A Psycholinguistic Evaluation of Language Models' Sensitivity to Argument Roles

Anonymous ACL submission

Abstract

We present a systematic evaluation of large language models' sensitivity to argument roles, i.e., *who* did what to *whom*, by replicating psycholinguistic studies on human argument role processing. In three experiments, we find that language models are able to distinguish verbs that appear in plausible and implausible contexts, where plausibility is determined through the relation between the verb and its preceding arguments. However, none of the models capture the same selective patterns that human comprehenders exhibit during real-time verb prediction. This indicates that language models' capacity to detect verb plausibility does not arise from the same mechanism that underlies human real-time sentence processing.

1 Introduction

Humans rapidly make predictions when comprehending language. However, certain types of information do not immediately impact predictions, and a well-studied case of this in the sentence processing literature involves argument roles.

Argument roles refer to the roles of participants that take part in the event described by a sentence, such as who is the agent (do-er of the action) and who is the patient (undergo-er of the action). One hallmark of language understanding is the capacity to compute the meanings of arguments and their roles in relation to the verb given in a sentence. Studies with human participants have shown that in contrast to the lexical meanings of arguments, the roles assigned to the arguments by the structure are not immediately used to predict an upcoming verb. For example, given the context (1a), a verb like *served* is a highly expected continuation, whereas swapping the arguments (1b) makes the same verb *served* no longer appropriate. However, despite the difference in how likely the verb is given the preceding argument role context, human comprehenders show the same initial response to a verb when it appears in a role-appropriate and role-reversed context (Kim and Osterhout, 2005; Chow et al., 2016). This has been taken to indicate that argument roles have a delayed impact on verb prediction in human sentence processing.

1. a. The customer that the waitress...
   b. The waitress that the customer...

Recent work has used paradigms from experimental psycholinguistics to evaluate language models' representation of syntactic and semantic knowledge, and language models trained on next-word prediction alone have shown strong levels of correspondence with human behavioral and neural data. However, despite extensive work on the linguistic patterns they are able to learn, the extent to which they accurately encode argument role information and utilize it in distinguishing plausible and implausible sentences remains an open question. Previous work has mostly focused on analyzing whether models can distinguish plausible or implausible sentences that involve argument role manipulations (Ettinger, 2020; Papadimitriou et al., 2022; Wilson et al., 2023a; Kauf et al., 2023), where factors such as animacy are confounded, making it difficult to precisely observe models' sensitivity to argument role information.

In this paper, we take a new approach in evaluating role-sensitivity in large language models by focusing on models' representations of verbs that appear in either plausible or implausible sentences, where plausibility is determined based on the verb's relation with the preceding argument-role bindings enforced by the context.

We adapt materials used in psycholinguistic studies evaluating humans' sensitivity to argument roles, which allows us to use carefully constructed minimal pairs of sentences which only differ with respect to argument roles, while controlling for other factors like animacy. This serves as a rigorous test in examining models' ability to extract argument-role bindings based on sentence structure, as it requires models to go beyond simply learning relations between various arguments and verbs, i.e., between real-world events and participants that are likely to be involved in those events. We compare model performance on two different types of argument role manipulations, in addition to a baseline condition which has shown to elicit immediate sensitivity in humans, as a way to more systematically compare human and model behavior.

Through three experiments, we find that i) language models show weak sensitivity to argument role information relative to role-independent argument meanings, similar to human initial predictions, ii) models do not show the same consistency across different types of argument role manipulations as humans do, indicating a difference in the way argument roles are processed in models and humans, and iii) weak performance does not necessarily arise from a misrepresentations of arguments. These results overall indicate that even if models are able to distinguish plausible and implausible verbs to varying degrees of success, this lack of generalization across the conditions that share a structural relation suggests that the models do not use the same mechanism as humans to compute argument-verb relations.

2 Related Work

To evaluate language models' representations of argument roles, reversing the order of the verb's arguments is a common design, paralleling the stimuli in human experiments. Researchers then compare differences in the reversed and felicitous conditions, using various metrics from the models. There are two major issues with existing work that we address. First, existing work often relies on the animacy of the verbs' arguments. Second, work using different metrics often offer conflicting conclusions.

