UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Instruction tuning modulates discourse biases in language models

Permalink

https://escholarship.org/uc/item/5g02g0g8

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors

Kankowski, Florian
Solstad, Torgrim
Zarrie§, Sina
et al.

Publication Date

2025

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Instruction tuning modulates discourse biases in language models

Florian Kankowski* **(florian.kankowski@uni-bielefeld.de)**

Torgrim Solstad* **(torgrim.solstad@uni-bielefeld.de)**

Sina Zarrieß* **(sina.zarriess@uni-bielefeld.de)**

Oliver Bott* **(oliver.bott@uni-bielefeld.de)**
*Bielefeld University and CRC 1646

Abstract

Instruction tuning (IT) has been a fruitful technique for aligning Large Language Models with human preferences. However, the linguistic implications of IT remain unclear. In two experiments on coreference and coherence biases in the context of Implicit Causality, we investigate how IT modulates these discourse biases in relation to model size. Our results show that IT interacts with model size – instruction-tuned models display enhanced coherence biases and more human-like coreference patterns, sometimes exceeding human performance. However, this effect appears size-dependent, suggesting that IT causes some linguistic patterns to emerge that are dormant in the respective foundation models.

Keywords: Large language models (LLM); Instruction tuning; Implicit causality; Coreference; Discourse coherence

Introduction

As the capabilities of Large Language Models (LLMs) grow more advanced, the study of their abilities has shifted. While earlier research focused on their general ability to encode and express the rules of language (Blevins et al., 2018; Gulordava et al., 2018; Tenney et al., 2019), this field of study has more recently taken a back seat in favour of benchmarking models' logical and mathematical prowess (Lewis & Mitchell, 2024; Yuan et al., 2023), their cognitive abilities like theory of mind (Verma et al., 2024; Wu et al., 2024) and their inferential capabilities (Asher et al., 2023; Mahowald et al., 2024), investigating to which degree LLMs are not only models of human language, but of human cognition at large.

One fruitful area of linguistic research that has remained popular with the rising dominance of generative GPT-like LLMs is to adapt psychological and psycholinguistic experiments to study the (linguistic) behaviour of language models (Binz & Schulz, 2023; Ettinger, 2020; Hawkins et al., 2020). Among these, the prediction-based phenomenon of Implicit Causality (IC; e.g., Garnham, Child, & Hutton, 2020) has often been used to gauge the similarity of discourse behaviour between humans and LLMs. Given its predictive nature, IC is arguably of particular relevance for comparing the predictive capabilities of humans (Clark, 2013) and LLMs.

In this paper, we enrich existing research by identifying two gaps present in the study of Implicit Causality in LLMs: the effect of Instruction Tuning versus model size and the interplay of different discourse biases associated with Implicit Causality, as a more adequate measure of complex human behaviour. We ran two experiments across a selected range of models to investigate these gaps.

Background

Implicit Causality

IC, which has been widely researched in psycholinguistics, comprises two highly consistent and well-attested discourse biases associated with interpersonal verbs like *fascinate* and *admire*. The two biases are typically established in production studies prompting continuations for sentences as in (1):

(1) a. Mary fascinated Peter. ✍
(e.g., *She always had the most brilliant ideas.*)
b. Mary admired Peter. ✍
(e.g., *He was a great dancer*)

For one, IC verbs display a strong **coherence bias** towards explanations over all other discourse relations as a whole ($\approx$ 60% according to Kehler et al., 2008; Solstad & Bott, 2022, see also the bar chart in the upper left corner of Figure 3). That is, participants prefer a continuation as in (1a) to a result relation like ✍ *He decided to ask her out*. Thus, providing an explanation constitutes the basic discourse strategy for these verbs. Furthermore, IC verbs show strong **coreference biases** towards one of the arguments over the other in explanations and consequences. More specifically, stimulus-experiencer verbs like *fascinate* are biased towards the stimulus subject *Mary* when the prompt in (1a) is followed by the causal connective *because* – characterized as *Implicit Causality bias* – and equally strongly towards the experiencer object *Peter* when followed by *and so* – so-called *Implicit Consequentiality bias* (e.g., Crinean & Garnham, 2006; Garnham, Vorthmann, & Kaplanova, 2020; Solstad & Bott, 2022). This distribution is reversed for the two relations for experiencer-stimulus verbs like *admire* (Fig. 1).

