Large Language Models Are Partially Primed in Pronoun Interpretation

Suet-Ying Lam[†*] **Qingcheng Zeng** [‡*] **Kexun Zhang** [◇*] **Chenyu You** [♣] **Rob Voigt**[‡]

[†] Department of Linguistics, UMass Amherst
[‡] Department of Linguistics, Northwestern University
[◇] Department of Computer Science, UCSB
[♣] Department of Electrical Engineering, Yale University

Abstract

While a large body of literature suggests that large language models (LLMs) acquire rich linguistic representations, little is known about whether they adapt to linguistic biases in a human-like way. The present study probes this question by asking whether LLMs display human-like referential biases using stimuli and procedures from real psycholinguistic experiments. Recent psycholinguistic studies suggest that humans adapt their referential biases with recent exposure to referential patterns; closely replicating three relevant psycholinguistic experiments from Johnson and Arnold (2022) in an in-context learning (ICL) framework, we found that InstructGPT adapts its pronominal interpretations in response to the frequency of referential patterns in the local discourse, though in a limited fashion: adaptation was only observed relative to syntactic but not semantic biases. By contrast, FLAN-UL2 fails to generate meaningful patterns. Our results provide further evidence that contemporary LLMs discourse representations are sensitive to syntactic patterns in the local context but less so to semantic patterns. Our data and code are available at `https://github.com/zkx06111/llm_priming`.

1 Introduction

While neural network models, and particularly pretrained large language models have shown excellent performance at particular language processing tasks, many questions remain about the extent to which models optimized for such performance encode, as a side effect, human-like linguistic knowledge and cognitive biases. We know that they do to some extent; existing work has shown, for example, that neural models encode aspects of human-like long-distance number agreement (Gulordava et al., 2018), incremental syntactic state (Futrell et al., 2019), and syntactic generalization more broadly (Hu et al., 2020). In this paper, we examine whether FLAN-UL2 (Tay et al., 2023) and InstructGPT (Ouyang et al., 2022), two representative LLMs, display adaptation in pronoun interpretation when exposed to consistent referential patterns in the local discourse context.

Compared with syntactic or lexical knowledge, representing referential knowledge is possibly more complex; we know from psycholinguistic studies that human referential interpretation integrates multiple levels of linguistic structure. Humans do not interpret ambiguous pronouns at random but are guided by both syntactic and semantic information. It is well-established that absent other cues humans prefer a syntactic subject in choosing the antecedent of the ambiguous pronoun, i.e., **subject bias** (Ariel, 1990; Brennan, 1995). In example (1), *she* is more likely to be interpreted as the subject *Ada* than the non-subject *Eva* [¶], even though both referents are possible antecedents for the pronoun.

(1) Ada_1 talked with Eva_2. She_1...

People are also sensitive to the semantic structure of the sentence when choosing an antecedent for an ambiguous pronoun, in addition to syntactic information. In a transfer event that depicts a transfer-of-possession from one entity (the source) to another (the goal), they prefer the goal referent (Ada in (2), Eva in (3)) over the source referent (Eva in (2), Ada in (3)) to be the antecedent (Arnold, 2001, 1998), i.e., **goal bias**:

(2) Goal-source (gs) verb:
Ada_1 received a letter from Eva_2. She_1...

(3) Source-goal (sg) verb:
Ada_1 sent a letter to Eva_2. She_2...

[*] Equal contribution by alphabetical order. Correspondence to qcz@u.northwestern.edu

[¶] Indices of pronouns in examples indicate the preferred referent.

Findings of the Association for Computational Linguistics: ACL 2023, pages 9493–9506
July 9-14, 2023 ©2023 Association for Computational Linguistics

In sum, people exhibit sensitivity to both syntax (subject bias) and semantics (goal bias) during pronoun interpretation. Importantly, these levels of linguistic structure are frequently entwined since both can influence referential interpretation. At times these influences may push in different directions: for instance, *Ada* in [3] is both the syntactic subject and the semantic source.

Building on a long tradition investigating preferences in pronoun interpretation, recent psycholinguistic studies have probed into the deeper question of the origin of these biases. One hypothesis is that referential biases come from linguistic experience: when a bias occurs very frequently, people will tend to adapt this more frequent referential pattern, both in the immediate exposure as well as more large-scale past experience (Arnold, 1998, 2001). Recent evidence has provided support to this idea by demonstrating that recent exposure to certain referential patterns did change people's referential biases. In a series of psycholinguistic studies, Johnson and Arnold (2022) show that after reading numerous stories that consistently show a particular referential bias, e.g., always referring to the non-subject or source referent, people did have a stronger preference for these primed referents.

Given this line of psycholinguistic research, the current study investigates the extent to which LLMs adapt and vary referential biases in pronoun interpretation through exposure to referential patterns in the local context. To do so we replicated actual psycholinguistic experiments from Johnson and Arnold (2022) in LLMs using an ICL framework and asked whether the responses of LLMs display adaptation from exposure to referential patterns like human experimental participants. Comparing syntactically-motivated to semantically-motivated exposure conditions will allow us to first examine whether LLMs display human-like subject bias and goal bias and further understand the extent to which LLMs make use of local frequency changes in discourse representations in these categories.

In-context learning refers to LLMs' ability to learn from demonstrations written in natural language prompts. Compared to previous work that has largely examined the encoding of such discourse knowledge using zero-shot inference (Upadhye et al., 2020), ICL is particularly suitable for experimental simulation since it replicates the naturalistic context in which later responses draw upon exposure to previous examples.

Foreshadowing our results, we find that InstructGPT can adapt and thus vary its syntactic bias from exposure to referential patterns in the local context, but the same is not true for semantic bias. Given that InstructGPT still exhibits a goal bias in spite of local discourse priming, we argue this suggests LLMs only encode partial semantic knowledge in referential processing. To sum up, our contributions can be summarized as follows:

1) We extended a discourse understanding evaluation to state-of-the-art LLMs from a new perspective, asking whether LLM's referential bias can be modified by exposure to particular referential patterns, like how humans adapt referential bias from experience.

