Discourse Sensitivity in Attraction Effects: The Interplay Between Language Model Size and Training Data

Sanghee J. Kim
University of Chicago
sangheekim@uchicago.edu

Forrest Davis
Colgate University
fdavis@colgate.edu

Abstract

While work on the linguistic ability of language models (LMs) is driven by a variety of aims, one dominant motivation is using LMs to determine what linguistic knowledge can be learned from unstructured text. The current work aims to evaluate LMs on discourse sensitivity—the capability to distinguish between content that is more relevant and important to the discourse and that which is less so. We ground our evaluation of LMs by leveraging an existing psycholinguistics study on the *number agreement attraction effect*, one of the well-studied measures of human language comprehension. Based on human empirical findings on the modulation of the attraction effect by discourse, we establish three tests that LMs should pass if they demonstrate discourse sensitivity. A total of 25 models were evaluated that vary in (i) model size (small or large) and (ii) training type (dialogue-based, plain, and instruction-based). The models showed systematicity in discourse sensitivity, though in ways dissimilar to humans, either by over-relying on structural cues or overusing discourse cues. Notably, models that patterned most similarly to human performance were predominantly smaller and those trained on dialogue-targeted data. We discuss the implications of these findings and insights into human language processing.

1 Introduction

A growing body of work has investigated the linguistic capabilities of language models (LMs), tackling aspects of syntax, semantics, and pragmatics (for a survey, see Chang and Bergen, 2024). While work on the linguistic ability of LMs is driven by a variety of aims, one dominant motivation is using LMs to determine what linguistic knowledge can be learned from unstructured text (Linzen and Baroni, 2021). Some work has claimed LMs obtain abstract linguistic knowledge, resolving complex syntactic (e.g., Wilcox et al., 2024) and anaphoric dependencies (e.g., Hu et al., 2020), and exhibiting signs of pragmatic skills (e.g., Hu et al., 2023), though there is nuance in what can be inferred from these types of results (for a case study in the limitations of inferring full grammatical knowledge from overlap in behavior, see Lan et al., 2024).

Much of the work on linguistic evaluations of LMs focuses on linguistic phenomena treated in isolation. For example, linguistic knowledge benchmarks like BLiMP (Warstadt et al., 2020) and SyntaxGym (Gauthier et al., 2020) explore linguistic phenomena separately (e.g., subject-verb agreement, argument structure) rather than the interaction of multiple processes (e.g., interactions between argument structure and agreement; for discussion see Davis (2022b)). The current study aims to expand on this body of work by investigating the interaction of discourse structure with syntactic dependencies. We ask whether exposure to a massive amount of text and differing forms of training (e.g., instruction finetuning) yields "knowledge" of discourse.

Building on a large body of work investigating subject-verb agreement in language models (Linzen et al., 2016; Arehalli and Linzen, 2020; Warstadt et al., 2020; Yedetore and Kim, 2024, a.o.), we focus on structures like the following:

(1) The waitress *who sat near the girls* was unhappy.

In (1), the agreement between the main verb (*was*) and the subject (*The waitress*) can be made difficult because of an interfering noun, *girls*, which, if misidentified as the subject, would yield a different agreement pattern (e.g., *were*). This influence of interfering nouns, when the subject-verb agreement needs to be resolved, leads to an *interference effect* and has been widely used in both human studies (e.g., Wagers et al., 2009) and evaluations of language models (e.g., Arehalli and Linzen, 2020).

In our study, we manipulate the discourse status of the relative clause containing the interfering noun. In (1), the relative clause (*who sat near the girls*) is a restrictive relative clause. Restrictive relative clauses conventionally convey essential information to the discourse (as they function as selecting a specific referent). By adding commas surrounding the relative clause (i.e., *The waitress, who sat near the girls, was unhappy*), we can signal an appositive relative clause, which conveys side-commentary information and are not part of the main assertion (Potts, 2005; AnderBois et al., 2015; Syrett and Koev, 2015; Koev, 2022; cf. Potts, 2012). We make use of this contrast in discourse status between the two structures to examine the interaction of discourse structure and syntactic dependencies, specifically cases where human processing of subject-verb agreement is modulated by discourse status.

As argued for in Suijkerbuijk et al. (2024), we ground our evaluation of language models via comparison to an existing psycholinguistic study demonstrating human *discourse sensitivity* (Kim and Xiang, 2024). Drawing on the same materials, we established three tests that LMs should pass to exhibit human-like behavior. Concretely, we investigated 25 models, including plain (base) and instruction-tuned models, and models trained on dialogue and conversational goal-oriented datasets, and those that vary in model size (small or large).

To preview the findings, the results suggest that (i) models trained on datasets with dialogue and goal-oriented conversations outperform other models, (ii) larger models do not yield human-like discourse sensitivity, and (iii) instruction-based training does not necessarily benefit models compared to base training. Taking these findings, we suggest that the *qualitative* nature of training data (e.g., genre and the specific types of constructions) is critical in the success of discourse sensitivity. We conclude by discussing insights into human language processing from evaluating language models and why instruction-tuned models underperform compared to base models, despite their seemingly advantageous training.

2 Background

2.1 Discourse structure: the division of more and less important information

Discourse can be defined in multiple different ways. It can be illustrated as a coherence relation (Hobbs, 1985; Kehler, 2002), the conversational moves for a successful discourse (Lewis, 1979; Farkas and Bruce, 2010), a hierarchically structured representation of discourse units (Polanyi, 1988; Asher and Lascarides, 2003; Jasinskaja, 2016), or a set of organized question and answer pairs to the conversational topic (Roberts, 2012), to name a few.