Solstad and Bott (2022) proposed to account for this mirror-like pattern induced by discourse relations with recourse to two mechanisms: First of all, IC verbs introduce an explanatory *empty slot* which is associated with the stimulus argument for *fascinate*- and *admire*-type verbs. This slot is preferably filled by providing an explanation about the stimulus when possible, for both implicit and explicit explanations. If this strategy is not available, as in consequences after *and so*, one reverts to a general *contiguity principle* (Murray, 1997; Stevenson et al., 1994), specifying a subsequent eventuality associated with the experiencer argument.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

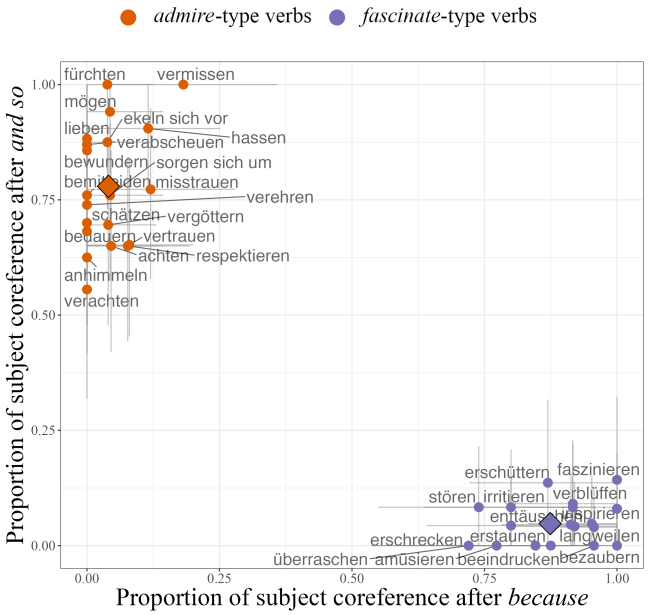


Figure 1: Implicit causality (*x* axis) and Implicit consequentiality (*y* axis) coreference biases for 20 *fascinate*- and 20 *admire*-type verbs in German (based on data from Solstad & Bott, 2022). *fascinate*-type verbs (in blue) have a strong subject bias after *because* (*x* coordinate) and a strong object bias after *and so* (*y* coordinate). *admire*-type verbs (in red) show a reversed pattern. Diamonds represent verb-type averages.

Based on the above, the coreference bias can be said to depend on the coherence bias in human sentence processing (e.g., Ehrlich, 1980; Kehler et al., 2008). We therefore contend that IC is particularly well-suited to test discourse-pragmatic abilities of LLMs, closing the discourse ability gap in previous research (see below). In particular, assessing both coreference and coherence biases in LLMs is a useful tool in determining the influence of lexical semantic and more general discourse-pragmatic influences on their linguistic performance (Kankowski et al., 2025).

Instruction Tuning

Instruction Tuning (IT; also known as Instruction Finetuning) has emerged as a crucial technique for improving the capabilities of LLMs in conversational and instruction-following settings.

IT refers to the process of fine-tuning a pre-trained language model to better follow natural language instructions and align with human preferences. This typically involves additional training on instruction-response pairs, often through supervised fine-tuning (SFT, Wei et al., 2021) or reinforcement learning from human feedback (RLHF, Christiano et al., 2017). While the primary goal of IT is to improve models' ability to follow instructions and generate more helpful and appropriate responses, it can also be used to inject specific world knowledge or skills into a model, like function-calling or coding.

The impact of IT on model behaviour is complex – while some studies suggest that basic RLHF can actually decrease performance on traditional NLP tasks (Ouyang et al., 2022), Wei et al. (2021) have found that targeted instruction finetuning using NLP tasks can enhance related capabilities on hold-out tasks. However, systematic investigation of IT's effects on linguistic behaviour has been limited, in part because popular instruction-tuned models like ChatGPT cannot be directly compared to their foundation models.

Implicit causality in LLMs

Since Implicit Causality is a very well investigated phenomenon in psycholinguistics that allows for proper experimental control, it serves as a suitable test bed for analysing the discourse capabilities of language models. As such, many studies investigated whether LLMs express an IC bias (Huynh et al., 2022; Upadhye et al., 2020, among others) or how IC effects are modulated by context or linguistic constraints (Davis & van Schijndel, 2021; Sieker et al., 2023; Zarrieß et al., 2022). More recently, Cai et al. (2024) used IC in a larger benchmark spanning multiple linguistic disciplines and Kankowski et al. (2025) established a benchmark of different IC-related discourse biases to gauge the discourse behaviour of LLMs – which has also been used in parts for this study.

Although existing research covers a lot of ground, we have identified two research gaps that we aim to fill with this paper:

discourse ability gap : Existing studies of Implicit Causality effects in language models focus almost exclusively on its coreference bias (Cai et al., 2024; Davis & van Schijndel, 2020; Sieker et al., 2023; Upadhye et al., 2020; Zarrieß et al., 2022). This, however, does not paint a full picture of the IC effects, as the **coreference** and **coherence** biases strongly interact in human processing. While Kankowski et al. (2025) compared four different IC biases in LLMs, we restrict ourselves to two experiments, investigating the coreference and coherence biases, since these two are unambiguously agreed upon in the psycholinguistic community and thus present an uncontroversial test bed for comparison.

model type gap : Older studies on smaller language models have consistently observed non-existent or weak IC effects (Upadhye et al., 2020; Zarrieß et al., 2022), with models not being sensitive to contextual modulations typically affecting IC bias in humans (Sieker et al., 2023). This stands in contrast to the strong effects reported by Cai et al. (2024), who found that both ChatGPT 3.5 (OpenAI, 2022) and the considerably smaller Vicuna (Chiang et al., 2023) not only match, but exceed human IC coreference biases. These models differ in that they are larger in parameter count and underwent Instruction Tuning. We investigate whether this large gap in model performance should be attributed to model size, finetuning or a combination of both, by conducting our experiments on a range of models, spanning different model sizes, both with and without Instruction Tuning.

