When *Men Bite Dogs*: Testing Good-Enough Parsing in Turkish with Humans and Large Language Models

Onur Keleş
Department of Linguistics
Boğaziçi University
onur.keles1@bogazici.edu.tr

Nazik Dinçtopal Deniz
Department of Foreign Language Education
Boğaziçi University
nazik.dinctopal@bogazici.edu.tr

Abstract

This paper investigates good-enough parsing in Turkish by comparing human self-paced reading performance to the surprisal and attention patterns of three Turkish Large Language Models (LLMs), GPT-2-Base, GPT-2-Large, and LLaMA-3. The results show that Turkish speakers rely on good-enough parsing for implausible but grammatically permissible sentences (e.g., interpreting sentences such as 'the man bit the dog' as 'the dog bit the man'). Although the smaller LLMs (e.g., GPT-2) were better predictors of human RTs, they seem to have relied more heavily on semantic plausibility than humans. Comparably, larger LLMs (e.g., LLaMA-3) tended to make more probabilistic parsing based on word order, exhibiting less *good-enough* parsing behavior. Therefore, we conclude that LLMs take syntactic and semantic constraints into account when processing thematic roles, but not to the same extent as human parsers.

Introduction

Human language comprehension is rapid and incremental, in which listeners or readers build up provisional interpretations of sentences *on the fly*. However, a growing body of work suggests that these interpretations can occasionally be shallow or incomplete, leading to syntactic misanalyses. *Good-enough parsing* (Ferreira et al., 2002; Ferreira and Patson, 2007; Christianson et al., 2001), argues that language comprehension sometimes relies on heuristics guided by real world knowledge and may not engage in detailed syntactic analyses. That is, a sentence, such as 'the dog was bitten by the man' can be interpreted as 'the dog bit the man.' Such errors are more prevalent when the event described is highly plausible in one particular direction (e.g., 'dogs biting people'), but is reversed in a sentence, such as 'the dog was bitten by the man' especially in noncanonical syntactic structure like passive constructions due to the conflict between real world knowledge and syntactic structure (Ferreira, 2000).

To our knowledge, *good-enough parsing* has not been examined in Turkish, a language that encodes key thematic roles through overt morphosyntactic marking that can potentially influence such misinterpretations. A passive construction, for instance, is marked not only by a syntactic operation (e.g., *Move-Merge* in generative approaches) but also through morphosyntactic operations on the verb (e.g., the passivizing suffix *-Il-*).

This paper investigates if Turkish speakers are prone to *good-enough parsing* effects and if Turkish LLMs, which are hitherto unexplored in the context of psycholinguistics, also exhibit these effects. Specifically, we investigate the extent to which Turkish speakers and Turkish LLMs use syntactic detail and real world plausibility in their processing of active and passive constructions.

To address these questions, we employed (i) a Self-Paced Reading (SPR) experiment with native Turkish speakers, using sentences adapted from Ferreira (2000) and (ii) computational simulations with three Turkish LLMs (GPT-2-Base, GPT-2-Large, and LLaMA-3). By comparing human reading times and model-based surprisal measures, along with each model's relative attention to morphological cues, we aim to show how factors like animacy, semantic plausibility, morphosyntax, and model capacity jointly shape the likelihood of *good-enough* interpretations in Turkish.

Psycholinguistics and LLMs

Recent studies have leveraged LLMs to explore how well they capture processing patterns in humans (Salicchi et al., 2023; Marvin and Linzen, 2018; Goldberg, 2019; Wilcox et al., 2023a; Wang et al., 2024). Surprisal is often used as a predictor of human reading times, showing strong correla-

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 219–231
May 3, 2025 ©2025 Association for Computational Linguistics

tions across languages and sentence types (Wilcox et al., 2023a; Oh and Schuler, 2022; Nair and Resnik, 2023). In particular, it has been claimed that processing difficulty at a word reflects how unexpected that word is, given prior context. LLMs have also been tested with respect to their ability to detect event plausibility in English (Pedinotti et al., 2021; Kauf et al., 2024). For example, Kauf et al. (2023) highlight that current LLMs can infer thematic relations and can recognize events more consistently that are truly impossible (e.g., 'the laptop bought the teacher') than those that are merely unlikely (e.g., 'the boy tutored the nanny'). Similarly, Amouyal et al. (2024) presented evidence that log probabilities from base and instruction-tuned models can be used as a proxy for human plausibility judgments, while Kauf et al. (2024) evaluated how these probabilities map onto graded semantic acceptability. Finally, Ettinger (2020) reported that although BERT can pick up on role reversal differences or identify thematic relations, its sensitivity is lower than that of human judgments. Together, these studies suggest that LLMs take into consideration the syntactic regularities and world knowledge, albeit not always matching the precise patterns found in human data and there seems to be variation among different models.

While most of the work on LLMs have centered on English, there is growing interest in multilingual settings and underrepresented languages, including Turkish. Recent work has tested Turkish LLMs for different linguistic tasks like indexical shift (Oğuz et al., 2024) and universal dependency annotation evaluation (Akkurt et al., 2024). However, as far as we are concerned, psycholinguistic evaluation in Turkish LLMs remains sparse. Our investigation therefore addresses a key gap by providing a direct comparison between human data and the predictions of autoregressive Turkish LLMs with differing sizes in an experiment designed to test *good-enough parsing* effects.

Methodology

Self-Paced Reading (SPR) Experiment

This experiment examined the predictions of *good-enough parsing* model (Ferreira and Patson, 2007) with native Turkish speakers. 26 native Turkish speakers ($M_{\text{age}} = 19$, all college students) participated in a word-by-word self-paced reading (SPR) experiment. The experimental sentences, exemplified in (1), were translated from the materials in Ferreira (2000) into Turkish and were slightly adapted to prevent ambiguity and ensure naturality (in (1) and elsewhere, PL: plural, ACC: accusative case, PST: past tense, PASS: passive voice, 1SG: first person singular marker).