2) To the best of our knowledge, we are the first study that replicates actual psycholinguistic experiments using the ICL framework and compares LLMs' behaviors with real human participants.

3) We present results in this context showing further evidence that InstructGPT can acquire abstract syntactic knowledge in referential interpretation to some extent, but not semantic knowledge.

2 Related Work

A growing body of literature has suggested that LLMs encode rich representations of linguistic structure at various levels, including aspects of syntax, semantics, and reference encoded throughout their representations (Tenney et al., 2019). One of the most well-documented lines of this work demonstrates that these models can acquire diverse elements of syntactic knowledge (Gulordava et al., 2018; Futrell et al., 2019; Hu et al., 2020).

This capacity for encoding linguistic understanding extends to priming effects with psycholinguistic analogies to humans. At the syntactic level, Sinclair et al. (2022) explored structural priming in various autoregressive LLMs and found priming effects despite a clear dependence on semantic information. At the semantic level, using English lexical stimuli in BERT (Devlin et al., 2019), Misra et al. (2020) found that BERT does display evidence of sensitivity to semantic priming, though this is localized to more unconstrained contexts and only certain semantic relations.

In analyses of LLMs' linguistic understanding modeled on psycholinguistic experiments, however, the question of discourse knowledge remains relatively under-explored. Recent existing work has presented partially conflicting accounts in this area, in particular with regard to how LLMs may or may not exhibit human-like biases in pronoun interpretation. For example, Davis and van Schijndel (2020) compared LSTM LMs and Transformer LMs behaviors and internal representations in dealing with implicit causality verbs, finding that surprisingly (contrary to humans) implicit causality only influences Transformer LMs' behavior for reference, but influences neither model for syntactic attachment. Sorodoc et al. (2020) also compared LSTM LMs and Transformer LMs in coreference resolution corpora, finding that although LMs are much better at grammar, they also captured referential aspects to some extent.

This existing work replicates sets of individual stimuli from psycholinguistic experiments in isolation; by contrast, our work takes a more behaviorally-oriented approach by replicating the stimuli, procedure, and even experimental design of a full set of human psycholinguistic studies. We do this to ask a further question: beyond exhibiting baseline human-like referential biases in pronoun interpretation, do LLMs display adaptation to the frequency of referential patterns in the local context like humans? This question is important because adapting to a referential bias over the course of an experiment requires a sustained representation of the frequency of the pattern, which involves a higher-level understanding of the discourse structure. Humans have shown abilities of this kind of adaptation at multiple linguistic levels. For instance, exposure to syntactic structures consecutively affects humans' choice of structures in both the short term and long term via priming effects (e.g., Branigan et al. 2005; Chang et al. 2000). Exposure to phonological patterns can also guide humans' segmentation patterns (e.g., Saffran et al. 1996a,b). At the semantic and pragmatic levels, humans can also adapt as listeners to speakers' variable choices of uncertainty expressions (Schuster and Degen, 2020).

3 Methods

In this work, we aim to replicate three experiments from Johnson and Arnold (2022), transferring their designs, procedures, and stimuli as faithfully as possible to the LLM context using in-context learning.

3.1 Source Experiments

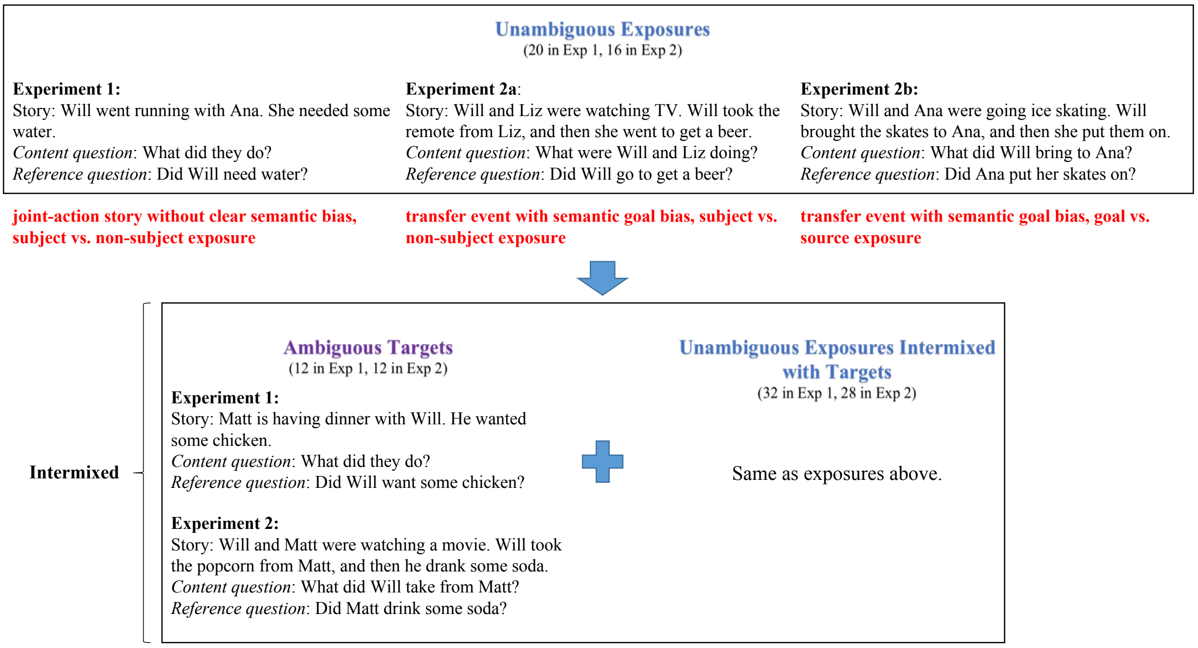
We first briefly summarize the experimental setup employed by Johnson and Arnold (2022). In each experimental setting, participants heard a series of two-sentence stories in which the first sentence contained two characters with gender-marked first names (Matt or Will for men, Liz or Ana for women). For each story, participants answered a content question to check comprehension, and then a reference question to check pronominal interpretation. In order to lower the ceiling and keep participants from falling into a pattern of simply answering "yes," reference questions were equally split between default and non-default phrasings (e.g. between the subject/non-subject interpretations in Experiment 1a and 2a and between the source/goal interpretations in Experiment 2b). Figure 1 illustrates sample stimuli and the procedure of the three experiments conducted.