Regardless of the approaches to analyzing discourse, however, a shared notion of discourse is that certain parts of discourse are more important than others—some components of discourse are more relevant to the discourse topic, and others are less so. The examples in (2) demonstrate this contrast, realized at a sentence level:

(2) a. The waitress *who sat near the girl* was unhappy. [RRC]
b. The waitress, *who sat near the girl*, was unhappy. [ARC]

The same content, *that the waitress sat near the girl* is primary discourse information in (2a), essential to specify the very waitress that is being discussed, whereas it is secondary information in (2b), adding side-commentary details to the discourse. This division is expressed with the contrast of restrictive (RRC) (2a) and appositive relative clauses (ARC) (2b). Throughout the paper, we use these two structures to distinguish between different types of discourse status, serving as stand-ins for discourse structure at the sentence level.

2.2 Human sensitivity to discourse

Theoretical and experimental studies have shown that humans are highly sensitive to distinctions in information status (Potts, 2005; Syrett and Koev, 2015). In ongoing discourse, content that is part of the main discourse structure is judged to be a more natural continuation than content belonging to a non-main or subordinate discourse structure, such as information in an appositive relative clause (Syrett and Koev, 2015; Göbel, 2019). Discourse salience, topichood, and coherence have also been shown to affect real-time language comprehension and production. Entities that are salient in discourse are easier to recall and retrieved (e.g., Birch and Garnsey, 1995; Sturt et al., 2004), those in topical or focused sentential positions are more likely to be selected as antecedents of pronouns (e.g., Arnold, 1998; Kaiser, 2011; Rohde and Kehler, 2014; Colonna et al., 2012), and discourse topic (or Question under Discussion) modulates the ease of comprehension (e.g., Clifton and Frazier, 2012;

Kehler and Rohde, 2017; Clifton and Frazier, 2018) and the resolution of syntactic ambiguity (e.g., Kehler, 2015). Additionally, these distinctions have immediate effects on processing, with studies demonstrating their active use in real-time language comprehension. For instance, when linguistic materials known to lead to processing difficulty (e.g., long embedded relative clauses) are part of less important discourse, they result in reduced processing difficulty (Dillon et al., 2014, 2017; Kroll and Wagers, 2019; Duff et al., 2023).

2.3 Language model sensitivity to discourse

While the linguistic evaluations of language models have been dominated by syntactic contrasts (for a survey, see Chang and Bergen, 2024), there has been a growing body focusing on discourse knowledge. This includes work on the interaction of discourse structure and pronouns (e.g., Davis and van Schijndel, 2020), discourse structure and at-issueness (e.g., Kim et al., 2022), implicatures and presuppositions (e.g., Jeretic et al., 2020), and discourse connectives (e.g., Cong et al., 2023; Pandia et al., 2021). Broadly, pre-trained language models appear to capture some contextual effects. However, there are still notable differences between model and human behavior, suggesting differences in their processing of discourse. More recently, the impact of instruction-tuning on the linguistic knowledge of models has been investigated, with some results showing that such fine-tuning results in models with a worse fit to human behavioral measures (Kuribayashi et al., 2024). Moreover, the exact fine-tuning strategy directly impacts the ability of models on discourse tasks, with some strategies yielding models with better pragmatic abilities (Ruis et al., 2024). These results suggest that, while instruction-tuning was proposed to align models with human discourse preferences, it may not always align with the linguistic behavior of humans. The present study finds additional support for this misalignment.

3 Metrics

3.1 Interference effect

To evaluate model performance on its discourse sensitivity, we compare the *interference effect*, a common way to show the cognitive process that underlies human language comprehension (Van Dyke and Lewis, 2003; Lewis and Vasishth, 2005). For example, the interference effect is observed in the different degrees of acceptance of the two sentences in (3), even when both are ungrammatical. Studies have found that (3b) is considered more acceptable than (3a), and reading times at the verb (*were*) are commonly found to be faster in (3b) compared to (3a) (Wagers et al., 2009; Parker and An, 2018, a.o). Such a difference between the two ungrammatical sentences derives from the interfering linguistic unit, *the girl(s)*, where the plural (number) feature of *the girls* matches the feature of the verb (*were*)—leading to a *number agreement attraction effect*.

(3) a. *The waitress who sat near *the girl* were unhappy.
b. *The waitress who sat near *the girls* were unhappy.

Empirical findings suggest that this effect primarily occurs in ungrammatical sentences—when the subject and verb do not match (e.g., *the waitress... were* instead of *was*) (e.g., Wagers et al., 2009; cf. Jäger et al., 2017)—commonly referred to as the *standard number agreement attraction effect*.