(1) a. Köpek-ler adam-ı **ısır-dı** sanır-ım
dog-PL man-ACC bite-PST think-1SG
'I think the dogs bit the man.'

b. Adam-lar köpek tarafından **ısır-ıl-dı**
man-PL dog by bite-PASS-PST
sanır-ım
think-1SG
'I think the men were bitten by the dog.'

c. Adam-lar köpeğ-i **ısır-dı** sanır-ım
man-PL dog-ACC bite-PST think-1SG
'I think the men bit the dog.'

d. Köpek-ler adam tarafından **ısır-ıl-dı**
dog-PL man by bite-PASS-PST
sanır-ım
think-1SG
'I think the dogs were bitten by the man.'

e. **Question:** Is the event in the sentence plausible?

The experimental sentences manipulated syntactic structure as active voice as in (1a, c) or passive voice as in (1b, d), word order as non-reversed as in (1a, b) or reversed as in (1c, d). Half of the sentences had animate arguments which were reversible but *biased* as in 'the man bit the dog' and the other half had one animate one inanimate argument which were *irreversible* as in 'the chef wore the apron.' Reversing the arguments in the reversible set would result in permissible but unlikely events and the reversal of arguments in the irreversible set would cause semantic anomaly. A *symmetrical* set (e.g., 'the boy kissed the girl') was used as control in which the two arguments were equally likely to be agents. All sets had 21 experimental sentences. Each sentence ended with a content-neutral word[1] to prevent *wrap-up* effects. There were 21 sentences each in reversible and irreversible sets, with the four conditions manipulating syntactic structure (active, passive) and word order (reversed, non-reversed), totaling up to 42 experimental items. The experimental sentences were distributed across four reading lists counterbalancing for syntactic structure (active, passive) and word order (reversed, non-reversed). In each list, the experimental sentences were intermingled with 21 additional controls and six practice items.

The experiment was prepared on the PCIbex

[1] Words expressing epistemic modality like *perhaps, maybe, probably*.

experiment building software (Zehr and Schwarz, 2018)[2] and an online link to it was shared with the participants, who read, on their own computer, the sentences word-by-word moving from one word to the next with a key-press. Their task was to indicate, by clicking on two possible options presented under the sentence, if the sentence described a plausible event (see 1e). Accuracy of the response to the plausibility question, word reading time, and end-of-sentence plausibility decision time were measured.

LLM Experiments

Models In addition to Turkish speakers, we tested the predictions of the *good-enough parsing* model on three decoder-only Turkish LLMs using the same experimental item set. We used the base and large variant of the GPT-2 (Radford et al., 2019) trained on Turkish (Kesgin et al., 2024b) and a LLaMA-3 Turkish (Kesgin et al., 2024a), which is an adapted version of the LLaMA-3 model (Dubey et al., 2024) fine-tuned using a 30GB dataset of Turkish. The models used in this study shared the same architecture as autoregressive models trained for next token prediction to align well with the SPR task reported earlier, which tests incremental processing. However, these models differ in size and performance: GPT-2-Base Turkish has 124 million parameters, 12 layers, and 12 attention heads; GPT-2-Large Turkish, 774 million parameters, 36 layers, and 20 attention heads; LLaMA-3-8B Turkish, 8 billion parameters, 32 layers, and 32 attention heads. To compare with human word reading time, we calculated model surprisal for each word in each sentence. We also report a heatmap visualization of the model's attention to examine if it shifts based on the likelihood of the event.

Surprisal We simulated the incremental processing behavior of Turkish speakers as in the SPR task and estimated surprisal values for each word in the sentence to examine if model surprisal in Turkish could predict reading times. To do so, the experimental items were first tokenized using the byte-pair encoding (BPE) tokenizer (Sennrich, 2015), resulting in sub-word sequences. Each word w_i in the sentence was then incrementally presented to the model, conditioned on its preceding context $w_{1:i-1}$. Formally, surprisal $S(w_i)$ is defined as the negative log-probability of w_i given $w_{1:i-1}$:

$$S(w_i) = -\log P(w_i \mid w_{1:i-1}) \qquad (1)$$

where $w_{1:i-1}$ represents the words preceding w_i. In practice, the first word w_1 has no context ($w_{1:0} = \emptyset$), the second word w_2 depends on w_1, the third word w_3 depends on w_1, w_2, and so on.

To account for sub-word segmentation, we aggregated the surprisal estimates of all the sub-words belonging to a single word (following Wilcox, 2020; Oh and Schuler, 2023, and others). If a word w is decomposed into sub-words $(s_1, s_2, \ldots, s_k)$, its word-level surprisal $S(w)$ is computed as the following:

$$S(w) = \sum_{i=1}^{k} S(s_i) \qquad (2)$$

Attention weights Following Li et al. (2024)'s approach (also see Clark et al., 2019; Voita et al., 2019, for similar uses), we created a heatmap visualization of attention, which is a common strategy for probing model interpretability. In the transformer architecture (Vaswani et al., 2017), each layer contains multiple self-attention heads. These heads compute weighted dot products among token representations (query, key, value), allowing the model to capture a wide range of linguistic relationships. This way, we can try to understand what the model is attending to or *looking at* when processing a word w_i.

Our focus was the degree of attention to the post-position *tarafindan* 'by', which introduces the agent in passive constructions in Turkish. For each condition, we computed (for all attention heads and layers) and subtracted the attention weights of the NP local to the post-position from the NP that is distant, which gives us how much more/less attention the two NPs received compared to each other. We then computed the difference between these relative attention patterns when word order was reversed and evaluated how reversing word order (i.e., the condition when event becomes less likely or impossible) affected these relative attention patterns. Increased attention toward the distant NP in the reversed condition may point to a less accurate interpretation (indicating that the sentence was processed in a *good-enough* manner), where the model relies more on semantic/real-world plausibility cues than on strict syntactic structure. Conversely, preference for the local NP would indicate that *tarafindan* was successfully mapped to the correct agent phrase that it introduces (as in Figure 1).