Papadimitriou et al. (2022) claim language models are able to effectively make use of word order-related information when arguments are switched for verbs with transitive subjects and objects, reflecting these distinctions imposed by selectional constraints on the verb in their representations. For instance, the models they evaluated would represent *The chef chopped the onion* differently from *The onion chopped the chef*. For this evaluation, they automatically switch the order of arguments in naturalistic corpora. Thus, it is unclear if these positive results are based on properties of the lexical items (i.e. frequency, animacy) that are learned more easily from distributional information, or more abstract representations of argument roles.[1] A more reliable way to measure the linguistic capacity of language models is to effectively treat them as psycholinguistic subjects (Futrell et al., 2019; Ettinger, 2020, among others) across a range of configurations (see reviews by Linzen and Baroni (2021); Pavlick (2022) and Mahowald et al. (2024)). Work in this vein presents models with minimal pairs of sentences and analyzes differences in language models' responses to each sentence. Language models' sensitivity to a variety of phenomena been evaluated with this paradigm (Linzen et al., 2016; Warstadt et al., 2020; Wilcox et al., 2023b). Looking at argument roles specifically, Kauf et al. (2023) find they are able to distinguish plausible events from implausible ones, assigning higher probabilities to sentences like *The teacher bought the laptop.* as opposed to *The laptop bought the teacher.*, but only when one participant is animate and the other is inanimate. Given the ability of language models to handle animacy even in atypical settings (Hanna et al., 2023), it is possible that the results of both Kauf et al. (2023) and Papadimitriou et al. (2022) may be tapping into this ability rather than a generalized representation of argument roles.

Ettinger (2020) presented a suite of psycholinguistically motivated diagnostics for BERT; one of these tests was on *argument role reversals*, which was similar in spirit to some of Kauf et al. (2023)'s stimuli but only tested animate participants. This study had different conclusions, finding that BERT was indeed sensitive to these role-related contrasts, generating role reversals in appropriate contexts, but not on par with humans. Working with this dataset, Li et al. (2021) evaluate the probabilities the models assign to the sentence at individual layers and finds that they are not sensitive to the role reversal sentences. These studies all use different methods of evaluation. Ettinger (2020) queried sentence completions made by BERT, while Kauf et al. (2023) determined whether the language models assigned lower probabilities to the implausible sentence of the pair.

We take a different approach to examine language models' sensitivity to argument roles by

[1] If such generalizations exist, they are largely tied to the presence of surface forms in the training data (Wilson et al., 2023b).

replicating psycholinguistic experiments with multiple conditions designed to isolate humans' representations of argument roles. These experiments track human processing in real time and specifically examine participants' responses to verbs, which reflect how the representation of the sentence is built up. To tighten the link to whether models are making human-like judgments, we also examine the models' responses to the verbs rather than sentence-level metrics through behavioral and representational methods in Experiments 1 and 2.

Furthermore, one reason why Transformers are hypothesized to capture many empirical patterns in human sentence processing is that their attention mechanisms are able to efficiently keep track of long distance dependencies (Ryu and Lewis, 2021). Despite findings localizing handling certain syntactic dependencies to individual attention heads (Clark et al., 2019; Vig and Belinkov, 2019; Jian and Reddy, 2023), little work has been done on connecting these measures to psycholinguistic findings. Ryu and Lewis (2021) specifically found an attention head that handled subject-verb agreement in GPT2, which corresponded with human processing of these dependencies. This approach has not been tried for argument roles in a more generalized setting.[2] We do so in Experiment 3.

3 Psycholinguistic Data

We use materials from previous psycholinguistic experiments which were carefully constructed to evaluate human comprehenders' sensitivity to argument roles in real-time sentence processing. These stimuli sets were designed to compare electrophysiological responses to verbs that appeared in different sentence contexts, and the different conditions have shown to elicit distinct N400 amplitudes, a neural response taken to reflect how strongly a target word was predicted based on the previous context (Kutas and Hillyard, 1980).

We use the materials from Chow et al. (2016) and Kim and Osterhout (2005), and label the conditions as `swap-arguments`, `change-verb`, and `replace-argument` (Table 1). Both studies were conducted in English on native speakers.

Both the `swap-arguments` and `change-verb` conditions include manipulations of argument roles and verb plausibility. In the `swap-arguments` condition, the two arguments preceding the verb in the plausible sentence are swapped to create the implausible sentence. In the `change-verb` condition, the verb form is changed to create the plausible and implausible sentences. Although the two conditions involve different changes, both have the same consequence: verb plausibility changes because of the way the argument(s) are assigned different roles, while the argument(s) that appear in the context remain the the same (e.g., *waitress-customer*, *meal*). In addition to the two role-related conditions, we also include a `replace-argument` condition (taken from Chow et al. (2016)), which involves replacing one of the arguments with an entirely different noun. This results in changing the argument meaning rather than argument roles, and this has shown to yield immediate predictability effects in human verb predictions, as opposed to the previous two conditions which both fail to elicit rapid sensitivity.