Experimental setup

Model selection

In order to investigate the model type gap sketched out above, models were compared with regard to two dimensions: model size and instruction tuning. For this purpose, we selected model pairs that are available as both *foundation* (henceforth: F) models, meaning that they were only trained for general next token prediction using non-domain-specific corpora, as well as *instruction-tuned* (henceforth: I) models. This way, the effect of pretraining is minimized, as all I models underwent the exact same pretraining stages as their F counterparts.

We chose three model pairs from the Llama 3 family (Grattafiori et al., 2024), namely Llama 3.2 1B, 3.2 3B and 3.1 8B. Although the models span two sub-generations, their weights are deeply related, because the smaller 1B and 3B models are derived from Llama 3.1 through pruning and knowledge distillation. IT was performed through iterative rounds of SFT and Reinforcement Learning, including human-annotated preference data as well as synthetic data, comprising general reasoning, instruction-following and skill-specific dialogical texts (instruction-response pairs).

We also included Mistral-NeMo 12B (Mistral AI & NVIDIA, 2024, henceforth: Mistral 12B) in our analysis as a model from a different model family and with a parameter count of over 10B. Like Llama 3, Mistral has both foundation and instruct versions, but there does not exist any official information by the authors on how the IT was achieved.[1]

In total, we included 8 models (4 F, 4 I), all with official German support. Model inference was performed in 16 Bit precision on an NVIDIA A100 GPU.

Data generation

Our approach differs from previous studies of IC effects in LLMs in that we let the models generate full sentence completions after the prompt instead of measuring relative token output probability. This is necessary for capturing anaphoric coreferences not immediately realized after the connective or for excluding non-grammatical generations – which is typically done in human studies. It also allows us to further analyse the generations in the future.

Full sentence generations require a decoding method, which can greatly impact model behaviour (Holtzman et al., 2019) and should thus be chosen carefully. In many LLM applications, a variation of stochastic sampling is used, but the optimal decoding strategy is heavily task-dependent (Shi et al., 2024). In the context of Implicit Causality, Sieker et al. (2023) observed that diverse beam search (Vijayakumar et al., 2016) aligns better with human biases. Following their results, we use a diverse beam search decoding with a beam size and beam group size = 10 and a diversity penalty $\lambda = .6$.

For foundation models, we do not include any context in the prompts, as pure sentence continuation is the "natural task" for foundation models. Instruction models receive a short system prompt instructing them to continue the sentence presented to them by the user.[2] We provide no further context about the experiment or instructions on the form or content of the continuation. The experimental items are then presented as the user prompts.

The model generations were semi-automatically annotated using spaCy (Honnibal & Montani, 2017) to label the data for coreference (Exp. 1) and discourse relation (Exp. 2).[3]

Experiment 1: Coreference Bias

Following the design in Solstad and Bott (2022) and Kankowski et al. (2025), we investigated coreference biases in German, manipulating the factors VERB TYPE (*fascinate*-type vs. *admire*-type verb; between items) and CONNECTIVE (*weil* 'because' vs. *sodass* 'and so'; within items). We also included GENDER ORDER of the proper names as a counterbalancing factor within items:

(2) a. Lea/Ted fascinated Ted/Lea *because/and so* ... ✍
b. Mia/Max admired Max/Mia *because/and so* ... ✍

Testing a total of 40 items with 20 verbs of each verb type, and coding the subject or object coreference of any referential expression occurring in participants' continuations, Solstad and Bott (2022) found the two verb types to form two distinct clusters (see Figure 1). While *fascinate*-type verbs have a bias towards the subject after *because* and towards the object after *and so*, *admire*-type verbs displayed the opposite patterns. Statistical analysis revealed an almost perfect negative correlation between the two bias patterns (Pearson's $r = 0.94$). Logit mixed-effects regression models predicting coreference with the subject argument as a function of *verb type* and *connective* correspondingly revealed a strong crossover interaction for these factors. In assessing the LLMs, we will focus on this interaction.

Methods 19 (instead of 20)[4] verbs of each VERB TYPE were used along with 40 female and 40 male first names chosen such that the LLMs would process the referents as unambiguously male or female. The names were paired and all pairs were tested in four conditions (CONNECTIVE × GENDER ORDER) for the 38 verbs, resulting in a total of 6,080 continuations for each LLM.