[2] http://farm.pcibex.net

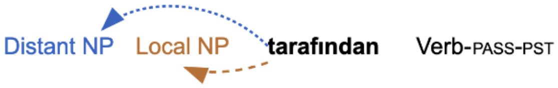


Figure 1: Possible model attention routes from the post-position *tarafindan* 'by' and two noun phrases. A structural dependency between *tarafindan* and the non-agentive distant NP is not possible in Turkish. The brown path indicates the correct dependency (resolved at the agent), whereas the blue path points to an incorrect dependency (resolved at the non-agent subject).

Results and Discussion

SPR Results

Accuracy Table 1 shows Turkish speakers' accuracy in their decisions. Following a strategy similar to Kauf et al. (2023), if speakers answered 'plausible' to reversed orders in biased and irreversible sets, we considered that response to be erroneous. For inference, we fit a mixed effect binomial model to Accuracy with Word Order (Reversed, Nonreversed), Structure (Active, Passive) as fixed effects using `lmer` (Bates et al., 2005) in R for biased and irreversible sets separately. Participants and Items were entered as random effects.

Overall, Turkish speakers were successful (with a mean accuracy of 90% or higher) in all conditions except for reversed sentences in the reversible (i.e., biased) condition. For these sentences, we observed an error rate of 25% for reversed constructions, where the reversed order had a significant negative effect on accuracy ($\textit{Odd Ratio} = 0.11, p < .001$). There was no reliable difference in accuracy for active (30% error) and passive constructions (20% error) in the reversed condition ($\textit{Odd Ratio} = 3.48, p = .13$). Reversing the word order did not result in a decrease in accuracy for irreversible sentences ($\textit{Odd Ratio} = 5.39, p = 0.13$). We attribute this to a *good-enough parsing* effect present in the biased set, whereby participants mistakenly preferred the interpretation that was more in line with their real world knowledge, but not with the syntactic structure of the sentence.

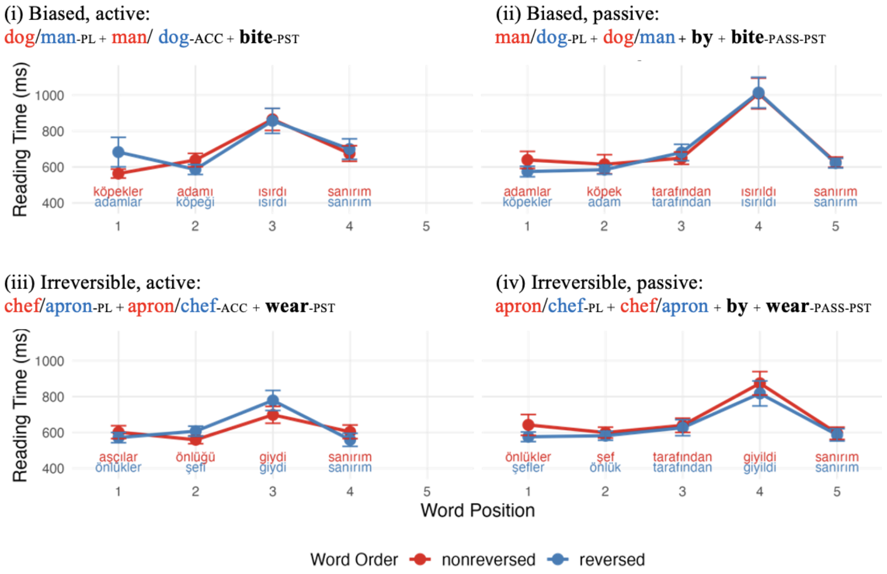
This did not happen for the irreversible sentences, though, possibly because the presence of an inanimate entity was a strong cue for the correct structure. Note that *good-enough parsing* effects were observed in passive constructions in English (Ferreira, 2000). We attribute the comparable decrease in accuracy in passive constructions (compared to that in active counterparts) in Turkish to the semantic content that the 'by'-phrase has in Turkish. Unlike its English counterpart 'by', *tarafindan* is a semantically transparent word carrying lexical content that could have provided additional cue to the correct parse. The observation of the decrease in accuracy in active sentences in Turkish (compared to the lack thereof in English) can be attributed to the relatively flexible word order in Turkish in which the order of agents and patients can change depending on the information structure of the sentence (İşsever, 2003).

Set	Word Order	Structure	Accuracy
Biased	Nonreversed	Active	96%
Biased	Nonreversed	Passive	99%
Biased	Reversed	Active	71%
Biased	Reversed	Passive	80%
Irreversible	Nonreversed	Active	97%
Irreversible	Nonreversed	Passive	92%
Irreversible	Reversed	Active	99%
Irreversible	Reversed	Passive	99%

Table 1: Turkish Speakers' Mean Accuracies on the Plausibility Task

Word reading time The RTs for each word can be examined in Figure 2. We fit a mixed effects regression model on the log-transformed RTs for the reversible and irreversible sets separately. Word Order (Reversed, Nonreversed), and Structure (Active, Passive) were fit as fixed effects and Participant and Item were random effects. In addition, all models also included Word Length, Previous Word Length, Region (Verbal, immediately Preverbal, Other) as additional predictors. All numeric factors were centered to prevent collinearity. The verb region had the highest RT in all conditions ($\beta = .23, p < .001$ for both sets). There was no significant interaction between Word Order and Region for the biased set ($\beta = -.05, p = .218$ for preverbal; $\beta = -.07, p = .07$ for verbal), meaning that reversal of the arguments did not yield an online surprisal effect for Turkish speakers.