The key human empirical pattern to which we compare language models' behavior is the relatively weak sensitivity to argument roles (`swap-arguments` & `change-verb`) compared to argument meanings (`replace-argument`).

4 Models & Experiments

We use the following pre-trained language models for our analyses: GPT2 (small, medium, and large) (Radford et al., 2019), BERT (base-uncased, large-uncased) (Devlin et al., 2019), and RoBERTa (base, large) (Liu et al., 2019). Details of the model properties are included in Appendix A. All models were accessed through the `transformers` (Wolf et al., 2020) or `minicons` library (Misra, 2022), built to work with the Huggingface API. Code and data will be made publicly available upon acceptance.

We carry out three experiments, evaluating language models' ability to differentiate plausible and implausible verbs given the sentence. We specifically focus on addressing the following questions: (i) Do the models show a human-like pattern across the different conditions? (ii) Are these contrasts reflected in the models' representations across the intermediate layers? (iii) Do patterns in the models' attention weights reflect argument role sensitivity?

5 Experiment 1: Surprisal Effects

One of the most well-established measures linking language models to cognitive hypotheses is surprisal, or the negative log probability of a word

[2]However, see improvements from Timkey and Linzen (2023) modeling this specific case and Oh and Schuler (2023a) which shows the success of attention in modeling broad-coverage sentence processing.

Condition	Items	Plausible	Implausible
`swap-arguments`	120	The restaurant owner forgot which *customer* the *waitress* **served** during dinner yesterday.	The restaurant owner forgot which *waitress* the *customer* **served** during dinner yesterday.
`change-verb`	96	The hearty meal was **devoured** with gusto.	The hearty meal was **devouring** by the kids.
`replace-argument`	120	The secretary confirmed which *illustrator* the author had **hired** for the new book.	The secretary confirmed which *readers* the author had **hired** for the new book.

Table 1: Example sentences (1 pair = 1 item) in each condition. The `swap-arguments` and `change-verb` conditions involve argument role manipulations, while `replace-argument` serve as a control. Humans show greater sensitivity in the `replace-argument` than in the `swap-arguments` and `change-verb` conditions.

given context. Surprisal theory (Hale, 2001; Levy, 2008) states that the difficulty associated with processing linguistic information can be operationalized with this measure. Language model surprisal has shown to strongly correlate with both human reading times (Smith and Levy, 2013; Shain et al., 2024) as well as the N400 EEG response (Frank et al., 2013; Michaelov et al., 2024). Current Transformer models perform more effectively than other methods of language modeling (Merkx and Frank, 2021), and this relationship with reading times has been established cross-linguistically (Wilcox et al., 2023a).[3]

5.1 Methods

For each item, we compute the **surprisal effect** at the verb. Even if we might expect models to assign lower probability, and thus higher surprisal, to implausible continuations, it is important to determine the surprisal effect on individual items, following work on the targeted syntactic evaluation of language models (Marvin and Linzen, 2018; Wilcox et al., 2023b). This allows us to quantify not just whether the model is successfully capturing distinctions between sentences, but to what extent it is able to do so. We operationalize this effect in Equation 1, such that $context_i$ and $context_p$ are implausible and plausible versions of the same context, respectively, and S_{LM} is the language model's surprisal in Equation 2.

$$S_{LM}(verb, context_i) - S_{LM}(verb, context_p) \tag{1}$$

$$S_{LM}(w, c) = -\log_2 P_{LM}(w|c) \tag{2}$$

Verb surprisal estimates were obtained with Equation 1, and the surprisal effect for each item was obtained by subtracting the surprisal of the verb in the implausible context from the plausible context in all experimental conditions. Therefore, a positive value indicates that the model correctly assigned lower surprisal to the target verb in the plausible context relative to the implausible context, i.e., role-sensitivity, while a value close to zero or negative indicates that the model incorrectly assigned similar or greater surprisal to the verb in the plausible context than the implausible context.

5.2 Results

We report the surprisal effect in all the models in Figure 1. In line with our expectations, the surprisal effect is larger for the `replace-argument` items than the `swap-arguments` items, showing that models are less sensitive to role reversals compared to `replace-arguments`. GPT2-small in particular did not exhibit any sensitivity to the role-reversed sentences, while showing considerably more sensitivity to the `replace-argument` sentences, consistent with Chow et al. (2016). However, one key difference between the model and human responses is that all the models' effects for `change-verb` were far higher than both the `swap-arguments` and the baseline `replace-argument` case. Instead of showing a smaller effect, like for `swap-arguments`, the surprisal effect for these sentences is far higher.