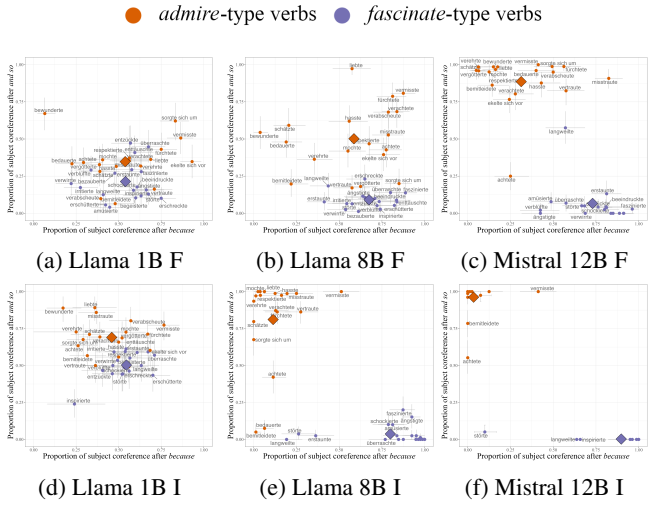
Annotation The data annotation ensured that every continuation submitted to statistical analysis could be assigned a syntactic parse. No other criteria of acceptability were ap-

[1] It should be noted that this model also differs from the Llama models in both architecture and training data, which should be taken into account when comparing experimental results.

[2] The system prompt also includes instructions to not produce anything else in their output, like greetings or comments, and to not repeat or alter the user prompt. This is done to ensure that the automatic annotation parses the continuation correctly.

[3] All code and generated data as well as annotationan guidelines are available in the accompanying OSF archive: https://osf.io/m6fgv/

[4] One verb of each type were excluded from our experiment, as Solstad and Bott (2022) found them to display unexpectedly low biases.



(a) Llama 1B F (b) Llama 8B F (c) Mistral 12B F

(d) Llama 1B I (e) Llama 8B I (f) Mistral 12B I

Figure 2: Implicit causality and consequentiality biases of individual verbs for selected LLMs. See caption of Figure 1 for details. Abbreviations: F = foundation, I = instruction.

plied. Annotation identified the subject in the continuation, allowing for the identification of anaphoric coreference. Filtering out continuations with no suitable referring expression led to the removal of 1–15% of the data. Inter-annotator agreement with the manual annotation for the data in Solstad and Bott (2022) was Cohen's $k = .85$ for coreference labeling.

Statistical analysis Logit mixed-effects binomial logistic regression models were fitted for each LLM with the R package lme4 (Bates et al., 2015, see the OSF archive for details), predicting *coreference to the subject referent* as a function of VERB TYPE and CONNECTIVE (and GENDER ORDER), including by-item random slopes for GENDER ORDER and CONNECTIVE and their interaction in addition to a random intercept for items. This maximal model was simplified in line with the findings in Solstad and Bott (2022). Model improvement was assessed by performing likelihood ratio tests of models with and without the effect in question. Of importance in the context of this study, the model involving the interaction VERB TYPE × CONNECTIVE TYPE was compared with a model with these factors as simple effects only. If this interaction turned out to be significant, the data were subsetted according to CONNECTIVE (*because* vs. *and so*), comparing a model with VERB TYPE with an "intercept only" model to assess the presence of verb type differences for the connectives individually.

Results and Discussion Figure 2 shows the results of three representative LLMs in their foundation and instruction versions. Across LLMs, performance improved with increasing size. Among the **foundation models**, only the smallest LLM (Llama 1B, Figure 2a) failed to differentiate between the two connectives and verb types altogether. Still, only Mistral 12B F (Figure 2c) matched the negative correlation between the biases after *because* and *and so* found in human data (see Table 1 for details). The other foundation LLMs only showed significant effects for the Implicit Consequentiality bias (after *and so*, see distance between verb-type average diamonds on *y* axis) and numerically stronger biases for *fascinate*-type verbs (towards the object; range: 81.4–92.3%) than for *admire*-type verbs (towards the subject; range: 51–88.6%).

Interestingly, **instruction tuning** improved performance: All LLMs but Llama 1B displayed a moderate negative correlation between Implicit Causality and Implicit Consequentialty (see Table 1). These LLMs also displayed significant differences between verb types for both connectives (as opposed to the situation for foundation LLMs). Remarkably, after *and so* subject biases for *admire*-type verbs (range: 81.1–96.2%) were even stronger than in humans (77.7%). For *because* biases, only Mistral 12B I matched (and slightly surpassed) human biases. In sum, the largest model, Mistral 12B I, was again closest to the human data (Table 1), even outperforming humans on average with very low in-class variance (compare Figures 1 and 2f). Thus, for *admire*-type verbs, Mistral 12B I displayed an object bias after *because* of 96.2% (humans: 94.2%) and a subject bias after *and so* of 96.2% (humans: 77.7%).[5]

Regarding the two research gaps, we see that the best-performing model, Mistral 12B, closes the model type gap to ChatGPT and Vicuna reported in Cai et al. (2024). By including the dimension of consequentiality (previously only addressed by Kankowski et al., 2025), this experiment also contributes towards closing the discourse ability gap.