Interference effect in human reading times In this study, we use the number agreement attraction effect as a signal of human processing, typically measured by the difference in reading time (RT) between the singular (3a) and plural (3b) conditions, subtracting the singular from the plural condition:

$$\text{Interference effect} = RT_{\text{plural}} - RT_{\text{singular}} \quad \text{(Eq. 1)}$$

Interference effect in models Following previous work (e.g., van Schijndel and Linzen, 2021), we used language model surprisal in correspondence to human reading time. Surprisal (Hale, 2001) is defined as (Eq. 2), calculated at the critical verb position given prior context left of the verb. The surprisal was calculated from the logits of the model.[1]

$$\text{Surprisal} = -\log P(\text{verb} \mid \text{left context}) \quad \text{(Eq. 2)}$$

In examining the number agreement attraction effect, we evaluated the difference in surprisal at the verb between the context with a plural distractor (i.e., plural context) and a singular distractor (i.e., singular context), as shown in (Eq. 3). For example, the difference in surprisal for the verb *were* when the left context was *The waitress who sat near the*

[1]For GODEL-based models, we calculated the surprisal using the decoder of the encoder-decoder model.

girls and when the left context was *The waitress who sat near the girl* was calculated.

$$\Delta\text{Surprisal(verb)} = \\ -\log P(\text{verb} \mid \text{plural context}) \\ -(-\log P(\text{verb} \mid \text{sigular context})) \quad \text{(Eq. 3)}$$

For each model, the presence or absence of the effect was determined by comparing the bootstrapped 95% confidence interval (CI) of the average interference effect as in (Eq. 4).[2]

$$\text{Average Interference Effect} = \\ \frac{1}{N}\sum_{i=1}^{N}\Delta\text{Surprisal}_i(\text{verb}), \quad \text{(Eq. 4)}$$

where N is the number of samples.

The absence of an interference effect was determined by whether the CI overlapped with zero (i.e., there was no difference between the plural and singular conditions).

3.2 Evaluation of discourse sensitivity

The attraction effect has been reported to be robustly found when the distractor noun is linearly close to the verb (*the girls ... were* as in (4a)) and even when it is distant (*the musicians ... praise* as in (5a)) (e.g., Wagers et al., 2009). Studies have further found that the attraction effect, however, can be modulated by the discourse status of the distractor noun, where in one case, the standard attraction effect disappears (4b) (Ng and Husband, 2017; McInnerney and Atkinson, 2020; Duff et al., 2023; Kim and Xiang, 2024) but it sustains in the other (5b) (Kim and Xiang, 2024).

(4) a. *The waitress who sat near *the girls* were unhappy.
b. *The waitress, who sat near *the girls*, were unhappy.

(5) a. **The musicians* who the reviewer praise highly will win a Grammy.
b. **The musicians*, who the reviewer praise highly, will win a Grammy.

When the distractor (e.g., *the girls*) is part of secondary information as in (4b), it does not interfere when the subject-verb dependency needs to be resolved, and hence the number agreement attraction effect is absent. On the contrary, when the distractor (e.g., *the musicians*) is related to the discourse topic (or Question under Discussion as in Roberts (2012)) at retrieval (e.g., *praise*) in (5b)), the distractor interferes and leads to a number agreement attraction effect (Kim and Xiang, 2024). This modulation of the interference effect due to the discourse status of the distractor noun will be used as a signal for *discourse sensitivity*.

Discourse sensitivity in human reading times The key aspects of discourse sensitivity in humans in interference effects are summarized in Table 1. First, in both constructions (Experiments 1 and 2), a standard attraction effect is found with the baseline RRC condition. This is identified by significant reading differences between the singular and plural distractor conditions in the ungrammatical condition but not in the grammatical condition (Eq. 5).

$$\text{Standard number agreement attraction effect} = \\ \begin{cases} RT_{\text{plural}} - RT_{\text{singular}} < 0 & \text{if ungrammatical,} \\ RT_{\text{plural}} - RT_{\text{singular}} \simeq 0 & \text{if grammatical.} \end{cases} \quad \text{(Eq. 5)}$$

Secondly, the standard attraction effect should be present in structures as in (4b) (Experiment 1) but absent in structures as in (5b) (Experiment 2).

Discourse sensitivity in models Using the above-mentioned human reading time results identifying discourse sensitivity as a baseline (Kim and Xiang, 2024), we evaluate model outputs based on the three following tests:[3]

- **Discourse Attraction**. In Experiment 1, the standard attraction effect is exhibited in the RRC structure (4a) but not in the ARC structure (4b). For RRCs, the average difference in surprisal between the plural distractor and the singular distractor should be negative, indicating that plural distractors lower the surprisal of plural verbs. For ARCs, the difference should not differ significantly from zero.

- **Standard Attraction**. In Experiment 2, the standard effect is exhibited in both the RRC (5b) and ARC (5b) structures. The average difference in surprisal should be negative, indicating that plural distractors lower the surprisal of plural verbs. More specifically, we divide this test into two subcases. With

[2] Bootstrapping was done with 1000 samples and resampling.

[3] For code and data: https://github.com/sangheek16/discourse-sensitivity-attraction-effect.git.

Exp.	Clause	Grammaticality	Input (subject-verb is **bold-faced**; distractor is underlined)	Effect
1	RRC	Grammatical	**The waitress** who sat near the girl(s) **was** unhappy.	✗
1	RRC	Ungrammatical	**The waitress** who sat near the girl(s) **were** unhappy.	✓
1	ARC	Grammatical	**The waitress**, who sat near the girl(s), **was** unhappy.	✗
1	ARC	Ungrammatical	**The waitress**, who sat near the girl(s), **were** unhappy.	✗
2	RRC	Grammatical	The musician(s) who **the reviewer praises** will win a Grammy.	✗
2	RRC	Ungrammatical	The musician(s) who **the reviewer praise** will win a Grammy.	✓
2	ARC	Grammatical	The musician(s), who **the reviewer praises**, will win a Grammy.	✗
2	ARC	Ungrammatical	The musician(s), who **the reviewer praise**, will win a Grammy.	✓

Table 1: Human baseline: presence (✓) vs. absence (✗) of interference effect (Kim and Xiang, 2024).

the stronger version of this test (**Standard Attraction-Strong**), the magnitude of the interference effect between RRC and ARC should be comparable. In the weaker version (**Standard Attraction-Weak**), the size of the interference effect does not matter as long as both exhibit an attraction effect.

