)**:

Die Begründung wirkt **natürlich** und liest sich so, als ob sie von einer/m deutschen Muttersprachler/in geschrieben wurde.

stimme gar nicht zu ○ 1 ○ 2 ○ 3 ○ 4 ◉ 5 stimme voll zu

Die Begründung ist **sinnvoll**, es gibt einen logischen Zusammenhang zwischen dem Satzanfang und der Fortsetzung.

stimme gar nicht zu ○ 1 ○ 2 ○ 3 ○ 4 ◉ 5 stimme voll zu

Die Begründung ist **überraschend**, dadurch könnte der Satz insgesamt ein interessanter Beginn einer Geschichte sein.

stimme gar nicht zu ○ 1 ◉ 2 ○ 3 ○ 4 ○ 5 stimme voll zu

In diesem Fall wäre es wohl natürlich, auf den ersten beiden Skalen eine Position weiter rechts auszuwählen, denn die Fortsetzung liest sich flüssig und bietet eine sinnvolle Erklärung für die Aussage im Satzanfang. Auf der dritten Skala würde man vielmehr eine Position weiter links auswählen, da die generierte Erklärung für diesen Satzanfang eher erwartbar und damit weniger überraschend ist.

Auf der nächsten Seite bekommen Sie Gelegenheit, sich mit der Aufgabe vertraut zu machen, bevor das eigentliche Experiment anfängt.

Wichtig: Jeder Durchgang im Experiment steht für sich allein und Ihre Urteile sollten sich deshalb stets nur auf den gerade vorliegenden Satz und die gerade vorliegende Begründung beziehen.

Nachdem Sie ein Urteil gefällt haben, bestätigen Sie die Eingabe bitte mit dem "Weiter"-Knopf. Bitte benutzen Sie im Experiment **NIE** die "Zurück"-Taste im Browser, da dies zum sofortigen Abbruch des Versuchs führen kann.

Sind Sie bereit für zwei Übungsdurchgänge?

Weiter

Figure 3: Instructions used in the online experiment. Participants were given an example items as well as an explanation of what would have been a reasonable rating on each of the three evaluation criteria (naturalness, coherence and informativity).

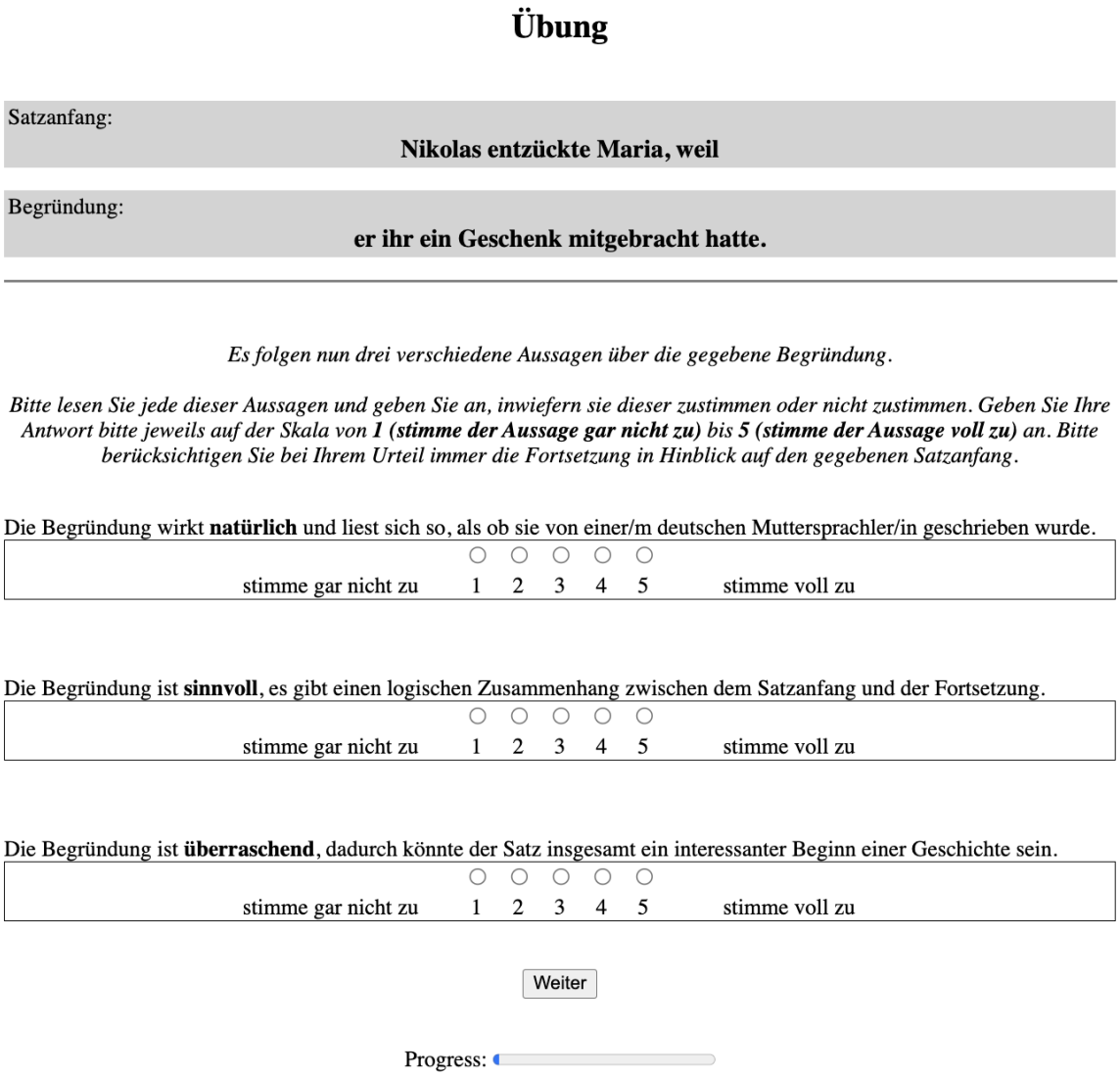
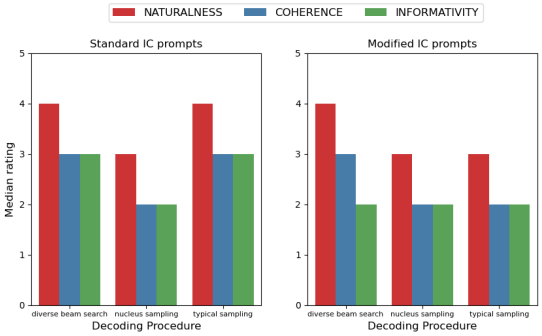