In each experiment participants were first shown a series of stories with *exposure* reference questions to establish a referential pattern; in these, the characters had different genders so pronoun interpretation was unambiguous. After the exposure phase (20 stories in Experiment 1, 12 stories in Experiment 2), further stories with unambiguous exposure questions were intermixed with 12 stories accompanied by *critical* reference questions; in these stories, the characters had the same gender, so pronoun interpretation was ambiguous and required reliance on discourse cues.

Across experiments, there were two key conditions that were manipulated for exposures. Under a *subject exposure* condition, the unambiguous intended referents of all exposure questions are subjects of the preceding clauses; in the corresponding *non-subject exposure* condition, they are the objects of the preceding clauses. Similarly, under a *goal exposure* condition the unambiguous intended referents of all exposure questions are goal referents while in a *source exposure* condition they are source referents.

Aiming to transfer these experiments as faithfully as possible to the LLM context, we used this experimental paradigm to evaluate the LLM by providing the model with the full text of each story prompt and content/reference question, then generated tokens in response which we interpreted as answers.

For clarity, we use the same experiment numbers and identifiers as in Johnson and Arnold (2022).



Experiment 1:
Story: Will went running with Ana. She needed some water.
Content question: What did they do?
Reference question: Did Will need water?

Experiment 2a:
Story: Will and Liz were watching TV. Will took the remote from Liz, and then she went to get a beer.
Content question: What were Will and Liz doing?
Reference question: Did Will go to get a beer?

Experiment 2b:
Story: Will and Ana were going ice skating. Will brought the skates to Ana, and then she put them on.
Content question: What did Will bring to Ana?
Reference question: Did Ana put her skates on?

joint-action story without clear semantic bias, subject vs. non-subject exposure
transfer event with semantic goal bias, subject vs. non-subject exposure
transfer event with semantic goal bias, goal vs. source exposure
Intermixed
Ambiguous Targets
(12 in Exp 1, 12 in Exp 2)

Experiment 1:
Story: Matt is having dinner with Will. He wanted some chicken.
Content question: What did they do?
Reference question: Did Will want some chicken?

Experiment 2:
Story: Will and Matt were watching a movie. Will took the popcorn from Matt, and then he drank some soda.
Content question: What did Will take from Matt?
Reference question: Did Matt drink some soda?

Unambiguous Exposures Intermixed with Targets
(32 in Exp 1, 28 in Exp 2)
Same as exposures above.

Figure 1: Illustrated Experimental Procedure. Closely following the setup with human participants in Johnson and Arnold (2022), LLMs were primed via exposure to story/question pairs with unambiguous referents, and tested for their responses on ambiguous target pairs.

Note that we did not replicate experiments 1b and 1c, which investigated whether people are still sensitive to referential patterns using different types of referential expressions (e.g., third-person names and first- and second-person pronouns). While these experiments provided insights into the linguistic structure at which people generalize referential biases, they were less relevant to the objective of this study and thus were not included.

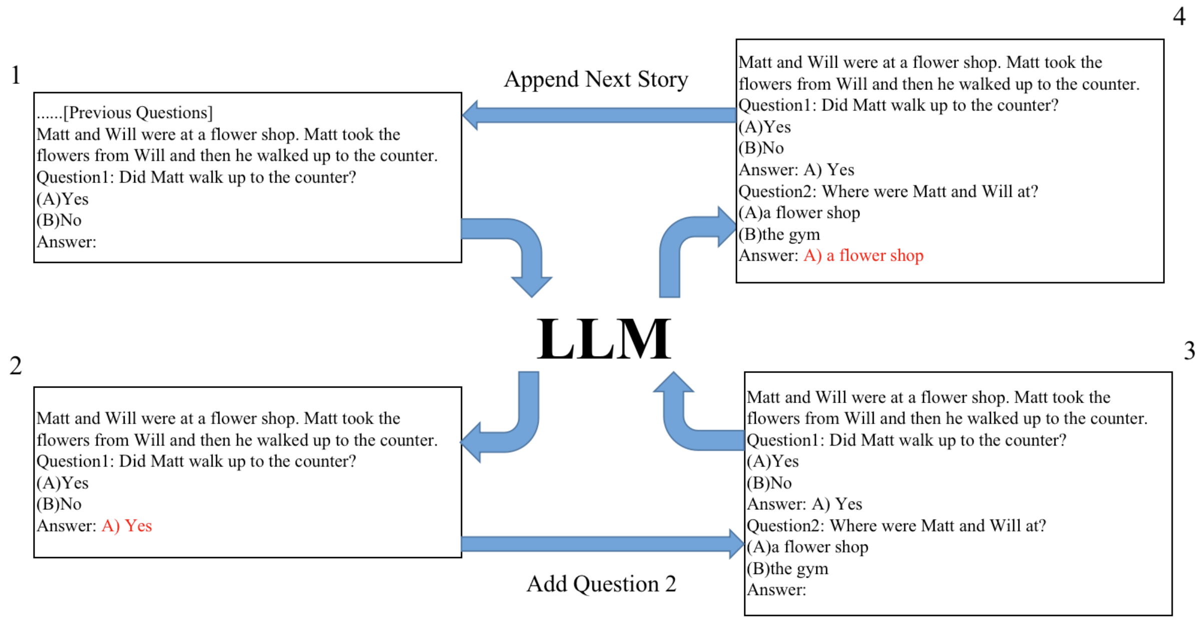
Experiment 1a In this experiment, all story prompts contain "joint-action" verbs using "with" in the form "X did something with Y." Since these verbs lack a clear semantic bias (e.g., Arnold et al. 2018), this context allows us to evaluate LLMs' sensitivity to syntax-based biases in discourse by asking whether exposure to only subject-bias or object-bias examples will influence following answers on the ambiguous critical items. If LLMs are sensitive to syntax-based referential patterns, we expect more subject responses under the subject exposure condition and more non-subject responses under the non-subject exposure condition.