However, both the verbal ($\beta = .12, p = .001$) and preverbal ($\beta = .04, p = .029$) regions had significantly increased RTs in the reversed condition when the events were irreversible (as in 'aprons wearing chefs'). This suggests that semantic anomaly was detected at the verbal and preverbal region. In the preverbal region, which corresponds to the direct object with accusative marking (e.g., *önlüğ-ü* 'apron-ACC') in active sentences and the post-position *tarafindan* in passive



(i) Biased, active:
dog/man-PL + man/ dog-ACC + **bite**-PST

(ii) Biased, passive:
man/dog-PL + dog/man + **by** + **bite**-PASS-PST

(iii) Irreversible, active:
chef/apron-PL + apron/chef-ACC + **wear**-PST

(iv) Irreversible, passive:
apron/chef-PL + chef/apron + **by** + **wear**-PASS-PST

Figure 2: Word-by-word Reading Time (RT) in the Self-Paced Reading (SPR) Task by Set (Biased, Irreversible), Structure (Active, Passive), and Word Order (Reversed, Nonreversed).

sentences, participants encountered critical morphosyntactic/semantic cues that signaled the upcoming structure. In active sentences, the presence of accusative case-marking on an inanimate argument probably served as a strong predictive cue as to the theme/patient of the upcoming verb. Similarly, in passive sentences, encountering the agentive post-position *tarafindan* could have provided information about the upcoming passive verb. These morphosyntactic and semantic cues might have allowed the readers to construct syntactic and thematic dependencies before reaching the verb, where these predictions were ultimately resolved. This was not the case for biased sentences, in which both arguments were animate. This animacy information may have let readers to entertain both arguments as potential agents and patients.

Decision time Figure 3 shows the sentence-final decision times for each condition. We fit a mixed effects regression model to log-transformed decision times with Word Order (Reversed, Nonreversed) and Structure (Active, Passive) as fixed effects for each set. Participants and Items were random effects. Overall, in biased conditions, both in active and passive constructions, the participants took longer to decide in reversed conditions than in nonreversed conditions ($\beta = .18, p = .007$). In irreversible conditions, the opposite pattern was observed and the participants took less time to decide for reversed sentences ($\beta = -.014, p = .014$), and spent more time on sentences with a canonical word order.

Let us first consider the reversible condition. When one argument (e.g., ‘the dog’) is more likely to do an action (e.g., ’biting’) than the other (e.g., ‘the man’), reversing their order resulted in delay in decision times. Together with accuracy data (reduced accuracy in reversed than non-reversed conditions), we interpret this delay to *good-enough parsing* effects. Although the participants faced some processing difficulty due to the implausibility of the event (e.g., ‘the man biting the dog’), some participants, to some extent, appear to have interpreted such sentences as their plausible counterparts (e.g., ‘the dog biting the man’). The pattern in the irreversible conditions was not predicted but is explicable. In irreversible conditions, the agent (e.g., ‘the chef’) referred to an animate entity and the patient (e.g., ‘the apron’) was inanimate. Reversing their order was predicted to cause processing difficulty but it appears that the participants were quick to integrate the animacy information in their decisions and to detect the implausibility when the order of the arguments was reversed.

LLM Results

Surprisal The estimated surprisals from each of the three Turkish models are given in Figure 4. To investigate if the same critical regions resulted in difference in model surprisal, mixed effects linear models were fit for each model and for each set, resulting in 6 models. Word Length, Preceding

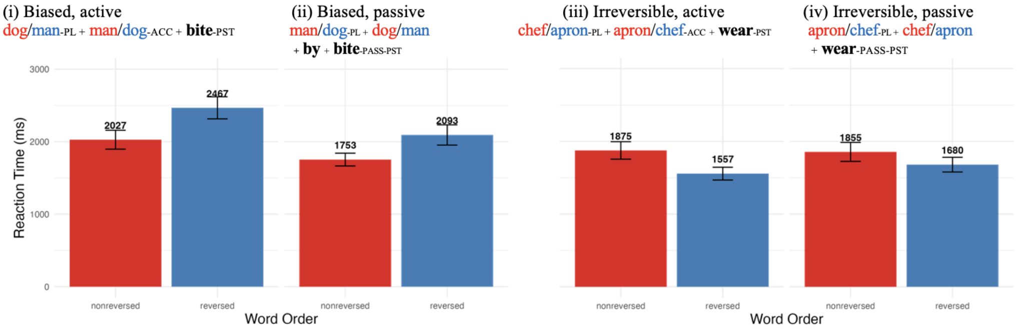


Figure 3: Semantic Plausibility Decision Time by Set (Biased, Irreversible), Structure (Active, Passive), and Word Order (Reversed, Nonreversed)

Word Length, Region (Verbal, Preverbal, Other), Word Order (Reversed, Nonreversed), and Structure (Active, Passive) were entered as fixed effects. Item was entered as a random effect. For the biased set, both variants of the GPT-2 (base and large) had only preverbal region significant in the passive voice when the arguments were reversed ($\beta = .37$, $p = .04$ for GPT-2-Base; $\beta = .28$, $p = .04$ for GPT-2-Large). In contrast, LLaMA-3 had significantly larger surprisal estimates only for the verb region in the reversed condition ($\beta = .42, p < .001$).

As to the irreversibles, the large and base variants of GPT-2 behaved almost identically except for the surprisal for the verbal region in the irreversible passive condition. Verbs that had reversed arguments in the passive voice had a larger surprisal estimate (p's $< .001$) than the non-reversed argument in GPT-2-Large. This means that the larger model was able to detect the semantic bias on both critical regions whereas GPT-2-Base did so only for the preverbal region with no significant effect of the verbal region ($\beta = .31, p = .07$). All critical regions significantly increased surprisal estimates in other conditions (p's $< .001$). In contrast, LLaMA-3 also yielded significantly larger surprisals for all conditions in the irreversible set (p's $< .005$).

Overall, LLaMA-3 demonstrated a broader sensitivity to structural and semantic cues compared to the GPT-2 family, particularly in irreversible sentences. In contrast, GPT-2 models showed sensitivity in specific regions (preverbal for the base model, and both preverbal and verbal for the larger variant). This suggests that LLaMA-3 might have better captured structural and semantic dependencies than the smaller models. Additionally, the differences in how GPT-2 variants process passive constructions with reversed arguments suggest that model size might be influencing the ability to integrate multiple linguistic cues in Turkish,[3] with the larger GPT-2 model estimating higher surprisal for both critical regions.