In light of the two-mechanism account in Solstad and Bott (2022), the overall better performance for Implicit Consequentiality biases (after *sodass* 'and so') may be taken as evidence that LLMs are better at mimicking general discourse-pragmatic principles like the *contiguity principle* than discourse biases driven by more fine-grained lexical-semantic properties, as for Implicit Causality (after *weil* 'because'). There could also be other reasons for this difference, however. For example, explanations are semantically more diverse than result relations (e.g., Sweetser, 1990) and also display more variation with regard to their syntactic position relative to the effect clause, making it harder to recognize patterns. It may also be added that *weil* 'because' is not always the most natural connective occurring with these verbs (similar observations can be made for English).

Experiment 2: Coherence Bias

LLMs' coherence biases were investigated using prompts with commas instead of connectives, again manipulating the factor VERB TYPE and counterbalancing for GENDER ORDER (see Experiment 1):

(3) a. Lea/Ted fascinated Ted/Lea, …
b. Mia/Max admired Max/Mia, …

[5]It should be noted, however, that even in the best-performing LLMs, a number of verbs do not pattern consistently within the verb types established in psycholinguistic investigations of human sentence production (see Figures 1 and 2f).

Data	Correl. Biases	V.TYPE × CONNECT.	V.TYPE *because*	V.TYPE *and so*
	r	$\chi^2(1)$	$\chi^2(1) =$	$\chi^2(1)$
HUMAN	✓ −.94	✓ 1161.3	✓ 681.3	✓ 487.8
LLama 1B F	✗ .07	✗ .5	✗ —	✗ —
LLama 3B F	✗ .09	✓ 7.8	✗ 1.8	✓ 21.6
Llama 8B F	✗ −.08	✓ 27.2	✗ 1.1	✓ 39.7
Mistral 12B F	✓ −.73	✓ 58.6	✓ 25.2	✓ 60.2
LLama 1B I	✗ −.16	✓ 11.9	✗ .3	✓ 21.2
Llama 3B I	✓ −.72	✓ 52.8	✓ 27.0	✓ 54.4
Llama 8B I	✓ −.74	✓ 59.8	✓ 55.2	✓ 39.4
Mistral 12B I	✓ −.92	✓ 78.8	✓ 59.0	✓ 77.7

Table 1: LLM coreference bias performance as compared to the human data. Columns (f.l.t.r.): i) Correlation between *because* and *and so* biases (Pearson's r), ii) Interaction VERB TYPE × CONNECTIVE, iii) Main effect of VERB TYPE for *because* and *and so* continuations. Values prefixed with ✓ indicate a significant result in the direction of the human data, ✗ an insignificant result. See OSF archive for details.

While the type of discourse relation is usually assessed by human annotators using an insertion test checking which connective is most suitable, state of the art transformer language models make surprisingly many errors when inferring appropriate discourse connectives (Chang & Bergen, 2024). We reasoned that a comma would lead to a significant production of subordinate clauses with clause-initial connectives that we could use for automatic coding of the discourse relation based on connectives' semantics. For comparison, we used the results for the human data in Kankowski et al. (2025), who found participants to display an even stronger coherence bias of ≈ 80% for both *fascinate*- and *admire*-type verbs after a comma, with no significant difference between verb types. For the assessment of LLMs' performance against human data, however, we used a less strict measure, investigating whether explanations make out significantly more than 50% of discourse relations (as in Solstad & Bott, 2022, see "Statistical analysis" below).

Methods Testing the forty pairs of names from Experiment 1 for both GENDER ORDER conditions with the 38 verbs of both VERB TYPES resulted in 3,040 continuations per LLM.

Annotation After filtering out continuations that could not be assigned a syntactic parse and those for which a discourse relation could not be coded (removing 8–51% of data, such as main clauses and relative clauses), we used DiMLex (Stede & Umbach, 1998) to map the explicit discourse connectives to their discourse relation, relabelling relations to fit the coding scheme used in Solstad and Bott (2022). Excluding main clause continuations, a comparison of automatic and manual annotation of the data yielded a very good inter-annotator agreement (Cohen's $\kappa = .86$) for the binary label explanation vs. non-explanation, which is what is assessed below.

Statistical analysis The statistical analysis followed the same principled approach as described for Experiment 1. Ignoring the counter-balancing factor GENDER ORDER, we conducted logit mixed-effects regression analyses modelling the *likelihood to produce an explanation* as a function of VERB TYPE including by-item random intercepts. As in Solstad and Bott (2022), we confirmed that explanations made up more than half of the relations by investigating the intercept, which should differ significantly from zero with a positive estimate. We also investigated effects of VERB TYPE by comparing the just-mentioned model with an "intercept-only" model.