Experiment 2a Experiment 2a forms a bridge between adaptation to syntactic and semantic constraints. Are LLMs able to track patterns in both categories, for instance learning an exposure bias in one category while ignoring variation in the other? In this experiment, all story prompts contain source/goal verbs like "give" and "receive," but these are distributed equally throughout exposures. The manipulation remains the same as Experiment 1a, in which LLMs are exposed to consistent and unambiguous subject interpretations in the subject exposure condition and non-subject interpretations in the non-subject exposure condition.

Experiment 2b This experiment focuses solely on source/goal biases, in which all story prompts contain source/goal words, but the unambiguous exposure items are manipulated to contain only source references in the source exposure condition and only goal references in the goal exposure condition.

3.2 In-context Learning

We propose that since these experiments rely on short-term learning effects of exposure in an experimental context, they can only be effectively simulated with LLMs by using in-context learning recursively. Specifically, for each question, the model is provided access to all previous items in answering a new question, including its own previous responses. This is intended to mirror the process of human experimental participants making judgments in the light of recent exposure to input and



......[Previous Questions]
Matt and Will were at a flower shop. Matt took the flowers from Will and then he walked up to the counter.
Question1: Did Matt walk up to the counter?
(A)Yes
(B)No
Answer:

Matt and Will were at a flower shop. Matt took the flowers from Will and then he walked up to the counter.
Question1: Did Matt walk up to the counter?
(A)Yes
(B)No
Answer: A) Yes

Matt and Will were at a flower shop. Matt took the flowers from Will and then he walked up to the counter.
Question1: Did Matt walk up to the counter?
(A)Yes
(B)No
Answer: A) Yes
Question2: Where were Matt and Will at?
(A)a flower shop
(B)the gym
Answer:

Matt and Will were at a flower shop. Matt took the flowers from Will and then he walked up to the counter.
Question1: Did Matt walk up to the counter?
(A)Yes
(B)No
Answer: A) Yes
Question2: Where were Matt and Will at?
(A)a flower shop
(B)the gym
Answer: A) a flower shop

Figure 2: The ICL Simulation Framework. Black text is provided to the models as prompts, red text is generated by the model. With the last step's output appended to previous prompts, we ask LLMs the next question.

their own past responses.

We manually checked the correctness of LLM responses to content questions intended to check comprehension, and as in the human experimentation context removed answers for which the LLMs provided incorrect answers. The answers to critical target questions were recorded for further statistical analysis. Our ICL procedure is shown in Figure 2.

3.3 Models and Experimental Settings

We used *text-davinci-003* from the OpenAI API and open-sourced FLAN-UL2 as the LLMs of interest. Though these models are of course not exhaustive of the current landscape of LLMs, they provide some diversity since they differ in both structure (decoder-only vs. encoder-decoder) and parameter count (175B vs. 20B). To introduce more randomness into the experiment and allow enough sample sizes for statistical comparisons across conditions, we made slight modifications to the *temperature* hyperparameter on each run. Specifically, we assigned each run a random and unique temperature value between 0.2 to 1.0.

We also attempted to simulate 'participants' using a natural language prompt to approach different speaker identities following a similar methodology to Aher et al. (2022). We developed prompts with slots for titles, names, and country of origin to establish different character backgrounds simulating native English speakers from the United States, Britain, and Australia, following the participants' demographics in Johnson and Arnold (2022). However, these prompts did not induce greater diversity in responses than temperature modification, so only results using temperature modification are presented below. We present a further analysis for both methodologies in Appendix B.

In the end, we simulated 24 'participants' each in Experiment 1a and Experiment 2a, and 60 'participants' in Experiment 2b. We included more in Experiment 2b because this experiment has a lower response variability and thus needs more data points for the statistical analysis.

3.4 Measures

Following the analytic approach of Johnson and Arnold (2022), we used a regression-based approach to analyze whether the responses of LM are consistent with the subject or gogoalal bias of the context for each experiment. We can then compare our findings with theirs by asking whether the effect of the main predictors is the same. In all experiments, predictors included exposure type (unambiguous exposures to subject/non-subject or goal/source), reference question type (whether the reference question is asked about the subject/non-subject or goal/source), and the interaction effect between them. Experiment 2 included verb type, as well as its two-way and three-way interaction effects with the other two predictors as addi-