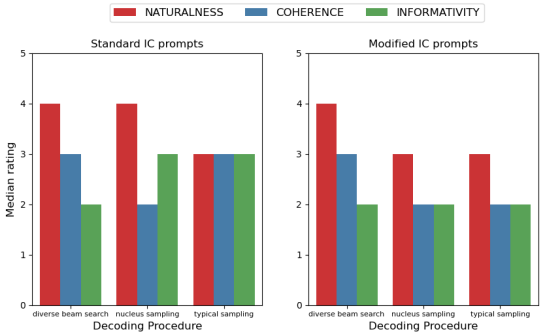
Figure 4: One of two training items used in the online experiment for the participants to get familiar with the task and the rating scales.

	GPT2			mGPT		
	BLEU	ROUGE-L	BERTScore	BLEU	ROUGE-L	BERTScore
Standard IC Prompt						
Diverse Beam Search	0.47	0.074	**0.592**	**0.565**	**0.089**	0.544
Nucleus Sampling	0.485	0.062	0.539	0.428	0.069	0.551
Typical Sampling	0.346	0.063	0.569	0.383	0.072	0.577
Modified IC Prompt						
Diverse Beam Search	0.432	0.056	0.588	**0.472**	**0.076**	0.587
Nucleus Sampling	0.377	0.061	0.578	0.429	0.064	**0.59**
Typical Sampling	0.404	0.059	0.612	0.37	0.056	0.58

Table 5: Automatic evaluation results for each decoding procedure and each IC prompt condition. Bold values indicate conditions with the best values for that metric.



(a) Median ratings for continuations of GPT-2 model.



(b) Median ratings for continuations of mGPT model.

Figure 5: For each model, median ratings for each decoding procedure, each prompt condition and each text evaluation criteria of the human evaluation.

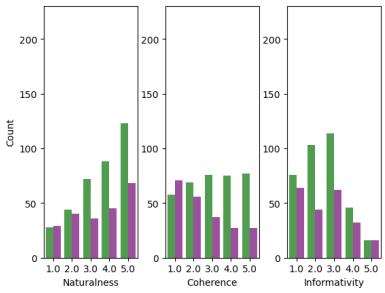


Figure 6: Diverse Beam Search & Simple IC prompts

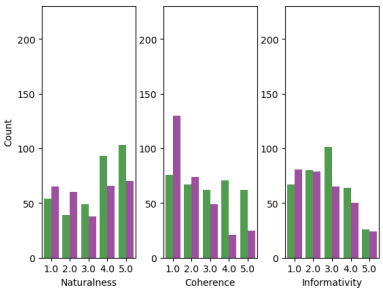


Figure 7: Nucleus Sampling & Simple IC prompts

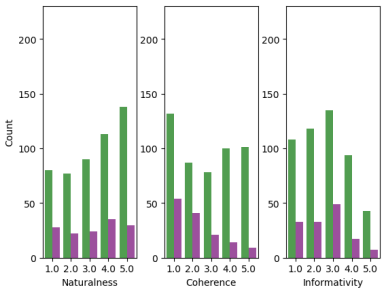


Figure 8: Typical Sampling & Simple IC prompts

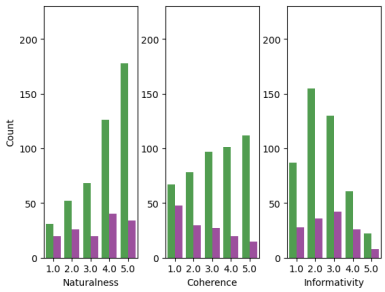


Figure 9: Diverse Beam Search & Modified IC prompts

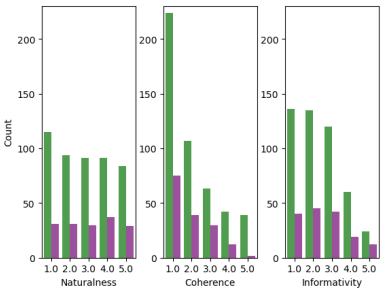


Figure 10: Nucleus Sampling & Modified IC prompts

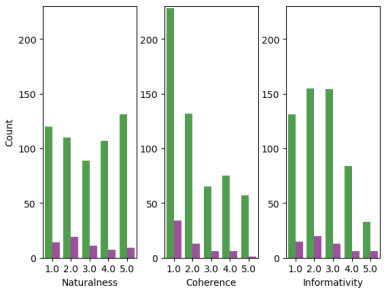


Figure 11: Typical Sampling & Modified IC prompts

Figure 12: Relationship between bias congruency and the text continuation criteria for each decoding method and for each prompt condition, aggregated across SE verbs and ES verbs and across the generations of both LLMs. Bias congruent continuations are depicted in green, bias incongruent continuations are depicted in purple.