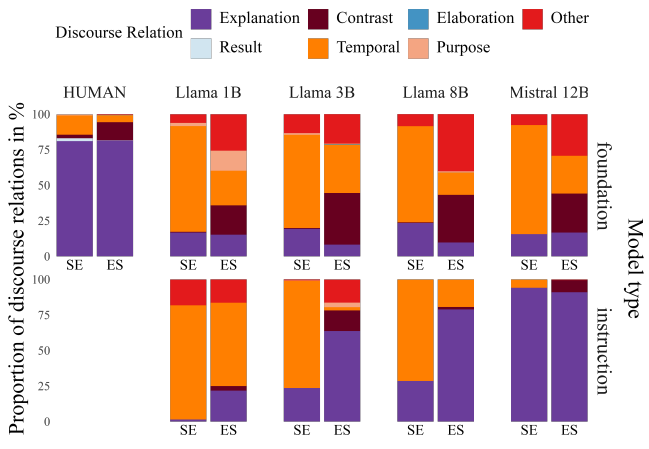


Figure 3: Coherence biases of *fascinate*- (SE) and *admire*-type (ES) verbs in humans (upper left corner) and LLMs (upper row: foundation, lower row: instruction-tuned models).

Results and discussion The proportions of discourse relations for the two verb types are plotted for human (upper left bar chart) and LLM data in Figure 3. LLMs produced a smaller range of different connectives (8–12) than humans (15), with a trend towards fewer discourse relations in the larger instruction models (e.g., three relations in LLama 8B I). As evidenced by Figure 3, foundation models produce far fewer explanations (15-17% across verb types) than humans (81%). For *fascinate*-type verbs, temporal relations (*als* 'as'/'when' oder *während* 'while') were clearly dominant, while the picture was more mixed for *admire*-type verbs, where Contrast (*aber* 'but') and Other (mostly *und* 'and') occurred frequently. Interestingly, apart from the smallest LLM, LLama 1B I (12% explanations), instruction-tuning caused a significant increase in the production of explanations (other LLMs: 41–93%). However, in most cases, the increase affected only *admire*-type verbs, Mistral 12B I being the only exception. This LLM even produced a larger proportion of explanations than humans (93% across verb types).

Importantly, this is, as far as we know, the first time a

model has shown a coherence IC bias in line with humans. None of the models analyzed by Kankowski et al. (2025) showed any such preference towards explanation relations.

Again, the fact that only one instruction LLM, Mistral 12B I comes close to capturing the discourse coherence bias consistently, can be taken as evidence that LLMs do not grasp discourse-pragmatic principles induced by lexical semantics as well as more general discourse-pragmatic principles. However, the non-explanatory discourse relations observed for the LLMs also do not correspond with those found for human participants in Solstad and Bott (2023) for verbs that lack a verb-based *empty slot*. In particular, LLMs did not produce any result relations whatsoever.

General Discussion

We set out to answer two main questions: (1) What is the influence of IT on LLM IC biases and how does it interact with model size? (2) How do the coherence and coreference biases – tightly linked in human processing – bear out in LLMs?

The results of both experiments highlight that IT significantly strengthens discourse biases, particularly for coherence relations, while a larger model size with only pretraining is insufficient to fully replicate human-like IC patterns.

A possible explanation for why foundation models struggle with IC, especially with coherence and coreference in causal (*because*-prompted) conditions is that these patterns are rare in standard training corpora. Asr and Demberg (2012) found that causal relations often remain implicit in corpus data, making them harder to acquire just by learning surface-level patterns, requiring instead a deeper, possibly extralinguistic understanding of causality.[6] In contrast, under the two-mechanism account of Solstad and Bott (2022), Implicit Consequentiality emerges from a more general contiguity principle that not only applies to IC contexts, making it easier for the models to pick up. In particular, this dichotomy maps onto the idea of functional vs. formal linguistic knowledge as described by Mahowald et al. (2024), whereby Implicit Causality requires the former, while Implicit Consequentiality – like the rules of grammar – corresponds to the latter.

On the other hand, it is likely that the dialogical and conversational data that makes up the majority of the finetuning data includes proportionally more explicit causal relations, making it easier for the models to pick up (possibly superficial) IC biases from surface-level correlations. On top of that, IT encourages models to adhere to human preferences (Ouyang et al., 2022), which include any subconscious biases like discourse effects from IC, resulting in a stronger learned bias, as datapoints with bias-incongruent contents would, for example, more likely be rated dispreferred in RLHF.

The influence of IT is especially evident for the coherence bias, which shows a very strong shift under IT. In fact, the foundation models hardly differ at all in terms of their proportion of explanations. A likely explanation is that instruction-tuned models are partially trained on datasets designed for reasoning and structured communication, which primes the models towards constructing causal relations. In contrast, foundation models tend to "continue the story" presented to them, resulting in a larger proportion of temporal and conjunction relations, especially given the fact that the prompt is presented without any context to refer back to.

Finally, our results indicate that IT alone is insufficient to develop robust IC biases. The failure of Llama 1B to acquire meaningful IC patterns suggests that IT may "teach" models the relevant biases for the two verb types, but does not instill any linguistic knowledge into the models. Thus, a smaller model, like Llama 1B F, which does not show any significant coreference bias and thus likely does not encode verb class internally, does not acquire any bias during IT. Importantly, our results show that IT cannot compensate for the absence of these fundamental representations. This suggests a two-stage developmental process: first, models must acquire a representation of verb classes through pretraining, and second, IT can refine and amplify these discourse biases, aligning the model behaviour more with human preferences.