The performance of GPT2-small for the `swap-arguments` condition mirrors the early stages of human processing more closely, as these role-reversed sentences do not elicit an N400 potential. However, humans are also not sensitive to the manipulation in the `change-verb` stimuli since they use an abstract, generalized representation of argument roles, which is a major contrast with the models' surprisal. Based on the compa-

[3] Correlations do not necessarily increase with language model scale (Steuer et al., 2023; Oh and Schuler, 2023b).

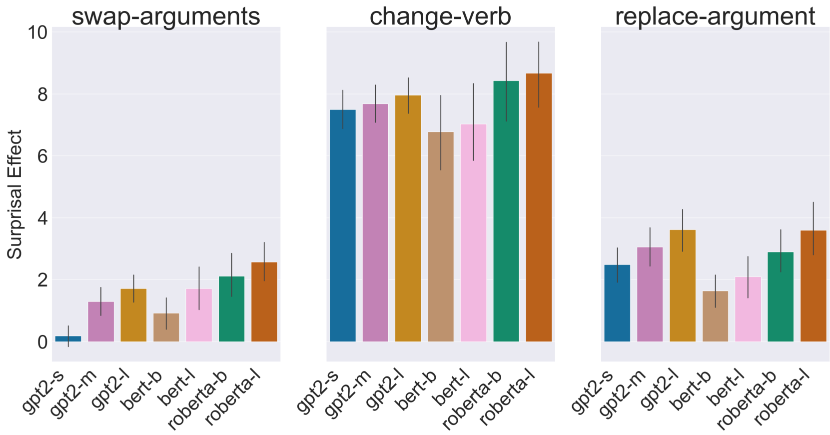

Figure 1: Surprisal effects plotted by condition and model.

rably better performance on the `change-verb` and `replace-argument` conditions relative to `swap-arguments`, it is likely that the models are making use of specific lexical cues to make their inferences rather than the structural relations humans are using. This is because the two conditions the model does better on introduce lexical variation in the stimuli, which is not the case for `swap-arguments`.

6 Experiment 2: Probing

6.1 Methods

While the surprisal estimates in Experiment 1 are computed based on the final layer of the models, in Experiment 2, we investigate which layers encode argument role information in verb representations by conducting a probing analysis. To show role-sensitivity at the verb, the model must correctly analyze the position of the arguments, represent the arguments with a role-specific meaning, and use that information to determine the plausibility of the verb that appears following the arguments. As these computations involve both syntactic and semantic processing, it is possible that such knowledge is encoded in earlier layers of the models which are not detectable in surprisal estimates based on final layer representations (Tenney et al., 2019; Jawahar et al., 2019). We investigate this by implementing layer-wise *probing classifiers* (Belinkov, 2022), on GPT2-small, which showed the most human-like pattern in the surprisal analysis, as well as GPT2-medium, BERT-large, and RoBERTa-large, which have the same number of layers and show better performance with the `swap-arguments` condition than GPT2-small.

For each condition, and for each layer, we train a logistic regression classifier on the models' representations of the target verbs, which predicts whether the verb is contextually appropriate or inappropriate. We choose to use a linear classifier because evidence points to conceptually relevant information being linearly separable in embedding space (Nanda et al., 2023). Target verbs in the plausible sentence were coded as 0 and the same target verbs in the implausible sentence were coded as 1.

Verb representations from each layer of each model were extracted using the `minicons` library. We report accuracies of each probe using 10-fold cross-validation with the `scikit-learn` implementation (Pedregosa et al., 2011). During training, we used a controlled method of splitting the train and test data sets, where the plausible and implausible verb pairs were always included in the same data set. This was to prevent the model from simply matching a verb in one context to the same verb in the counterpart context.

A high **classification accuracy** indicates that the verb representations extracted from the model contains information about the plausibility of the verb given the sentence it appears in - the model is able to distinguish contextually appropriate and inappropriate verbs.