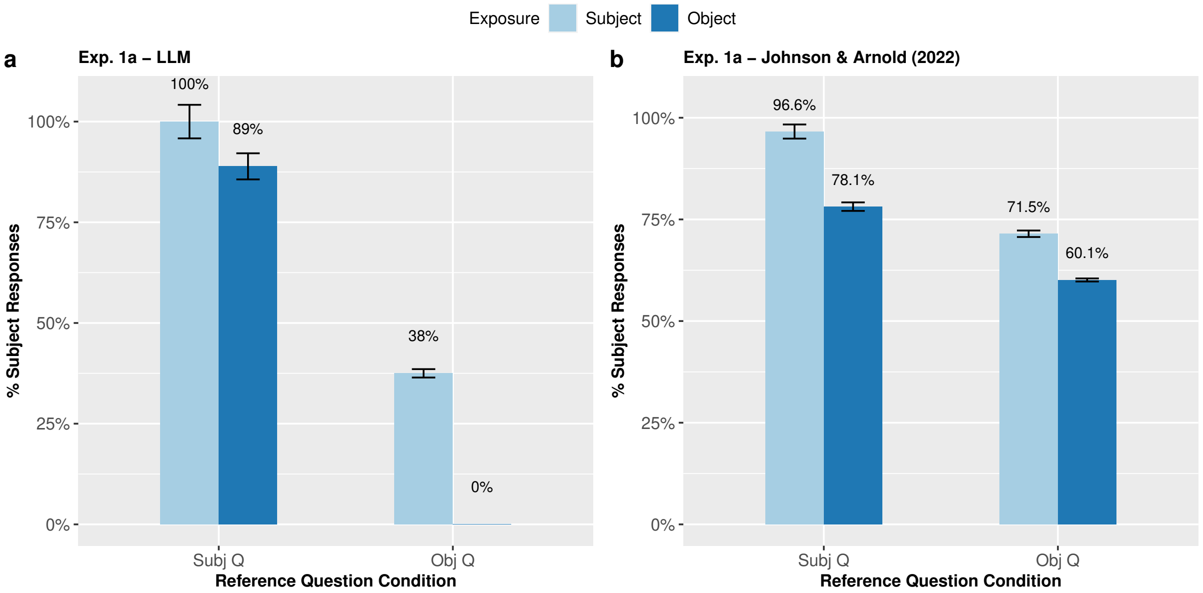


Figure 3: Subject responses of **LLM (left)** and **human participants (right)** in Experiment 1a, showing the percentage of subject responses for each type of reference question (Subj Q: subject reference question; Obj Q: object reference question), grouped by exposure type (subject exposure: light blue; object exposure: dark blue).

tional predictors. Given our small dataset, the results were analyzed using Bayesian mixed-effects Bernoulli logistic regression models in the R package *brms* (Bürkner, 2017) instead of a frequentist model. We report a Bayesian equivalent p-value (p_MAP) computed with the R package *bayestestR* (Makowski et al., 2019) to offer a straightforward interpretation of the results. Details of models are provided in Appendix 1.

4 Results

4.1 FLAN-UL2

We found that FLAN-UL2 was not capable of generating meaningful output for analysis under this design. First, FLAN-UL2 showed a much higher false rate in answering content questions versus InstructGPT: while InstructGPT replied 100% correctly to these questions, FLAN-UL2 replied to only 57% correctly. Second, for the ambiguous target items, FLAN-UL2 answered 'yes' 100% of the time, indicating an extremely strong bias towards simply answering 'yes' and producing no meaningful variation. By contrast, InstructGPT answered 'yes' to target questions 68% of the time, suggesting some amount of 'yes' bias but to a much weaker degree. From these findings we concluded that FLAN-UL2 did not produce sufficiently clean outputs for analysis; therefore, in the following sections, we will focus on results from InstructGPT.

4.2 Experiment 1a

Experiment 1a asks whether LLMs are sensitive to the frequency of referential patterns when subject referents are preferred. Figure 3 compares the subject responses of our results with Johnson and Arnold (2022)'s results. In both LLM and human data, we saw fewer subject responses in object exposure than in subject exposure. This is confirmed by the main effect of exposure type (*p_MAP* < .001) revealed in the statistical analyses. There was also a main effect of response question type (*p_MAP* = .009). As seen from the figure, there were more subject responses in the subject than object reference questions in both human and LLM data. This is because InstructGPT did answer 'yes' more frequently in general, which led to more subject interpretations when the question asked about the subject (where answering 'yes' indicates a subject interpretation), and fewer subject interpretations when the question was asked about the non-subject (where answering 'yes' indicates a non-subject interpretation).

However, we did not find any interaction effect between exposure type and reference question type. While Johnson and Arnold (2022) found that exposure type has a significant effect for subject-referent questions but only a marginal effect for the nonsubject-referent questions, the exposure effect was significant for both question types in our

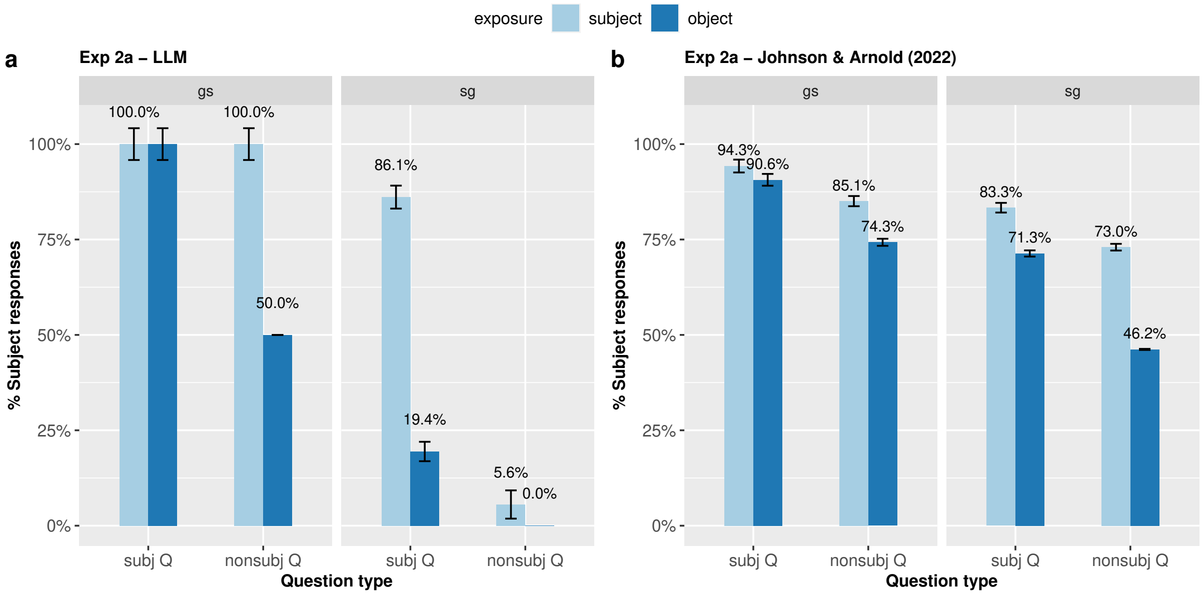


Figure 4: Subject responses of **LLM (left)** and **human participants (right)** in Experiment 2a. The proportion of subject responses is plotted against question type faceting by verb type, comparing goal-subject verbs (gs) like "receive," where the subject is the goal referent) to subject-goal verbs (sg) like "send," where the subject is the source referent.