Predicting RTs We also tested if model surprisal predicted Turkish speakers' word RTs. For this, we added two additional predictors to the word reading time model described earlier. We added current word's w_i Surprisal and PrevSurprisal (which corresponds to the surprisal from the previous word w_{i-1} for potential spillover effects) following Wilcox et al. (2023b). Then, the baseline model and surprisal models were compared with a likelihood ratio test for each set. We found that the models that included surprisal as an additional predictor had significantly greater likelihood for both the biased ($\chi^2(3) = 28.94$, $p < .001$) and irreversible sets ($\chi^2(1) = 8.63$, $p < .01$). The general finding that surprisal predicts reading times is in line with observations from prior studies (Demberg and Keller, 2009; Shain et al., 2022; Wilcox et al., 2023b).

To investigate which of these language models best predict human RTs, we split the data by LLM and fit linear mixed-effects models, one for each LLM across the two sets. The surprisal derived from GPT-2-Large predicted the RTs at the critical region (the verb) ($\beta = .04, p < .001$). The surprisal of GPT-2-Base had some predictive power but it did not reach statistical significance ($\beta = .04, p = .09$). LLaMA-3 surprisal did not predict the RTs at the critical region ($\beta = .00, p = .759$) but it did predict the RTs at the region preceding the verb (at the accusative-marked NPs in the active condition and at the postposition 'by' in the passive condition) ($\beta = .03$, p < .002). Overall, these find-

[3] For comparison, see Appendix A for the surprisal estimates by the same models in English.

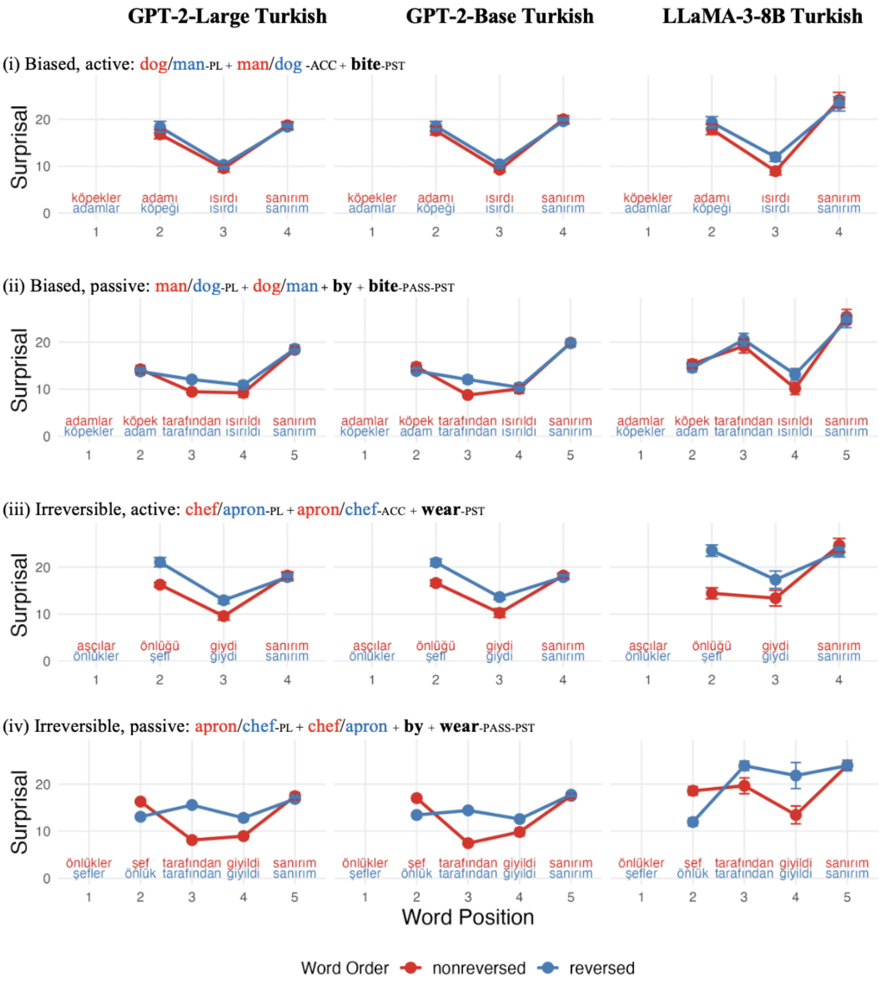


Figure 4: Replication of the SPR Task: Left-context Only Word-by-word Model Surprisal for GPT-2-Large Turkish, GPT-2-Base Turkish, LLaMA-3-8B Turkish by Set (Biased, Irreversible), Structure (Active, Passive), and Word Order (Reversed, Nonreversed).

ings indicate that the surprisal values from GPT-2 models (particularly the large variant) accounted for human RTs at the critical region better than those from LLaMA-3.

Attention weights Finally, we tested if models' attention patterns reflected sensitivity to argument structure, particularly in passive constructions with *tarafindan* ('by') in Turkish. We focused on the attention weights from *tarafindan* to its local, agent, NP and the more distant, patient, NP, comparing these patterns across canonical and reversed word orders. This analysis allowed us to examine whether the models correctly associated the agent marker with the more local and structurally correct NP, or if they showed increased attention to the distant NP in non-canonical orders, which might indicate shallower syntactic processing and more semantic or heuristic processing.

Figure 5 demonstrates the relative change in attention toward distant or local NP when the condition changed from non-reversed to reversed across all layers and attention heads. In other words, it shows us if the attention heads kept looking at the same NP or changed their attention weights in favor of either one of the two candidate NPs when the arguments were reversed. The red colors indicate change in attention in favor of the local NP, and the blue colors refer to a change in favor of the distant NP. A distant NP preference would imply that the model is associating *tarafindan* more with the non-agent possibly because of a semantic intrusion.