6.2 Results

The probes trained on the verb representations in the `change-verb` condition performed at ceiling for all models (Figure 2). This suggests that in all models, the systematic change in verb form (*-ed* vs. *-ing*) is robustly encoded in verb representations. This pattern corroborates the surprisal results, where the `change-verb` condition showed significantly large surprisal effects in all models, suggesting that the models can effectively distin-

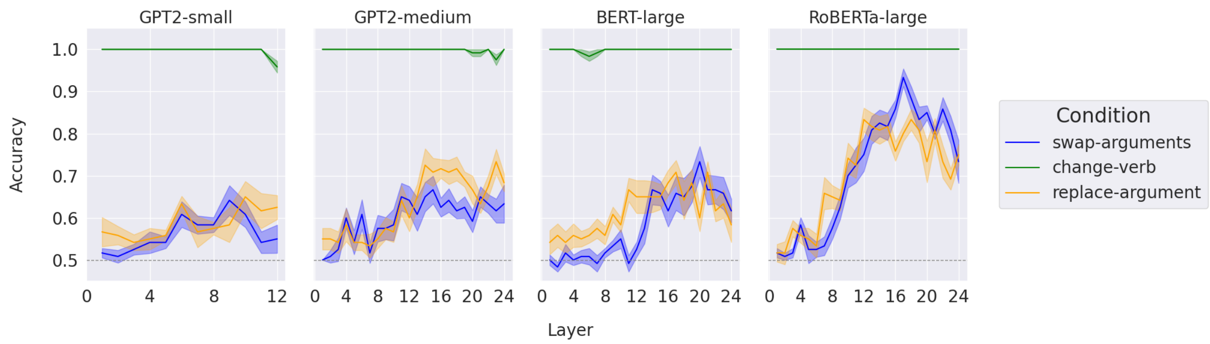

Figure 2: Classification accuracies for probes trained to distinguish plausible and implausible verbs under different conditions. Highlighted areas indicate standard errors of the mean across the 10 cross-validation folds. Dotted lines indicate at-chance accuracy.

guish verbs in the plausible and implausible contexts when the verb form differs between the two contexts.

Classification accuracy was generally lower for the conditions where the verb was kept the same and plausibility was determined by changing properties of the preceding context, i.e., `swap-arguments` & `replace-argument`, rather than verb form. GPT2-small did not improve greatly from chance-level performance. The larger models reached higher classification accuracy, with GPT2-medium and BERT-large reaching 70% accuracy, while RoBERTa showed the highest performance, reaching near 80-90% accuracy. For these larger models, decoding accuracy gradually increased throughout the layers and the particular increase in the middle layers suggests that verb plausibility information is more effectively represented from the middle layers.

While the accuracies between the `swap-arguments` and `replace-argument` conditions were overall comparable, the `replace-argument` condition showed slightly higher accuracy than the `swap-arguments` condition in earlier layers of BERT and RoBERTa, while the same contrast appeared in later layers of GPT2 (small and medium). This suggests that role-dependent verb plausibility information may be encoded at different stages of processing in uni- and bi-directional models. Finally, there was a tendency for the accuracies to fluctuate more and even decrease at the final layers, particularly for the `swap-arguments` condition in RoBERTa, which drops from 90% to 70% accuracy. This suggests that role-dependent plausibility information may become partially lost in models' representations.

7 Experiment 3: Attention

7.1 Methods

One question based on the previous experiment findings is what gives rise to models' relatively weak performance on determining verb plausibility based on argument role information, particularly when the argument role is manipulated by swapping the position of the arguments (`swap-arguments` condition). One possibility is that for these items, the models often incorrectly parse the argument roles indicated by the structure. It is possible that the models get confused about which noun is in which position and takes on which argument role. This could also offer a reason for why models perform better with the `change-verb` items, where argument position is fixed and held constant between the plausible and implausible conditions. In Experiment 3, we examine how models treat the preceding arguments by conducting an attention analysis that focuses on whether the models correctly allocate attention to the target subject at the verb position.

We adapt a similar method to that used in previous work. Ryu and Lewis (2021) inspected the attention patterns of GPT2 in order to probe whether the presence of a partially-matching distractor word interferes with the model's processing of a subject-verb dependency. The authors found an attention head that was specialized in finding the subject and examined whether the attention to the target subject differed between the intervening and non-intervening conditions.

We compare the attention profiles of GPT2-small and RoBERTa-large, the models that performed the worst and best, respectively, in the previous experiments. For each model, we first define an attention head that allocates the greatest attention weight from the verb to the subject in the sentence. For example, given the sentence, *The restaurant owner forgot which customer the waitress served during dinner yesterday*, we calculated the attention weight from the verb *served* to the subject *waitress* for each layer and head. We define the attention head that had the greatest attention weight to the subject as the subject attention head. The selected subject attention head was then used to calculate the attention from the verb to the subject and object, respectively. A high attention weight to the subject and a low attention weight to the object indicate that the model correctly distinguishes subjects from objects.