This is especially evident in the results of Experiment 2, where despite the similarity of distributions for foundational models, their IT counterparts differ immensely in terms of coherence bias, which can only be explained by the notion that IT only awakens causal tendencies that the model has already learned to understand. These observations are in line with the arguments of Zhou et al. (2023), who postulate that knowledge is mostly acquired during pretraining, with post-training largely only able to modulate model behaviour.

Conclusions

In this study, we examined the effects of instruction tuning (IT) on the discourse biases of LLMs, comparing them to the effects of raw parameter count. We addressed two gaps in the current research, showing that IT strongly modulates model behaviour, even for smaller models, and strengthened discourse effects already present in larger foundation models, closing the gap towards matching human preferences. We also provided data in multiple dimensions of IC, namely coherence biases and causal/consequential coreference biases.

Future work should investigate whether similar effects of IT are observed for other discourse phenomena beyond IC, or the consistency of IC effects across languages and verb classes (like *agent-patient* or *transfer-of-possession* verbs), for which human comparison data exists.

Analyzing the instruction data used for tuning could help determine whether these improvements stem from explicit discourse-related instructions or emerge more indirectly. To this end, models trained on open-source data like the OpenAssistant dataset (OpenAssistant, 2023) could be leveraged.

[6] In general, this question cannot be answered for any specific LLM without analyzing its training corpus. However, the exceedingly better performance of Mistral compared to the Llama models could be an indicator for the influence of not only model size, but also of distributional difference of pretraining data for the realization of the observed discourse effects.

Acknowledgements This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – CRC 1646, project number 512393437, subprojects Ö & B01.

References

Asher, N., Bhar, S., Chaturvedi, A., Hunter, J., & Paul, S. (2023, June). *Limits for Learning with Language Models* (No. arXiv:2306.12213). arXiv. doi: 10.48550/arXiv.2306.12213

Asr, F. T., & Demberg, V. (2012). Implicitness of discourse relations. In M. Kay & C. Boitet (Eds.), *Proceedings of coling 2012* (p. 2669-2684). Mumbai: The COLING 2012 Organizing Committee.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Binz, M., & Schulz, E. (2023, February). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120. doi: 10.1073/pnas.2218523120

Blevins, T., Levy, O., & Zettlemoyer, L. (2018, July). Deep RNNs Encode Soft Hierarchical Syntax. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 14–19). Melbourne, Australia: Association for Computational Linguistics. doi: 10.18653/v1/P18-2003

Cai, Z. G., Duan, X., Haslett, D. A., Wang, S., & Pickering, M. J. (2024, March). *Do large language models resemble humans in language use?* (No. arXiv:2303.08014). arXiv. doi: 10.48550/arXiv.2303.08014

Chang, T. A., & Bergen, B. K. (2024). Language model behavior: A comprehensive survey. *Computational Linguistics*, *50*(1), 293–350.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., … Xing, E. P. (2023, March). *Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality.*

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. doi: 10.1017/s0140525x12000477

Crinean, M., & Garnham, A. (2006). Implicit causality, implicit consequentiality and semantic roles. *Language and Cognitive Processes*, *21*(5), 636–648. doi: 10.1080/01690960500199763

Davis, F., & van Schijndel, M. (2020, November). Discourse structure interacts with reference but not syntax in neural language models. In R. Fernández & T. Linzen (Eds.), *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 396–407). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-1.32

Davis, F., & van Schijndel, M. (2021, June). *Uncovering Constraint-Based Behavior in Neural Models via Targeted Fine-Tuning* (No. arXiv:2106.01207). arXiv. doi: 10.48550/arXiv.2106.01207

Ehrlich, K. (1980). Comprehension of pronouns. *Quarterly Journal of Experimental Psychology*, *32*(2), 247–255. doi: 10.1080/14640748008401161

Ettinger, A. (2020, January). What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, *8*, 34–48. doi: 10.1162/tacl_a_00298

Garnham, A., Child, S., & Hutton, S. (2020). Anticipating causes and consequences. *Journal of Memory and Language*, *114*, Article 104130. doi: https://doi.org/10.1016/j.jml.2020.104130

Garnham, A., Vorthmann, S., & Kaplanova, K. (2020). Implicit consequentiality bias in English: A corpus of 300+ verbs. *Behavior Research Methods*, *53*, 1530–1550. doi: 10.3758/s13428-020-01507-z

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., … Ma, Z. (2024, November). *The Llama 3 Herd of Models* (No. arXiv:2407.21783). arXiv. doi: 10.48550/arXiv.2407.21783

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018, June). Colorless Green Recurrent Networks Dream Hierarchically. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1195–1205). New Orleans, Louisiana: Association for Computational Linguistics. doi: 10.18653/v1/N18-1108

Hawkins, R., Yamakoshi, T., Griffiths, T., & Goldberg, A. (2020, November). Investigating representations of verb bias in neural language models. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4653–4663). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.376