results. InstructGPT is sensitive to exposure type even for nonsubject-referent questions, as reflected in Figure 3: there were no subject responses when non-subject reference questions were asked under the non-subject exposure condition. This suggests that LLM may even be more sensitive to the exposure effect than humans. Overall, Experiment 1 shows that LLM can indeed learn from recent exposure to syntactically-oriented referential patterns, in a relatively more non-subject-biased way.

4.3 Experiment 2a

We ask in Experiment 2a if the sensitivity to subject reference patterns observed in Experiment 1a persists independent of semantic variability in the local context. Figure 4 illustrates the responses of LLM and human participants by exposure, reference question, and verb type. The behavior of InstructGPT is in line with what Johnson and Arnold (2022) found in human participants for Experiment 2a: as in Experiment 1, statistical analyses revealed a significant main effect of referent question type (*p_MAP* = .034), despite a marginally significant effect of exposure type (*p_MAP* = .085). InstructGPT did understand the pronoun as subject referents more after subject exposure and said ‘yes’ more often such that there were more subject interpretations when the critical question was asked about the subject. InstructGPT also showed a goal bias: there was a main effect of verb type (*p_MAP* = .001), such that it referred to the subject more with gs verbs (where the subject is the goal referent) than sg verbs (where the subject is the source referent). We observed no interaction effect.

In spite of these similarities, we still observe differences between LLM and human participants. Notice in Figure 4 that there is a larger difference between gs and sg verbs in InstructGPT than in human participants while keeping other conditions constant, suggesting that LLMs may have a larger goal bias than humans.

4.4 Experiment 2b

Experiment 2b examines whether InstructGPT is still sensitive to exposure to referential patterns that exhibit consistent preferences for a source or goal referent. Figure 5 compares the results from InstructGPT and human participants. Whereas Johnson and Arnold (2022) reported significant main effects from exposure type, reference question type, and verb type, as well as a marginal interaction effect between verb type and exposure, we did not find any effect from these predictors, but only an interaction effect between verb type and question type (*p_MAP* < .001). We further examined the effect of question type and exposure type for each verb type, but no significant effect was found for either predictor. These results suggest that exposure

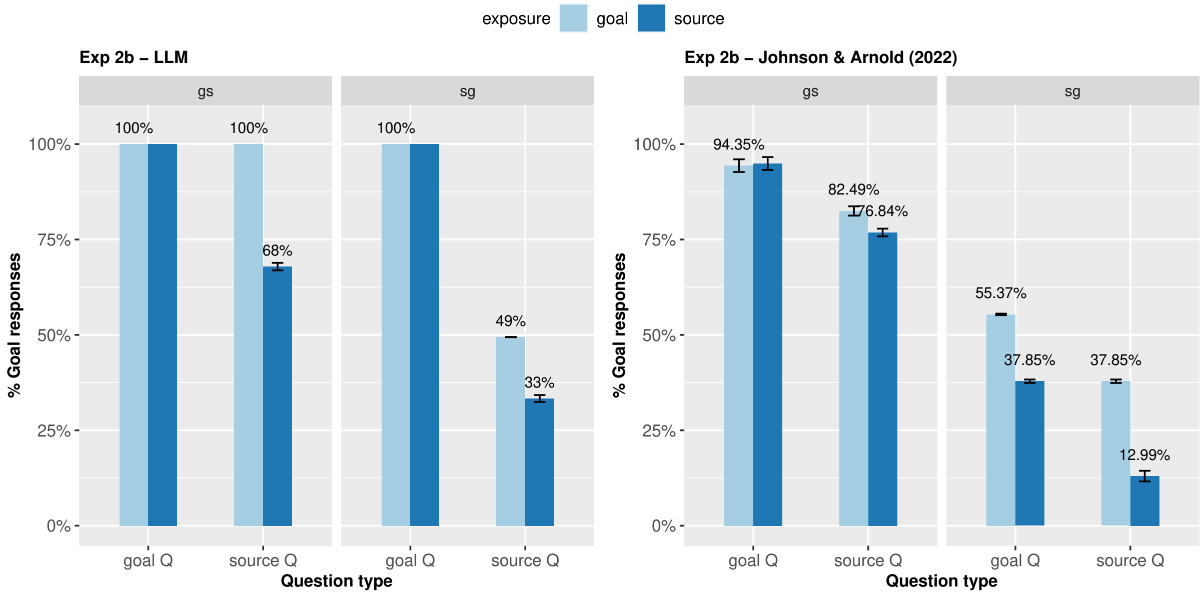


Figure 5: Goal responses of **LLM (left)** and **human participants (right)** in Experiment 2b.

did not necessarily change InstructGPT's behaviors: for goal questions, both the goal and source exposure conditions led to 100% goal responses. Reference question type also did not affect InstructGPT's responses, because it almost always interprets the pronoun as the goal referent under subject exposure. The lack of any significant effects, and indeed our observed universal goal interpretations to goal-focused reference questions, suggest that InstructGPT displays an extreme goal bias in pronoun interpretation.