For biased passive sentences (top row), GPT-2-Base showed mixed patterns with scattered attention shifts across layers and heads, while GPT-2-Large demonstrated more consistent but subtle changes (given the more faded colors) in attention distribution. LLaMA-3 exhibited the most uniform pattern, with minimal attention shifts between conditions. In irreversible passive constructions (bottom row), the patterns were more pronounced. In the GPT-2 models, there were increasing looks or sensitivity to the distant NP. While for the base model distant NP preference took place across different layers, for the large model it mainly occurred in the deeper layers, albeit less strongly.

LLaMA-3 maintained relatively stable attention patterns with only a few heads showing strong preferences for either NP. This suggests that LLaMA-3's processing of argument structure may be more immune to word order variations, particularly in syntactically constrained contexts. Deeper layers did not display significant attention shifts in either set. Furthermore, there was increasing attention weight at the local NP in the irreversible set. We infer that LLaMA-3 mostly did probabilistic and syntactically constrained processing, and presumably understood semantically implausible events with correct mapping of thematic relations. In contrast, both GPTs seem to have relied more on semantic cues and mapped *tarafindan* with the more plausible and animate NP in Turkish.[4]

General Discussion and Conclusion

Our findings offer converging evidence that Turkish speakers exhibit *good-enough parsing* effects, particularly for *biased* sentences in which real-world plausibility (e.g., 'dogs biting men' versus 'men biting dogs') competes with syntactic structure. The SPR experiment revealed that reversed sentences in the biased set generated a *good-enough parsing* effect, leading to increased error rates and slower plausibility decisions. However, this effect was greatly reduced in the irreversible events containing semantic anomaly. We conclude that the animacy cues help participants form accurate interpretations. Furthermore, all three autoregressive Turkish models (GPT-2-Base, GPT-2-Large, and LLaMA-3 Turkish) showed sensitivity to structural and semantic anomalies in their surprisal estimates. However, the larger GPT-2 model captured more linguistic cues than its smaller variant, and LLaMA-3 model appeared to be the most robust in assigning correct syntactic dependencies even under conditions of unlikely events. In addition, human reading times were significantly predicted by LLM-based surprisal, supporting earlier findings in the literature (Li et al., 2024; Wilcox, 2020). The surprisal analysis also revealed that GPT-2 better predicted human RTs than LLaMA-3. This finding aligns well with earlier observations made for English suggesting that smaller or mid-sized models (e.g., GPT-2 variants) can mirror human reading patterns (Oh and Schuler, 2023; Kuribayashi et al., 2023) more closely, in line with large-scale evidence for surprisal-based predictability effects (Shain et al., 2024). Moreover, attention weight analyses showed that GPT-2 models often shifted attention toward the more semantically plausible (but syntactically incorrect) noun phrase in reversed sentences, presumably reflecting good-enough heuristics across both biased and irreversible stimuli. Meanwhile, LLaMA-3 appeared more robust in capturing correct agent–patient mappings, which was more unlike the human data.

Crucially, Turkish speakers also relied on *good-enough parsing* strategies, but only for biased sentences, similar to the English speakers in Ferreira (2000). This may place their performance between the heuristic-driven patterns of the GPT-2 models and the more consistent syntactic mappings observed in LLaMA-3. GPT-2 models appear to rely more heavily on good-enough strategies than humans, whereas the attention patterns of LLaMA-3 suggest more syntactically detailed parsing. These results highlight the potential of LLMs as computational proxies for psycholinguistic phenomena and the need to incorporate semantic plausibility cues into neural parsing models. While human participants occasionally rely on shallow heuristics, larger models may attend to semantic and structural cues differently across representational scales. We conclude that both Turkish speakers and LLMs are sensitive to syntactic and semantic constraints, but differ in how they prioritize these linguistic cues.

These findings raise broader questions about the architecture of both human and model-based parsing. In the context of *good-enough* processing, transformer models appear to operate such that earlier layers capture syntax-level information and frequency-based cues, while deeper layers seem to encode information related to real world plausibility. (See similar observations for human sentence processing, not specifically in the context of *good-enough* parsing, e.g., Lowder and Gordon, 2015; and Frazier and Fodor, 1978.) The varia-

[4] For comparability, we also tested the attention weights of BERT and BERTurk (Schweter, 2020) (see Appendix B), bidirectional models for English and Turkish, respectively, and found similar changes in attention, but BERT (compared to BERTurk) had more shifts to the distant NP in the deeper layers, which might be related to the greater distance between the two NPs in English.

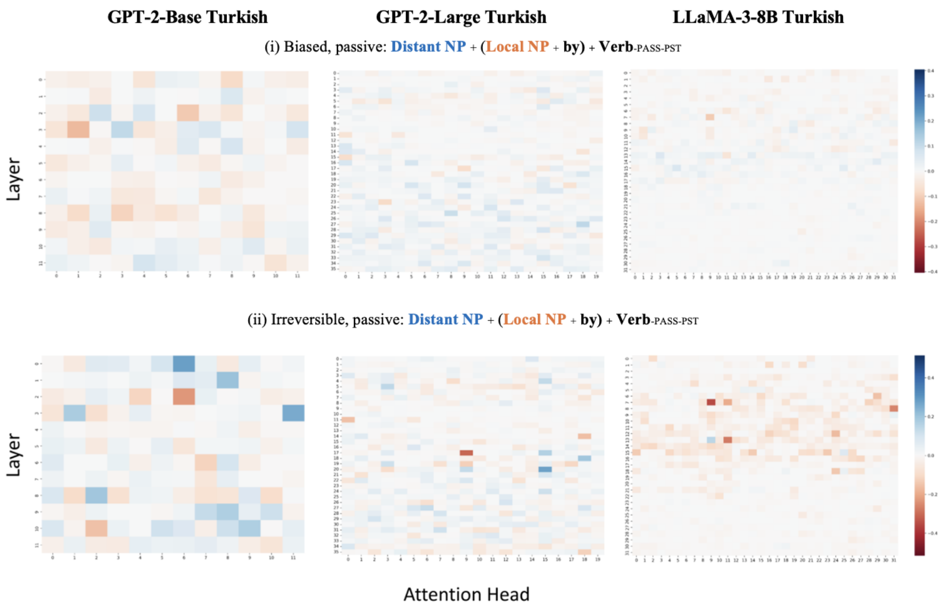


Figure 5: Attention difference between the NPs distant and local to the post-position (in Reversed - Non-reversed word order) for passive constructions. Results are shown for three Turkish LLMs: GPT-2-Base (12 layers, 12 heads), GPT-2-Large (36 layers, 20 heads), and LLaMA-3-8B (32 layers, 32 heads). Blue regions indicate stronger attention to Distant NP (incorrect dependency) in reversed word order, while red regions indicate stronger attention to Local NP (correct dependency. Top row shows differences for biased passive sentences, bottom row does so for irreversible passive sentences.