- **Grammatical Asymmetry**. As a signal for a standard number agreement attraction effect, there should be no interference effect (i.e., no difference based on whether the distractor is singular or plural) in all grammatical conditions, regardless of clause type and experiment.[4]

4 Model selection

We tested 25 models for evaluation, either base-trained or involving instruction-based tuning, and varying in size and type of training data. Models are categorized by their specifics, summarized in Table 2.

Categorization 1: Based on the number of parameters First, we compared models that vary in their number of parameters. Specifically, we examined whether larger models yield performance similar to that of humans. Given the current emphasis in the field of scale, we might straightforwardly predict that bigger models are more likely to pass the tests. However, some empirical results suggest that scale does not necessarily mean better prediction of human behavior (e.g., Oh and Schuler, 2023; Oh et al., 2024; Shain et al., 2024; Wilcox et al., 2024). As Oh et al. suggest, LLMs make good predictions on words with low frequency, which in turn is not what is expected in human data. If the same type of counter-advantage of large models applies to examining discourse sensitivity, then we could see better performance with smaller models.

Categorization 2: Based on the genre of data Second, we also examine whether the genre of data would affect the performance of LMs in discourse sensitivity. Earlier work has shown that models outperform others in dialogue and discourse settings when trained on data with conversation and naturalistic data (Wolf et al., 2019; Bao et al., 2020; Henderson et al., 2020; Wu et al., 2020; Zhang et al., 2020; Gu et al., 2021; Thoppilan et al., 2022). We acknowledge that the comparison of the genre of data between models is not totally straightforward, especially given the lack of accessibility to LLMs' training data. For example, the training data used for some LLMs may include the data used for the "dialogue-based models." However, we believe a useful comparison can still be sustained. If the dialogue-based models outperform the models trained on a variety of genres, then we take this as evidence that training data primarily composed of discourse-goal and dialogue-oriented data is of better quality, for alignment with human linguistic behavior, than a larger composition of varied genres and styles.

Categorization 3: Based on tuning/training type Finally, we also examine whether instruction models outperform base models in discourse sensitivity.

[4]We acknowledge prior findings showing that the grammatical asymmetry in attraction effects—typically observed in ungrammatical conditions—can be influenced by task factors such as response bias and answer ratios (e.g., Hammerly et al., 2019; Laurinavichyute and von der Malsburg, 2024). These studies found that the asymmetry is masked when the response bias is neutralized. However, in the study by Kim and Xiang (2024), which provides the human reading time data used for model evaluation in the current work, the task was explicitly designed to neutralize response bias. Therefore, while we recognize this as a general concern in the literature, we suspect it is less likely to impact the interpretation of our current results.

Given that instruction-tuned models are arguably better at capturing the user's (or the interlocutor's) needs and goals (see Zhang et al. (2023) for an overview), we speculate that models could benefit from such training methods to achieve better performance in discourse sensitivity, similar to understanding discourse goals. They could demonstrate patterns that align well with human expectations in discourse and dialogue settings. Yet, there is only little work on investigating how well instruction-based models align with human behavior. While instruction tuning can result in greater alignment at the high-level representation (e.g., between the LLM internal representation and human neural activity, see Aw et al., 2024), findings also suggest that at the behavioral level, there is no model-human alignment such as in human reading times or judgment tasks (Zhang et al., 2023; Kauf et al., 2024; Aw et al., 2024). Given that discourse sensitivity in the current work is measured through surprisal and is compared against human reading time data, it is possible that instruction-based models would not outperform the base models.

5 Results

Table 2 shows the list of models we evaluated and the results.

5.1 By each test

Discourse Attraction. With only one exception of DialoGPT-small, all dialogue-based models passed this test. While GPT-Neo-125M, GPT-Neo-2.7B, and Mistral-7B-v0.3 passed Discourse Attraction, the remaining models did not, showing no systematic correlation with training type or size.

Standard Attraction. All but three models passed Standard Attraction-Weak. The models that did not pass this test are all small dialogue-based models: DialoGPT-large, GODEL-base, and GODEL-large. When a stronger version (Standard Attraction-Strong) was applied, four additional models failed to pass: GPT-J-6B, Mistral-7B-v0.1, Mistral-7B-v0.3, and Mistral-7B-Instruct-v0.3.

Grammatical Asymmetry. None of the models passed Grammatical Asymmetry. All models exhibited an interference effect in the grammatical condition of at least one of the clause types in at least one of the experiments.

5.2 By combined tests

To better understand the results, we analyze them by each of the four combinations that can be found in passing Discourse Attraction and Standard Attraction. The models' failure to pass Grammatical Asymmetry is discussed in Section 6.