5 Discussion

The present study examined whether and to what extent LLMs display adaptation for pronominal interpretation after exposure to referential patterns by replicating three psycholinguistic experiments. Exposed to the same stimuli and study design, and analyzed with the same statistical procedures as the source experiments, we asked whether LLMs show human-like behaviors and compared the performance of LLMs-simulated participants with human participants.

We firstly found a difference between the capacity of contemporary models to replicate human psycholinguistic experimental designs in an ICL framework. While InstructGPT was able to correctly answer all the presented comprehension-check content questions, FLAN-UL2 was not. Both models displayed at least some bias towards answering 'yes' in ambiguous cases, but in the case of FLAN-UL2 this bias resulted in 100% 'yes' answers, rendering meaningful variation impossible to ascertain. This could be a result of structural differences between the models (decoder-only in the case of InstructGPT, encoder-decoder in the case of FLAN-UL2), or perhaps more likely a question of simple model size (175B for InstructGPT vs. 20B for FLAN-UL2).

Experiments 1a and 2a examined whether LLMs adapt their syntactic bias from recent exposure to referential patterns like humans, without and with the presence of possibly confounding semantic goal bias. We found that LLM's referential biases are indeed sensitive to such exposure in both experiments. In addition, LLM did exhibit a goal bias in Experiment 2a, in line with previous studies which argue for LLM's ability to exhibit human-like semantic bias in pronoun interpretation (e.g., Davis and van Schijndel 2020).

Experiment 2b examined whether LLM can adapt and vary their semantic bias from exposures. In this context, in contrast with the previous two experiments, exposure type did not affect LLM behavior at all. This raises the question of why LLM would be sensitive to exposure to only referential patterns that exhibit consistent syntactic bias but not semantic bias. An immediate and intuitive answer would be that LLM is unable to fully represent semantic knowledge in referential processing. However, given that LLM did display a human-like goal bias in pronoun interpretation independent of exposure, this explanation seems unlikely.

We suggest that our work provides evidence that LLM only partially represents the semantic knowledge involved in referential processing for two reasons.

First, adapting referential biases on the basis of exposure in the local context may require representations of higher-level knowledge than merely exhibiting a bias towards certain referents. While the latter may only require knowledge of which features associated with a referent are more frequent or likely in general, adaption requires a sustained awareness of referential pattern frequency as it changes in the local discourse context. Though representing knowledge about semantic relations was observed as early as the analogical reasoning task in Word2Vec (Mikolov et al., 2013), human-like extraction of abstract information like a persistent discourse state is more challenging. The model may only be able to identify thematic roles (source or goal) of a referent and associate them with pronoun interpretation, but not to identify a consistent pattern of thematic roles across a discourse. If so, this would explain the strong goal bias we observed in Experiment 2.

Second, it is possible that LLM has an extremely strong goal bias that masks the influence of exposures. If so, this suggests that LLM over-represents the semantic knowledge encoded in pronoun interpretation.

In either case, our results suggest that LLM do not encode semantic knowledge in a fully human-like way, even though they do demonstrate some human-like capacities for semantic understanding. Although we believe this gap can be mitigated via instruction fine-tuning or chain-of-thought prompting (Wei et al., 2022), these results still suggest we should consider incorporating semantically-informed objectives into self-supervised pre-training to a greater extent.

6 Conclusions and Future Work

By replicating a series of psycholinguistic experiments as closely as possible using in-context learning, this paper pioneered whether LLMs would adapt pronominal interpretation behaviors in a human-like way given exposure to referential patterns in the local discourse context. Our work suggests paths forward for replicating psycholinguistic experiments in a more faithful way that allows for comparisons between human and LLMs' behaviors.

Limitations

Several main limitations exist in our study in its current form. First, our reported results only simulated experimental participants by manipulating the *temperature* hyperparameter. We compared this approach with natural language prompting for Experiment 1, but that prompting did not increase "participant" diversity, so it was abandoned. Moreover, approaches for simulating psycholinguistic experimental "participants" could go far beyond what was tried here; our prompting method was relatively limited, and more detailed prompting could be included in future experimental simulations. Second, making a direct comparison with actual psycholinguistic experiments might not be the only method to investigate LLMs' discourse capacity. A comprehensive list of discourse probing tasks might play a similar role despite a different way (Koto et al., 2021). Third, this study is strictly behavioral: limited by both computational resources and obscure mechanisms of in-context learning, we do not dive into models' internal representations in our analyses.

Acknowledgement

We are very grateful to Jennifer Arnold for sharing the stimuli and design of Johnson and Arnold (2022) with us.

References

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2022. Using large language models to simulate multiple humans.

Mira Ariel. 1990. *Accessing noun-phrase antecedents*. Routledge.

Jennifer E Arnold. 1998. *Reference form and discourse patterns*. Stanford University.

Jennifer E Arnold. 2001. The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse processes*, 31(2):137–162.

Jennifer E Arnold, Iris M Strangmann, Heeju Hwang, Sandra Zerkle, and Rebecca Nappa. 2018. Linguistic experience affects pronoun interpretation. *Journal of Memory and Language*, 102:41–54.

Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278.

Holly P Branigan, Martin J Pickering, and Janet F McLean. 2005. Priming prepositional-phrase attachment during comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3):468.

Susan E Brennan. 1995. Centering attention in discourse. *Language and Cognitive processes*, 10(2):137–167.

Paul-Christian Bürkner. 2017. brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80:1–28.

Franklin Chang, Gary S Dell, Kathryn Bock, and Zenzi M Griffin. 2000. Structural priming as implicit learning: A comparison of models of sentence production. *Journal of psycholinguistic research*, 29(2):217–230.

Forrest Davis and Marten van Schijndel. 2020. Discourse structure interacts with reference but not syntax in neural language models. *arXiv preprint arXiv:2010.04887*.

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

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

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Elyce D Johnson and Jennifer E Arnold. 2022. The frequency of referential patterns guides pronoun comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3849–3864, Online. Association for Computational Linguistics.