tion in attention across layers and heads further suggests that different layers may support distinct parsing functions. Future research can examine if a cascaded cue integration approach can be extended to *good-enough* processing in human sentence comprehension. These directions can help clarify how humans and LLMs balance shallow and syntactically-detailed processing.

Limitations and Future Research

We can acknowledge several limitations that require further exploration: (i) We relied on surprisal and attention analyses, but did not include broader generative tasks (e.g., full-sentence completions, direct plausibility ratings from the models). Also, instead of plausibility ratings, agent and patient matching tasks could be used to investigate *good-enough* parsing. Future work could integrate these methods to probe whether humans and LLMs interpret semantically implausible sentences accurately. (ii) Our experiments focused on decoder-only architectures (GPT-2 and LLaMA-3). Other decoder-models available in Turkish like Kanarya (Safaya et al., 2022) or alternative model families, such as T5-based architectures (e.g., TURNA, Uludoğan et al., 2024) and possibly multilingual models like mGPT (Shliazhko et al., 2023) and Aya (Üstün et al., 2024), might yield different patterns of surprisal or attention particularly for language comprehension tasks in Turkish. (iii) Finally, recent work (Giulianelli et al., 2024) has raised concerns about tokenization granularity and argued that token-level language models should ideally be (approximately) marginalized into character-level representations before being used in psycholinguistic studies. Since the current work relies on token-level surprisals, we acknowledge that this may introduce a degree of misalignment with human processing.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments, Mehtap Güven Çoban for her contribution to the translations of the material, the participants of the SPR experiment for their time, Allison Dods and Ümit Atlamaz for valuable feedback on earlier versions of this work.

References

Furkan Akkurt, Onur Gungor, Büşra Marşan, Tunga Gungor, Balkiz Ozturk Basaran, Arzucan Özgür, and Susan Uskudarli. 2024. Evaluating the quality of a corpus annotation scheme using pretrained language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6504–6514, Torino, Italia. ELRA and ICCL.

Samuel Joseph Amouyal, Aya Meltzer-Asscher, and Jonathan Berant. 2024. Large language models for psycholinguistic plausibility pretesting. *arXiv preprint arXiv:2402.05455*.

Douglas Bates et al. 2005. Fitting linear mixed models in r. *R news*, 5(1):27–30.

Kiel Christianson, Andrew Hollingworth, John F. Halliwell, and Fernanda Ferreira. 2001. Thematic Roles Assigned along the Garden Path Linger. *Cognitive Psychology*, 42(4):368–407.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does bert look at? an analysis of bert's attention. *Preprint*, arXiv:1906.04341.

Vera Demberg and Frank Keller. 2009. A computational model of prediction in human parsing: Unifying locality and surprisal effects. In *Proceedings of the annual meeting of the cognitive science society*, volume 31. Issue: 31.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

Fernanda Ferreira. 2000. Syntactic Vulnerability and Thematic Misinterpretation. page 64.

Fernanda Ferreira, Karl GD Bailey, and Vittoria Ferraro. 2002. Good-enough representations in language comprehension. *Current directions in psychological science*, 11(1):11–15.

Fernanda Ferreira and Nikole D Patson. 2007. The 'good enough' approach to language comprehension. *Language and linguistics compass*, 1(1-2):71–83.

Lyn Frazier and Janet Dean Fodor. 1978. The sausage machine: A new two-stage parsing model. *Cognition*, 6(4):291–325. Publisher: Elsevier.

Mario Giulianelli, Luca Malagutti, Juan Luis Gastaldi, Brian DuSell, Tim Vieira, and Ryan Cotterell. 2024. On the proper treatment of tokenization in psycholinguistics. *arXiv preprint arXiv:2410.02691*.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *Preprint*, arXiv:1901.05287.

Selçuk İşsever. 2003. Information structure in turkish: the word order–prosody interface. *Lingua*, 113(11):1025–1053.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, Anna A Ivanova, et al. 2024. Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. Association for Computational Linguistics (ACL).

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.

H Toprak Kesgin, M Kaan Yuce, Eren Dogan, M Egemen Uzun, Atahan Uz, Elif İnce, Yusuf Erdem, Osama Shbib, Ahmed Zeer, and M Fatih Amasyali. 2024a. Optimizing large language models for turkish: New methodologies in corpus selection and training. In *2024 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6. IEEE.

H Toprak Kesgin, M Kaan Yuce, Eren Dogan, M Egemen Uzun, Atahan Uz, H Emre Seyrek, Ahmed Zeer, and M Fatih Amasyali. 2024b. Introducing cosmosgpt: Monolingual training for turkish language models. *arXiv preprint arXiv:2404.17336*.

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

Andrew Li, Xianle Feng, Siddhant Narang, Austin Peng, Tianle Cai, Raj Sanjay Shah, and Sashank Varma. 2024. Incremental comprehension of garden-path sentences by large language models: Semantic interpretation, syntactic re-analysis, and attention. *arXiv preprint arXiv:2405.16042*.

Matthew W Lowder and Peter C Gordon. 2015. Focus takes time: Structural effects on reading. *Psychonomic Bulletin & Review*, 22:1733–1738.

Rebecca Marvin and Tal Linzen. 2018. Targeted Syntactic Evaluation of Language Models. *arXiv preprint*. ArXiv:1808.09031 [cs].