7.2 Results

For GPT2-small, we identified layer 3 head 10 (head indices: 2, 9) as the subject attention head, and for RoBERTa-large, we identified layer 13 head 16 (head indicies: 12, 15) as the subject attention head. The attention weight to the subject averaged across all items was .52 for GPT2-small and .68 for RoBERTa, indicating that these attention heads allocated most of the attention from the verb to the subject across the experiment items.

The results are shown in Table 2. We found similar attention patterns between the `swap-arguments` and `replace-argument` conditions. For both GPT2 and RoBERTa, the subject attention head correctly allocates most of its attention to the subject rather than the object. However, RoBERTa gives less attention overall to the object than GPT2 does, with the attention weight to the object remaining below 10%.

The results show that even GPT2-small, which did not show clear sensitivity to argument roles in the surprisal and probing analyses, correctly allocates attention to subjects with the subject head, though its attention is also distributed to the object more than the better performing RoBERTa-large. The attention analysis, therefore, suggests that it is unlikely that weak role-sensitivity at the verb arises from being confused about which argument is in which position or which argument is assigned which role. Rather, the weak performance could be due to how the models encode the preceding argument role information into the representations of the verb. Models may be able to correctly distinguish argument roles but less capable of using this information to represent role-compatible and role-inappropriate verbs in different ways.

8 Discussion

While previous studies have examined language models' knowledge of argument roles by testing their capacity to distinguish plausible and implausible sentences, we take a new approach by examining whether models' representations of verbs in sentences encode plausibility based on preceding argument role information. This method, in combination with the controlled sets of materials used in psycholinguistic studies that examine human comprehenders' role-sensitivity, offers a rigorous and systematic test of language models' sensitivity to argument roles and a way to directly compare human and model behavior. In the surprisal and probing analyses, we find that language models generally exhibit greater sensitivity to changes in argument meanings than to changes in argument roles, similar to humans' initial predictions. However, unlike humans, they fail to show the same pattern across different types of argument role manipulations. Whether the argument role and verb compatibility is manipulated by swapping the argument positions or by changing the verb form, humans show the same processing pattern, whereas language models treat the two cases differently.

The relatively weak sensitivity to verb plausibility when the preceding arguments are swapped, which we observed in Experiments 1 and 2, is unlikely due to a misrepresentation of the context, as the models' attention patterns in Experiment 3 suggest that roles are accurately represented. Rather, we suggest it arises from the difficulty in evaluating whether a verb is plausible given the particular argument-role bindings enforced by the preceding context. This involves a more complex analysis than simply computing context-independent argument and verb co-occurrences, which is potentially why humans' predictions fail to make use of such information rapidly during real-time prediction (Chow et al., 2016).

A key divergence between the model and human behaviors was with regard to which conditions caused more difficulty than others. Human comprehenders show the same pattern in the `swap-arguments` and `change-verb` conditions (i.e., no immediate N400 role-sensitivity), both of which involve the computation of argument-verb relations with respect to argument roles. In all the models we tested, we observed greater performance in the `change-verb` condition than the `swap-arguments` condition. This suggests that lan-

Model	Condition	Attention to Subject		Attention to Object	
		Plausible	Implausible	Plausible	Implausible
GPT2-small	`swap-arguments`	.53 (.15)	.53 (.17)	.18 (.10)	.19 (.06)
GPT2-small	`replace-argument`	.51 (.12)	.50 (.13)	.19 (.09)	.21 (.08)
RoBERTa-large	`swap-arguments`	.68 (.18)	.70 (.20)	.06 (.10)	.05 (.09)
RoBERTa-large	`replace-argument`	.65 (.16)	.68 (.16)	.06 (.08)	.04 (.02)

Table 2: Results of the attention analysis. The values represent the subject attention head's average attention from the verb to the subject and its attention from the verb to the object under each condition. Standard deviations are in parentheses.

guage models treat the two cases differently, indicating an absence of a shared underlying process of computing argument-verb relations. This kind of contrast between model and human behavior with respect to the variability across different constructions, has also been observed with human reading time data (Arehalli et al., 2022; Huang et al., 2024), and offers another way of evaluating models against human processing mechanisms.