Discourse Attraction:✓, Standard Attraction:✓. This is a case where models were most sensitive to discourse division. Passing both of these tests signals a division of primary versus secondary information driven by the syntactic difference between RRC and ARC structures—as in Discourse Attraction—while not simply making distinctions between RRC and ARC structures based on their syntactic form—as in Standard Attraction. *Models*: DialoGPT-medium, GPT2-small, GPT-Neo-125M, and GPT-Neo-2.7B. The models that passed both of these tests (Discourse Attraction and Standard Attraction-Strong) were all small GPT-based models.

Discourse Attraction:✓, Standard Attraction:✗. This is a case where models were sensitive to the division between RRC and ARC and were applying the same division to resolving the linguistic dependency in Experiment 2. However, as we have seen in human performance, it is not simply the syntactic division between RRC and ARC to pass Standard Attraction; the interference effect with the ARC condition that was absent in Experiment 1 should be present in Experiment 2. The models under this category did not exhibit that contrast, suggesting that while they have grasped the syntactic division, the nuanced discourse division was not captured. *Models*: DialoGPT-large, GODEL-base, GODEL-large. These models were exclusively small, dialogue-based models.

Discourse Attraction:✗, Standard Attraction:✓. This is a case where models exhibited an interference effect in both experiments in both clauses. While all models showed the baseline interference effect in the RRC condition, the failure to pass Discourse Attraction was driven by the presence of the interference effect in the ARC condition. The results can be interpreted in that while the models showed an interference effect, they lacked discourse or syntactic division. *Models*: DialoGPT-small, GPT2-medium, GPT2-large, GPT2-XL, GPT-Neo-1.3B, GPT-J-6B, Llama-2-7B, Llama-2-13B, Llama-3-8B, Llama-3.1-8B, Mistral-7B-v0.1, Llama-2-7B-Chat, Llama-2-13B-Chat, Llama-3-

Training Type	Size	Model	Size	Discourse	Standard-weak	Standard-strong	Grammatical
Dialogue	Small	DialoGPT-small	117M	✗	✓	✓	✗
		DialoGPT-medium	345M	✓	✓	✓	✗
		DialoGPT-large	762M	✓	✗	✗	✗
		GODEL-base	220M	✓	✗	✗	✗
		GODEL-large	770M	✓	✗	✗	✗
Plain		GPT2-small	124M	✓	✓	✓	✗
		GPT2-medium	355M	✗	✓	✓	✗
		GPT2-large	774M	✗	✓	✓	✗
		GPT2-XL	1.5B	✗	✓	✓	✗
		GPT-Neo-125M	125M	✓	✓	✓	✗
		GPT-Neo-1.3B	1.3B	✗	✓	✓	✗
		GPT-Neo-2.7B	2.7B	✓	✓	✓	✗
		GPT-J-6B	6B	✗	✓	✗	✗
	Large	Llama-2-7B	7B	✗	✓	✓	✗
		Llama-2-13B	13B	✗	✓	✓	✗
		Llama-3-8B	8B	✗	✓	✓	✗
		Llama-3.1-8B	8B	✗	✓	✓	✗
		Mistral-7B-v0.1	7B	✗	✓	✗	✗
		Mistral-7B-v0.3	7B	✓	✓	✗	✗
Instruction		Llama-2-7B-Chat	7B	✗	✓	✓	✗
		Llama-2-13B-Chat	13B	✗	✓	✓	✗
		Llama-3-8B-Instruct	8B	✗	✓	✓	✗
		Llama-3.1-8B-Instruct	8B	✗	✓	✓	✗
		Mistral-7B-Instruct-v0.1	7B	✗	✓	✓	✗
		Mistral-7B-Instruct-v0.3	7B	✗	✓	✗	✗

Table 2: Model comparison: passed (✓) vs. failed (✗) the test.

8B-Instruct, Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.1, Mistral-7B-Instruct-v0.3. These include all the instruction-based models, most of the large models, and most of the small plain models.

Discourse Attraction:✗, Standard Attraction:✗. This would be the case where models demonstrated no interference effect. No model exhibited this behavior, which confirms that they were influenced by the presence of a distractor noun in at least some conditions. All models demonstrated the baseline effect of interference in the ungrammatical RRC condition. *Models*: None.

6 Discussion

No models passed all three tests. However, all models were influenced by distractors, facilitating the use of interference effects to test whether discourse structure influenced model's predictions. While the models did not pass all three tests, they showed systematicity in their performance on Discourse Attraction and Standard Attraction. In one case ({Discourse Attraction: ✗, Standard Attraction: ✓}), the presence of a distractor led to an interference effect, but this effect was not modulated by discourse division. In the other case ({Discourse Attraction: ✓, Standard Attraction: ✗}), the models were guided by the discourse, or the syntactic division of RRC and ARC, but they were overapplying this division. Four of the tested models performed in the most principled way, where they passed Discourse Attraction and Standard Attraction: DialoGPT-medium, GPT2-small, GPT-Neo-125M, and GPT-Neo-2.7B. In the following, we elaborate on the less principled models.

Lack of discourse division (Discourse: ✗, Standard: ✓). This is a systematic pattern where the models show the standard number agreement attraction effect without showing sensitivity to discourse division. Models showed the attraction effect in all constructions in (4)–(5), indicating that the different discourse status of the distractor in (5b) was not considered. This pattern was prevalent in most

of the models, except for the small dialogue-based models. This is in line with earlier studies that have shown cases where grammatically irrelevant words modulate the surprisal at the critical word (in subject-verb agreement (Ryu and Lewis, 2021; Arehalli and Linzen, 2020) as well as reflexive pronoun resolution (Ryu and Lewis, 2021; Davis, 2022a). The influence of linearly closer, but grammatically irrelevant words, remains a feature of even the larger models. That is, increases in scale and other training approaches have not made models robust to interference effects.