Dominique Makowski, Mattan S Ben-Shachar, and Daniel Lüdecke. 2019. bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, 4(40):1541.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2020. Exploring BERT's sensitivity to lexical cues using tests from semantic priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4625–4635, Online. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Jenny R Saffran, Richard N Aslin, and Elissa L Newport. 1996a. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Jenny R Saffran, Elissa L Newport, and Richard N Aslin. 1996b. Word segmentation: The role of distributional cues. *Journal of memory and language*, 35(4):606–621.

Sebastian Schuster and Judith Degen. 2020. I know what you're probably going to say: Listener adaptation to variable use of uncertainty expressions. *Cognition*, 203:104285.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022. Structural Persistence in Language Models: Priming as a Window into Abstract Language Representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Ionut-Teodor Sorodoc, Kristina Gulordava, and Gemma Boleda. 2020. Probing for referential information in language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4177–4189, Online. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. Ul2: Unifying language learning paradigms.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. *CoRR*, abs/1905.05950.

Shiva Upadhye, Leon Bergen, and Andrew Kehler. 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)*, pages 977–982, Online. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models.

A Statistical Information

For each experiment, we analyzed referent choice (Subject = 1, Non-subject = 0) using a mixed-effects Bernoulli regression model from the R package *brms* (Bürkner, 2017), with the maximal random structure justified by design (Barr et al., 2013). Predictors are coded in the same way as in Johnson and Arnold (2022). All models were specified with a weakly informative prior using the Cauchy distribution with center 0 and scale 2.5. Models were fitted using six chains, each with 4,000 iterations of which the first 1,000 are warmup to calibrate the sampler, resulting in 18,000 posterior examples.

The model for Experiment 1a included question type (QtypeC, sum-coded: Subject = 0.5, Non-subject = -0.5) and exposure type (PC, effects-coded: Subject-biased = 0.51, Object-biased = -0.49) as fixed predictors, random intercepts for participants and items, random slopes of question type, exposure type and their interaction for items, and a random slope of question type for the participant. Exposure type was not included as a random slope for participants because the condition does not vary within participants.

```
brm (Rc ~ QtypeC*PC +
        (1+QtypeC*PC|ID)+
        (1+QtypeC|Subject),
        data=(e1a,Exposure!="None"),
        family="bernoulli",
        chains=6,
        iter=4000,
        warmup=1000,
        control =
            list(adapt_delta = 0.95),
        prior=
          c(set_prior ("cauchy(0,2.5)")))
```

The model for Experiment 2a included question type (QtypeC, sum-coded: Subject = 0.5, Non-subject = -0.5), exposure type (PC, sum-coded: Subject-biased = 0.5, Object-biased = -0.5), and verb type (Vc, sum-coded: gs-verb = 0.5, sg-verb = -0.5) as fixed predictors, random intercepts for participants and items, random slopes of question type, exposure type and their interaction for items, and a random slope of question type for the participant. As in Experiment 1a, exposure type was not included as a random slope for participants because the condition does not vary within participants. Similarly, verb bias was not included as a random slope for items here because it does not vary within items.

```
brm (Rc ~ QtypeC*PC*Vc +
    (1+PC*Qtypec|Item)+
    (1+QtypeC*Vc|Subject),
    data=e2a,
    family="bernoulli", chains=6,
    iter=4000, warmup=1000,
  control = list(adapt_delta = 0.98),
    cores = 6,
    prior=
    c(set_prior ("cauchy(0,2.5)")))
```

The model for Experiment 2b included question type (QtypeC, sum-coded: Goal = 0.5, Source = -0.5), exposure type (PC, sum-coded: Goal-biased = 0.5, Source-biased = -0.5), and verb type (Vc, sum-coded: gs-verb = 0.5, sg-verb = -0.5) as fixed predictors. The random effect structure was the same as that of Experiment 2a.

```
brm (Rc ~ PC*Vc*QtypeC
   +(1+PC*QtypeC|Item)
   +(1+Vc*QtypeC|Subject),
   data=e2b,
   family="bernoulli",
   chains=6,
   iter=4000,
   warmup=1000,
   control =
       list(adapt_delta = 0.999),
   cores = 6,
   prior=
    c(set_prior ("cauchy(0,2.5)")))
```

B Temperature vs. Prompt

Our prompt-based simulation of multiple participants embedded names, countries, prefixes, and genders into a carrier sentence: *{Prefix + Name} is a native English speaker living in {Country}. {Gender} is asked in a psycholinguistic experiment to answer the following questions.* For example, *Mr. Smith is a native English speaker living in England. He is asked in a psycholinguistic experiment to answer the following questions.*

Specifically, we calculated the variance and ran the Levene's test for any significant difference between humans and InstructGPT in each experiment. In experiment 1, the human responses' variance is 0.055, the temperature-based responses' variance is 0.024, and the prompt-based responses' variance is 0.017. Human responses were significantly higher than both (Temperature-based: p = *.049*; Prompt-based: p = *.007*). Yet, temperature-based were not significantly higher than prompt-based. Due to the limitation of API pricing, we only ran temperature-based in the following experiments. In experiment 2a, the human responses' variance is 0.034 and the temperature-based responses' variance is 0.043. Yet, Levene's test did not reveal any significant difference. In Experiment 2b, the human responses' variance is 0.020 and the temperature-based responses' variance is 0.008 (p < *.001*).

We used different techniques to introduce randomness and include more experimental data in our experiments. We realized these were not well-designed prompts to elicit different linguistic backgrounds. Given the lack of investigation on simulating multiple participants in psycholinguistics studies, we recognize this as a future direction of possible work.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Last section without the number

□ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
The introduction is in the first section.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Experiments were described in section 3.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In the method section, we described our parameters.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. We did not do a hyperparameter search.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We provide detailed results in the results section.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
We provide existing packages' information in the methods section.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.