One notable observation was that GPT2-small showed stronger correspondence with the human N400 data patterns, while larger models showed the higher performance in all experiments, which outperformed humans' initial predictive processing capacities. GPT2 and variants have shown to be more effective at predicting human behavior compared to larger autoregressive models (Oh and Schuler, 2023c; Kuribayashi et al., 2023). Steuer et al. (2023) find a similar pattern, where smaller models predict human reading times better than larger ones that do better on syntactic and semantic judgments. Our results suggest that smaller models capture more immediate, online processing profiles of humans, and resemble human N400 patterns which reflect initial stages of predictive processing. Conversely, the measures derived from larger models more closely pattern with offline, final interpretations of humans. Nevertheless, no models capture the consistency between the two argument role manipulations which has been found with humans. These results offer insights into drawing connections with human empirical findings, especially for psycholinguists aiming to use language models, with regard to determining which models to use when simulating experiments. Additionally, the improved performance of larger models raises the question of whether scale is sufficient to learn these complex role-specific relationships; evaluating the argument role-reversal and `replace-argument` contrast for larger models like LLaMa (Touvron et al., 2023), as well as tracing the ability based on the number of parameters of a language model, e.g., the Pythia family of models (Biderman et al., 2023), can facilitate these types of investigations.

Our work provides a critical perspective to language models' representations of argument roles from a psycholinguistic perspective. Future directions could involve applying causal interpretability methods (Meng et al., 2022; Arora et al., 2024) to these sets of sentences. It may be the case that larger-scale models that assign correct plausibility ratings are implementing the similar computations for `replace-argument` and reversal items, which will take us further towards determining whether linguistic knowledge in language models is as robust as it seems.

Limitations

Cross-Linguistic Coverage

Our investigation was focused on English, but the role reversal effect has also been shown in languages like Mandarin (Chow et al., 2018) and German (Stone and Rabovsky, 2024). Although it is linguistically robust across humans, Xu et al. (2023) found that language model surprisal exhibits different trends in each of these three languages. Testing whether similar effects appear in other language models as well as monolingual or multilingual language models could be a way to establish whether the models' inferences are are based on language-specific factors or whether generalized representation of argument roles is an emergent phenomenon.

Interpretability

We identified attention heads that tracked dependencies using a purely correlational mechanism, based on the weights between the verb and its arguments. Although this measure is easily understandable, it is likely that the attention heads do not just track subject-verb and object-verb dependencies. A key future direction is to build on work in interpretability (Lakretz et al., 2021; Meng et al., 2022) which identifies causal mechanisms in language models responsible for specific computations. Arora et al. (2024) apply some of these measures to pairs of grammatical and ungrammatical sentences handling various syntactic phenomena. In future work, we hope to not just extend their methods, but derive measures of cognitive effort based on how the language models causally compute argument roles.

Ethical Considerations

All data and language models we used were publicly available, and our experiments do not rely on any specialized computing hardware.

References

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 301–313, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. 2024. Causalgym: Benchmarking causal interpretability methods on linguistic tasks. *arXiv preprint arXiv:2402.12560*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Wing-Yee Chow, Ellen Lau, Suiping Wang, and Colin Phillips. 2018. Wait a second! delayed impact of argument roles on on-line verb prediction. *Language, Cognition and Neuroscience*, 33(7):803–828.

Wing-Yee Chow, Cybelle Smith, Ellen Lau, and Colin Phillips. 2016. A "bag-of-arguments" mechanism for initial verb predictions. *Language, Cognition and Neuroscience*, 31(5):577–596.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

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.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2013. Word surprisal predicts n400 amplitude during reading.

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.

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*.

Michael Hanna, Yonatan Belinkov, and Sandro Pezzelle. 2023. When language models fall in love: Animacy processing in transformer language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12120–12135, Singapore. Association for Computational Linguistics.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Jasper Jian and Siva Reddy. 2023. Syntactic substitutability as unsupervised dependency syntax. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2341–

2360, Singapore. Association for Computational Linguistics.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event knowledge in large language models: the gap between the impossible and the unlikely. *Cognitive Science*, 47(11):e13386.

Albert Kim and Lee Osterhout. 2005. The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of memory and language*, 52(2):205–225.

Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. 2023. Psychometric predictive power of large language models. *arXiv preprint arXiv:2311.07484*.

Marta Kutas and Steven A Hillyard. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205.

Yair Lakretz, Dieuwke Hupkes, Alessandra Vergallito, Marco Marelli, Marco Baroni, and Stanislas Dehaene. 2021. Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, 213:104699.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Bai Li, Zining Zhu, Guillaume Thomas, Yang Xu, and Frank Rudzicz. 2021. How is BERT surprised? layerwise detection of linguistic anomalies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4215–4228, Online. Association for Computational Linguistics.

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

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Danny Merkx and Stefan L. Frank. 2021. Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22, Online. Association for Computational Linguistics.