Heavy reliance on syntactic/discourse division (Discourse: ✓, Standard: ✗). In line with the finding discussed above, it is still the case that all models under this category have exhibited the standard number agreement attraction effect in the baseline RRC condition (as in (4a) & (5a)). Nonetheless, the effect was not present with the ARC structure in both Experiment 1 (as in (4b)) and Experiment 2 (as in (5b)), suggesting that it is possible that models heavily relied on the linguistic cue that distinguishes the main content from the subordinate content in the sentences with the ARC structure. Earlier work using a probing task showed that LMs successfully classify (with greater than 99% accuracy) the main content differently from the subordinate content (Kim et al., 2022). Hence, it is possible that the structural difference (or even simply the presence of commas) of ARCs compared to RRCs has resulted in the absence of the attraction effect.

However, there is another possibility beyond the models tracking the superficial cues or the syntactic representation: the models were (overly) applying discourse division cues. Recall that the only three models that fell under this category are DialoGPT-large, GODEL-base, and GODEL-large, all trained on dialogue-based data. We conjecture that it is not coincidental that the overapplication of the division of main versus subordinate content to attraction effect was only found in the dialogue-based models. We speculate that the specific training process has led to an effect of models exhibiting abstract signals about discourse structure, either (a) naturally following from the abstract structural representation through training, or (b) demonstrating a separate pattern that is learned in addition to the abstract structural representation.

Given the promising performance of recent instruction-based models, it is perhaps unexpected that they fall short in exhibiting the level of discourse sensitivity in humans. This discrepancy may stem from the training methods of these instruction-based models, which are optimized for extracting and producing the most relevant information efficiently and concisely. During training, they are directed to perform tasks such as summarization and a clear question and answering (Zhang et al., 2023). However, human discourse includes purposeful digressions—often for the richness of conversation—and layers of primary (main) and secondary (subordinate) information. The different conversational goal perhaps accounts for the reason why instruction-based models diverge from the discourse division that humans show.

Why didn't any of the models pass Grammatical Asymmetry? Grammatical Asymmetry examined whether models exhibit the standard number agreement attraction effect, i.e., whether the attraction effect is found only in the ungrammatical and not in the grammatical condition. One of the ways to account for the asymmetric attraction effect in humans is an error-driven process, where the interference effect is realized only when there is a mismatch between the retrieval target (i.e., subject) and the retrieval site (i.e., verb)—that is, when the sentence is ungrammatical (Wagers et al., 2009; Lago et al., 2015; Schlueter et al., 2019). However, such an error-driven process seems unlikely for the models. As we saw in the results with Grammatical Asymmetry, the presence of the distractor in the subject-verb dependency led to an attraction effect, even when the subject and the verb agreed—that is, when the sentence was grammatical, and hence there were no "errors."

The contrast between human and model performance has implications for interpreting the models, where the distractor does not have an equal status in language processing. Humans may be applying a top-down approach (by incorporating the discourse status of distractors) (e.g., Kutas et al., 2011) while incorporating bottom-up linguistic information (Momma and Phillips, 2018) (such as number information). While prediction and expectation on the verb that agrees with the subject are formed in real time in humans, models are strongly driven by a bottom-up incremental process, where the linear sequence of the incoming linguistic units is influential on the retrieval process.

7 Conclusion

The current work examined the discourse sensitivity of language models by investigating the interaction between discourse structure and syntactic dependency. Leveraging findings from human experiments on the number agreement attraction effect, we compared language model behavior to human behavior. Critically, the pattern we targeted was the presence of a standard attraction effect in Experiment 1 (Discourse Attraction), its absence in Experiment 2 (Standard Attraction), and the presence of a grammatical asymmetry (Grammatical Asymmetry). As discussed in Kim and Xiang (2024), humans show a modulated attraction effect across the two experimental setups, driven by their sensitivity to the active discourse question (akin to the Question Under Discussion).[5] None of the 25 models fully overlapped with humans: some models associated structural cues with discourse, while others overapplied discourse cues. Larger models exhibited the attraction effect in both Experiment 1 and Experiment 2, indicating insensitivity to the nuanced discourse status of distractor and target noun phrases. In contrast, smaller models trained on dialogue-based data showed the best performance—even outperforming large, instruction-based models. These smaller models exhibited a modulated attraction effect, suggesting they may have learned some abstract representation of discourse, though not fully matching human retrieval patterns, as shown by their failure in Grammatical Asymmetry. As discussed in the Discussion section, we conjecture that larger models may underperform relative to smaller models in capturing human-like patterns due to the scale of their training data. Furthermore, instruction-tuned models may lack alignment with human discourse goals and conversational dynamics given their training objective.

Future work could solidify these claims by surveying a larger variety of instruction-tuning approaches and carefully controlling the training data to tease apart the effect of data quality on model performance (as in Misra and Mahowald, 2024). Ultimately, the contrast between language processing in humans and machines highlights a disconnect in their abilities to integrate multiple sources of information. While humans combine syntactic and discourse information, and top-down and bottom-up linguistic signals, models overrely on one of these sources.

[5] See Kim and Xiang (2024) for a detailed explanation of how the discourse question modulates retrieval processes that leads to the observed attraction effect.