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019). The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.* (To appear)

Huynh, H., Lentz, T. O., & van Miltenburg, E. (2022, December). *Implicit causality in GPT-2: A case study*

(No. arXiv:2212.04348). arXiv. doi: 10.48550/arXiv.2212.04348

Kankowski, F., Solstad, T., Zarriess, S., & Bott, O. (2025, January). *Implicit Causality-biases in humans and LLMs as a tool for benchmarking LLM discourse capabilities* (No. arXiv:2501.12980). arXiv. (Comment: 38 pages, 8 figures) doi: 10.48550/arXiv.2501.12980

Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2008). Coherence and coreference revisited. *Journal of Semantics*, *25*(1), 1–44. doi: 10.1093/jos/ffm018

Lewis, M., & Mitchell, M. (2024). Using Counterfactual Tasks to Evaluate the Generality of Analogical Reasoning in Large Language Models. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*(0).

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024, June). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, *28*(6), 517–540. doi: 10.1016/j.tics.2024.01.011

Mistral AI, & NVIDIA. (2024, July). *Mistral NeMo.* https://mistral.ai/news/mistral-nemo/.

Murray, J. D. (1997). Connectives and narrative text: The role of continuity. *Memory & Cognition*, *25*(2), 227-236. Retrieved from `https://doi.org/10.3758/BF03201114` doi: 10.3758/BF03201114

OpenAI. (2022, November). *Introducing ChatGPT.* https://openai.com/index/chatgpt/.

OpenAssistant. (2023). *OpenAssistant.* https://open-assistant.io.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... Lowe, R. (2022, March). *Training language models to follow instructions with human feedback* (No. arXiv:2203.02155). arXiv. doi: 10.48550/arXiv.2203.02155

Shi, C., Yang, H., Cai, D., Zhang, Z., Wang, Y., Yang, Y., & Lam, W. (2024, October). *A Thorough Examination of Decoding Methods in the Era of LLMs* (No. arXiv:2402.06925). arXiv. (Comment: EMNLP 2024 Main) doi: 10.48550/arXiv.2402.06925

Sieker, J., Bott, O., Solstad, T., & Zarrieß, S. (2023). Beyond the bias: Unveiling the quality of implicit causality prompt continuations in language models. In *Proceedings of the 16th international natural language generation conference* (pp. 206–220).

Solstad, T., & Bott, O. (2022). On the nature of implicit causality and consequentiality: The case of psychological verbs. *Language, Cognition and Neuroscience*, *37*(10), 1311–1340. doi: 10.1080/23273798.2022.2069277

Solstad, T., & Bott, O. (2023). Implicit causality and consequentiality of action verbs. *Frontiers in Language Sciences*, *2*(1143214). doi: 10.3389/flang.2023.1143214

Stede, M., & Umbach, C. (1998). Dimlex: A lexicon of discourse markers for text generation and understanding. In *Coling 1998 volume 2: The 17th international conference on computational linguistics.*

Stevenson, R. J., Crawley, R. A., & Kleinman, D. (1994). Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, *9*(4), 473–592.

Sweetser, E. (1990). *From etymology to pragmatics: The mind-body metaphor in semantic structure and semantic change*. Cambridge: Cambridge University Press.

Tenney, I., Das, D., & Pavlick, E. (2019, August). *BERT Rediscovers the Classical NLP Pipeline* (No. arXiv:1905.05950). arXiv. (Comment: Presented at ACL 2019) doi: 10.48550/arXiv.1905.05950

Upadhye, S., Bergen, L., & Kehler, A. (2020). Predicting reference: What do language models learn about discourse models? In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 977–982).

Verma, M., Bhambri, S., & Kambhampati, S. (2024, March). Theory of Mind abilities of Large Language Models in Human-Robot Interaction : An Illusion? In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 36–45). (Comment: Accepted in alt.HRI 2024) doi: 10.1145/3610978.3640767

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D., & Batra, D. (2016). Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., ... Le, Q. V. (2021, October). Finetuned Language Models are Zero-Shot Learners. In *Proceedings of the 11th International Conference on Learning Representations.*

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., ... Kim, Y. (2024, March). *Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks* (No. arXiv:2307.02477). arXiv. (Comment: NAACL 2024) doi: 10.48550/arXiv.2307.02477

Yuan, Z., Yuan, H., Tan, C., Wang, W., & Huang, S. (2023, March). *How well do Large Language Models perform in Arithmetic tasks?* (No. arXiv:2304.02015). arXiv. doi: 10.48550/arXiv.2304.02015

Zarrieß, S., Groener, H., Solstad, T., & Bott, O. (2022). This isn't the bias you're looking for: Implicit causality, names and gender in German language models. In R. Schaefer, X. Bai, M. Stede, & T. Zesch (Eds.), *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)* (pp. 129–134). Potsdam, Germany: KONVENS 2022 Organizers.

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., ... Levy, O. (2023, December). LIMA: Less Is More for Alignment. *Advances in Neural Information Processing Systems*, *36*, 55006–55021.