James A Michaelov, Megan D Bardolph, Cyma K Van Petten, Benjamin K Bergen, and Seana Coulson. 2024. Strong prediction: Language model surprisal explains multiple n400 effects. *Neurobiology of language*, pages 1–29.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30.

Byung-Doh Oh and William Schuler. 2023a. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921, Singapore. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2023b. 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 and William Schuler. 2023c. 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.

Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. 2022. When classifying grammatical role, BERT doesn't care about word order... except when it matters. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 636–643, Dublin, Ireland. Association for Computational Linguistics.

Ellie Pavlick. 2022. Semantic structure in deep learning. *Annual Review of Linguistics*, 8:447–471.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Soo Hyun Ryu and Richard Lewis. 2021. Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–71, Online. Association for Computational Linguistics.

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.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Julius Steuer, Marius Mosbach, and Dietrich Klakow. 2023. Large GPT-like models are bad babies: A closer look at the relationship between linguistic competence and psycholinguistic measures. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 142–157, Singapore. Association for Computational Linguistics.

Kate Stone and Milena Rabovsky. 2024. The role of syntactic and semantic cues in preventing illusions of plausibility.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

William Timkey and Tal Linzen. 2023. A language model with limited memory capacity captures interference in human sentence processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8705–8720, Singapore. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

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 Wilcox, Clara Meister, Ryan Cotterell, and Tiago Pimentel. 2023a. Language model quality correlates with psychometric predictive power in multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7503–7511, Singapore. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2023b. Using computational models to test syntactic learnability. *Linguistic Inquiry*, pages 1–44.

Michael Wilson, Jackson Petty, and Robert Frank. 2023a. How abstract is linguistic generalization in large language models? experiments with argument structure. *Transactions of the Association for Computational Linguistics*, 11:1377–1395.

Michael Wilson, Jackson Petty, and Robert Frank. 2023b. How abstract is linguistic generalization in large language models? experiments with argument structure. *Transactions of the Association for Computational Linguistics*, 11:1377–1395.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Weijie Xu, Jason Chon, Tianran Liu, and Richard Futrell. 2023. The linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15711–15721.

A Computational Resources

All experiments were run on a single CPU and took no more than two hours to run. We report metrics from a single run.

B Control Items

As a control, we examined a set of items included in each study (Chow et al., 2016; Kim and Osterhout, 2005), where the plausibility of the verb was manipulated by simply replacing the target verb with another verb or associating the target verb with another argument. These materials have shown to

Model	Parameters	#L	#U	#H
GPT2 S	124M	12	768	12
GPT2 M	355M	24	1024	16
GPT2 L	774M	36	1280	20
BERT B	110M	12	768	12
BERT L	340M	24	1024	16
RoBERTa B	125M	12	768	12
RoBERTa L	355M	24	1024	16

Table 3: Summary of Model Architectures. L, U, H refers to number of layers, hidden units, and attention heads.

elicit immediate neural responses in human comprehenders, indicating sensitivity to the likelihood of a target word appearing in a plausible context.

Experiment	High Cloze	Low Cloze
Chow et al. (2016)	Abby brushed her teeth after every **meal** and every snack.	Abby brushed her teeth after every **game** and every snack.
Kim and Osterhout (2005)	The hungry boys were **devouring** the plate of cookies when Jack arrived.	The dusty tabletops were **devouring** with gusto.

Table 4: Examples of control items.

We computed the surprisal effect for plausible and implausible variants of the same item for both studies, finding a much higher surprisal effect for both sets of control items relative to the experimental conditions.

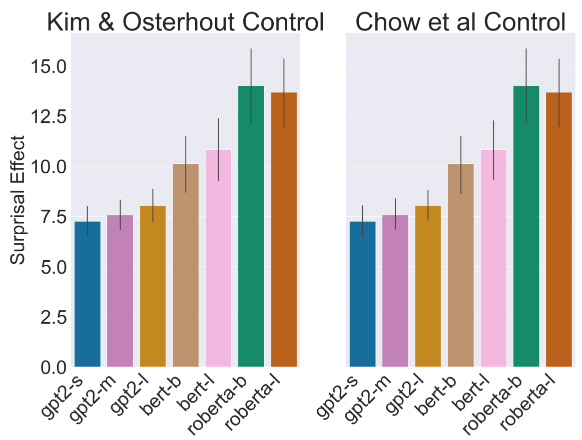

Figure 3: Compare `swap-arguments` and `replace-argument` to Chow et al, `change-verb` to Kim and Osterhout