8 Limitations

The current study used a discrete categorization based on the absence or presence of the number agreement attraction effect. While this approach offers ease of interpretation, we acknowledge that it limits the ability to perform more quantitative evaluations. Future work could adopt a quantitative approach to compare the magnitude of the attraction effect in human reading times and surprisal across experiments. Furthermore, we focused on one specific case study to investigate models' discourse sensitivity, rather than a suite of tests. As such, the conclusions drawn from the current findings may be limited to this particular form of discourse sensitivity. The authors are developing broader tests to evaluate discourse sensitivity beyond the modulated attraction effect to assess the generalizability of the current findings.

Acknowledgments

We would like to thank Marten van Schijndel and the anonymous reviewers for their constructive comments and suggestions.

References

Scott AnderBois, Adrian Brasoveanu, and Robert Henderson. 2015. At-issue proposals and appositive impositions in discourse. *Journal of Semantics*, 32(1):93–138.

Suhas Arehalli and Tal Linzen. 2020. Neural language models capture some, but not all, agreement attraction effects. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, pages 370–376.

Jennifer E. Arnold. 1998. *Reference form and discourse patterns*. Ph.D. thesis, Stanford University, Stanford, CA.

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge: Cambridge University Press.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. 2024. Instruction-tuning aligns LLMs to the human brain. In *First Conference on Language Modeling*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association*

for Computational Linguistics, pages 85–96, Online. Association for Computational Linguistics.

Stacy L Birch and Susan M Garnsey. 1995. The effect of focus on memory for words in sentences. *Journal of Memory and Language*, 34(2):232–267.

Tyler A. Chang and Benjamin K. Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350.

Charles Jr. Clifton and Lyn Frazier. 2012. Discourse integration guided by the 'Question under Discussion'. *Cognitive Psychology*, 65(2):352–379.

Charles Jr. Clifton and Lyn Frazier. 2018. Context effects in discourse: The question under discussion. *Discourse Processes*, 55(2):105–112.

Saveria Colonna, Sarah Schimke, and Barbara Hemforth. 2012. Information structure effects on anaphora resolution in German and French: A crosslinguistic study of pronoun resolution. *Linguistics*, 50(5):991–1013.

Yan Cong, Emmanuele Chersoni, Yu-Yin Hsu, and Philippe Blache. 2023. Investigating the effect of discourse connectives on transformer surprisal: Language models understand connectives, Even so they are surprised. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 222–232, Singapore. Association for Computational Linguistics.

Forrest Davis. 2022a. Incremental processing of Principle B: Mismatches between neural models and humans. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 144–156.

Forrest Davis. 2022b. *On the Limitations of Data: Mismatches between Neural Models of Language and Humans*. Ph.D. thesis, Cornell University.

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

Brian Dillon, Charles Clifton Jr., and Lyn Frazier. 2014. Pushed aside: Parentheticals, memory and processing. *Language, Cognition and Neuroscience*, 29(4):483–498.

Brian Dillon, Charles Clifton Jr., Shayne Sloggett, and Lyn Frazier. 2017. Appositives and their aftermath: Interference depends on at-issue vs. not-at-issue status. *Journal of Memory and Language*, 96:93–109.

John Duff, Pranav Anand, Adrian Brasoveanu, and Amanda Rysling. 2023. Pragmatic representations and online comprehension: Lessons from direct discourse and causal adjuncts. *Glossa Psycholinguistics*, 2(1):1–52.

Donka F Farkas and Kim B Bruce. 2010. On reacting to assertions and polar questions. *Journal of Semantics*, 27(1):81–118.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Alexander Göbel. 2019. Final appositives at the right frontier: An experimental investigation of anaphoric potential. In *Proceedings of Sinn und Bedeutung 23*, pages 451–467. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès).

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2021. DialogBERT: Discourse-aware response generation via learning to recover and rank utterances. *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, 35(14):12911–12919.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Christopher Hammerly, Adrian Staub, and Brian Dillon. 2019. The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. *Cognitive Psychology*, 110:70–104.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online. Association for Computational Linguistics.

Jerry R. Hobbs. 1985. *On the coherence and structure of discourse*. Stanford, CA: CSLI Technical Report 85-37.

Jennifer Hu, Sherry Yong Chen, and Roger Levy. 2020. A closer look at the performance of neural language models on reflexive anaphor licensing. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 323–333, New York, New York. Association for Computational Linguistics.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Lena A Jäger, Felix Engelmann, and Shravan Vasishth. 2017. Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, 94:316–339.

Katja Jasinskaja. 2016. Not at issue any more. Unpublished manuscript, University of Cologne.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Elsi Kaiser. 2011. Focusing on pronouns: Consequences of subjecthood, pronominalisation, and contrastive focus. *Language and Cognitive Processes*, 26(10):1625–1666.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. 2024. Comparing plausibility estimates in base and instruction-tuned large language models. *arXiv preprint arXiv:2403.14859*.

Andrew Kehler. 2002. *Coherence, reference, and the theory of grammar*. Stanford, CA: CSLI Publications.

Andrew Kehler. 2015. On QUD-based licensing of strict and sloppy ambiguities. In *Semantics and Linguistic Theory (SALT)*, pages 512–532.

Andrew Kehler and Hannah Rohde. 2017. Evaluating an expectation-driven question-under-discussion model of discourse interpretation. *Discourse Processes*, 54(3):219–238.

Sanghee J Kim and Ming Xiang. 2024. Incremental discourse-update constrains number agreement attraction effect. *Cognitive Science*, 48(9):e13497.

Sanghee J Kim, Lang Yu, and Allyson Ettinger. 2022. "No, they did not": Dialogue response dynamics in pre-trained language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 863–874, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Todor Koev. 2022. *Parenthetical meaning*. Oxford Studies in Semantics and Pragmatics. Oxford: Oxford University Press.

Margaret Kroll and Matthew W. Wagers. 2019. Working memory resource allocation is not modulated by clausal discourse status. Unpublished manuscript, University of California, Santa Cruz.

Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. 2024. Psychometric predictive power of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1983–2005, Mexico City, Mexico. Association for Computational Linguistics.

Marta Kutas, Katherine A DeLong, and Nathaniel J Smith. 2011. A look around at what lies ahead: Prediction and predictability in language processing. In Moshe Bar, editor, *Predictions in the Brain: Using our Past to Generate a Future*. New York, NY: Oxford University Press.

Sol Lago, Diego E Shalom, Mariano Sigman, Ellen F Lau, and Colin Phillips. 2015. Agreement attraction in spanish comprehension. *Journal of Memory and Language*, 82:133–149.

Nur Lan, Emmanuel Chemla, and Roni Katzir. 2024. Large language models and the argument from the poverty of the stimulus. *Linguistic Inquiry*, pages 1–28.

Anna Laurinavichyute and Titus von der Malsburg. 2024. Agreement attraction in grammatical sentences and the role of the task. *Journal of Memory and Language*, 137:104525.

David Lewis. 1979. Scorekeeping in a language game. *Journal of Philosophical Logic*, 8:339–359.

Richard L Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(Volume 7, 2021):195–212.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Andrew McInnerney and Emily Atkinson. 2020. Syntactically unintegrated parentheticals: Evidence from agreement attraction. The 33rd Annual CUNY Human Sentence Processing, University of Massachusetts Amherst: Amherst, MA. March 19–21 (oral presentation).

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 913–929, Miami, Florida, USA. Association for Computational Linguistics.

Shota Momma and Colin Phillips. 2018. The relationship between parsing and generation. *Annual Review of Linguistics*, 4(1):233–254.

Anne Ng and Matthew Husband. 2017. Interference effects across the at-issue/not-at-issue divide: Agreement and NPI licensing. The 30th Annual CUNY Human Sentence Processing, MIT: Cambridge, MA. March 30–April 1 (poster presentation).

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. *arXiv preprint arXiv:2402.02255*.

Lalchand Pandia, Yan Cong, and Allyson Ettinger. 2021. Pragmatic competence of pre-trained language models through the lens of discourse connectives. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 367–379, Online. Association for Computational Linguistics.

Dan Parker and Adam An. 2018. Not all phrases are equally attractive: Experimental evidence for selective agreement attraction effects. *Frontiers in Psychology*, 9:1566.

Livia Polanyi. 1988. A formal model of the structure of discourse. *Journal of Pragmatics*, 12(5–6):601–638.

Christopher Potts. 2005. *The logic of conventional implicatures*. Oxford: Oxford University Press.

Christopher Potts. 2012. Conventional implicature and expressive content. In Claudia Maienborn, Klaus von Heusinger, and Paul Portner, editors, *Semantics: An international handbook of natural language meaning*, volume 3, pages 2516–2536. Berlin: Mouton de Gruyter.

Craige Roberts. 2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69.

Hannah Rohde and Andrew Kehler. 2014. Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience*, 29(8):912–927.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2024. The goldilocks of pragmatic understanding: fine-tuning strategy matters for implicature resolution by llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Soo Hyun Ryu and Richard L Lewis. 2021. Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. *arXiv preprint arXiv:2104.12874*.

Zoe Schlueter, Dan Parker, and Ellen Lau. 2019. Error-driven retrieval in agreement attraction rarely leads to misinterpretation. *Frontiers in Psychology*, 10:1002.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Patrick Sturt, Anthony J Sanford, Andrew Stewart, and Eugene Dawydiak. 2004. Linguistic focus and good-enough representations: An application of the change-detection paradigm. *Psychonomic Bulletin & Review*, 11(5):882–888.

Michelle Suijkerbuijk, Naomi T Shapiro, Peter de Swart, and Stefan L Frank. 2024. The need for human data when analysing the human-likeness of syntactic representations in neural language models: The case of english wh-island constraints.

Kristen Syrett and Todor Koev. 2015. Experimental evidence for the truth conditional contribution and shifting information status of appositives. *Journal of Semantics*, 32(3):525–577.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. LaMBDA: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Julie A Van Dyke and Richard L Lewis. 2003. Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, 49(3):285–316.

Marten van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45(6):e12988.

Matthew W Wagers, Ellen F Lau, and Colin Phillips. 2009. Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, 61(2):206–237.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ethan Gotlieb Wilcox, Michael Hu, Aaron Mueller, Tal Linzen, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Ryan Cotterell, and Adina Williams. 2024. *Bigger is not always better: The importance of human-scale language modeling for psycholinguistics*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. TransferTransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Aditya Yedetore and Najoung Kim. 2024. Semantic training signals promote hierarchical syntactic generalization in transformers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4059–4073, Miami, Florida, USA. Association for Computational Linguistics.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.