Sathvik Nair and Philip Resnik. 2023. Words, Subwords, and Morphemes: What Really Matters in the Surprisal-Reading Time Relationship? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11251–11260, Singapore. Association for Computational Linguistics.

Metehan Oğuz, Yusuf Umut Ciftci, and Yavuz Faruk Bakman. 2024. Do llms recognize me, when i is not me: Assessment of llms understanding of turkish indexical pronouns in indexical shift contexts. *arXiv preprint arXiv:2406.05569*.

Byung-Doh Oh and William Schuler. 2022. Entropy- and distance-based predictors from gpt-2 attention patterns predict reading times over and above gpt-2 surprisal. *arXiv preprint arXiv:2212.11185*.

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.

Paolo Pedinotti, Giulia Rambelli, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. 2021. Did the Cat Drink the Coffee? Challenging Transformers with Generalized Event Knowledge. *arXiv preprint*. ArXiv:2107.10922 [cs].

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.

Ali Safaya, Emirhan Kurtuluş, Arda Goktogan, and Deniz Yuret. 2022. Mukayese: Turkish NLP strikes back. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 846–863, Dublin, Ireland. Association for Computational Linguistics.

Lavinia Salicchi, Emmanuele Chersoni, and Alessandro Lenci. 2023. A study on surprisal and semantic relatedness for eye-tracking data prediction. *Frontiers in Psychology*, 14:1112365. Publisher: Frontiers.

Stefan Schweter. 2020. Berturk-bert models for turkish. *Zenodo*, 2020:3770924.

Rico Sennrich. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

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.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Philip Levy. 2022. Large-scale evidence for logarithmic effects of word predictability on reading time. Publisher: PsyArXiv.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2023. mgpt: Few-shot learners go multilingual. *Preprint*, arXiv:2204.07580.

Gökçe Uludoğan, Zeynep Yirmibeşoğlu Balal, Furkan Akkurt, Melikşah Türker, Onur Güngör, and Susan Üsküdarlı. 2024. Turna: A turkish encoder-decoder language model for enhanced understanding and generation. *arXiv preprint arXiv:2401.14373*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, \Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *Preprint*, arXiv:1905.09418.

Daphne Wang, Mehrnoosh Sadrzadeh, Miloš Stanojević, Wing-Yee Chow, and Richard Breheny. 2024. How can large language models become more human? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 166–176.

Ethan G Wilcox. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.

Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023a. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023b. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470. Publisher: MIT Press One Broadway, 12th Floor, Cambridge, Massachusetts 02142, USA .

Jérémy Zehr and Florian Schwarz. 2018. PennController for Internet Based Experiments (IBEX).

A Model Surprisal in English

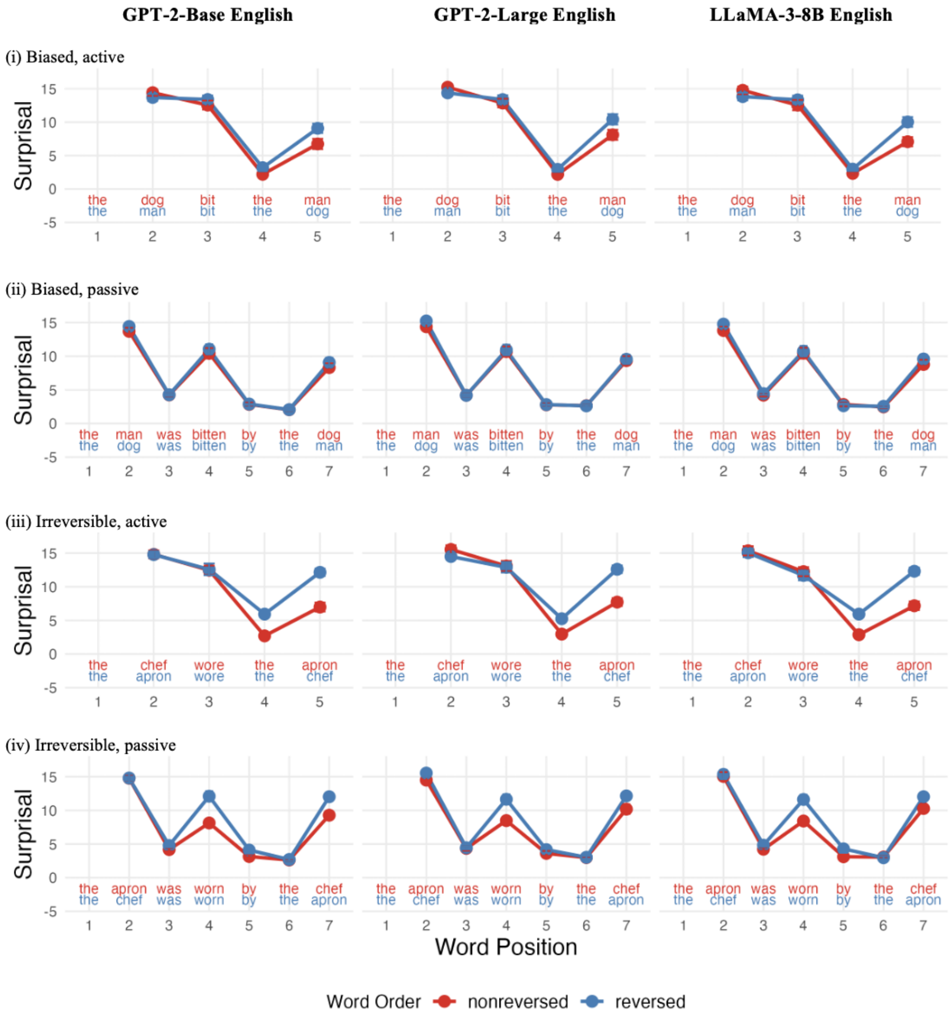


Figure 6: Word-by-word Model Surprisal for the English BERT and BERTurk (Turkish BERT) by Set (Biased, Irreversible), Structure (Active, Passive), and Word Order (Reversed, Nonreversed)

B Comparing English and Turkish: Encoder Model Attention

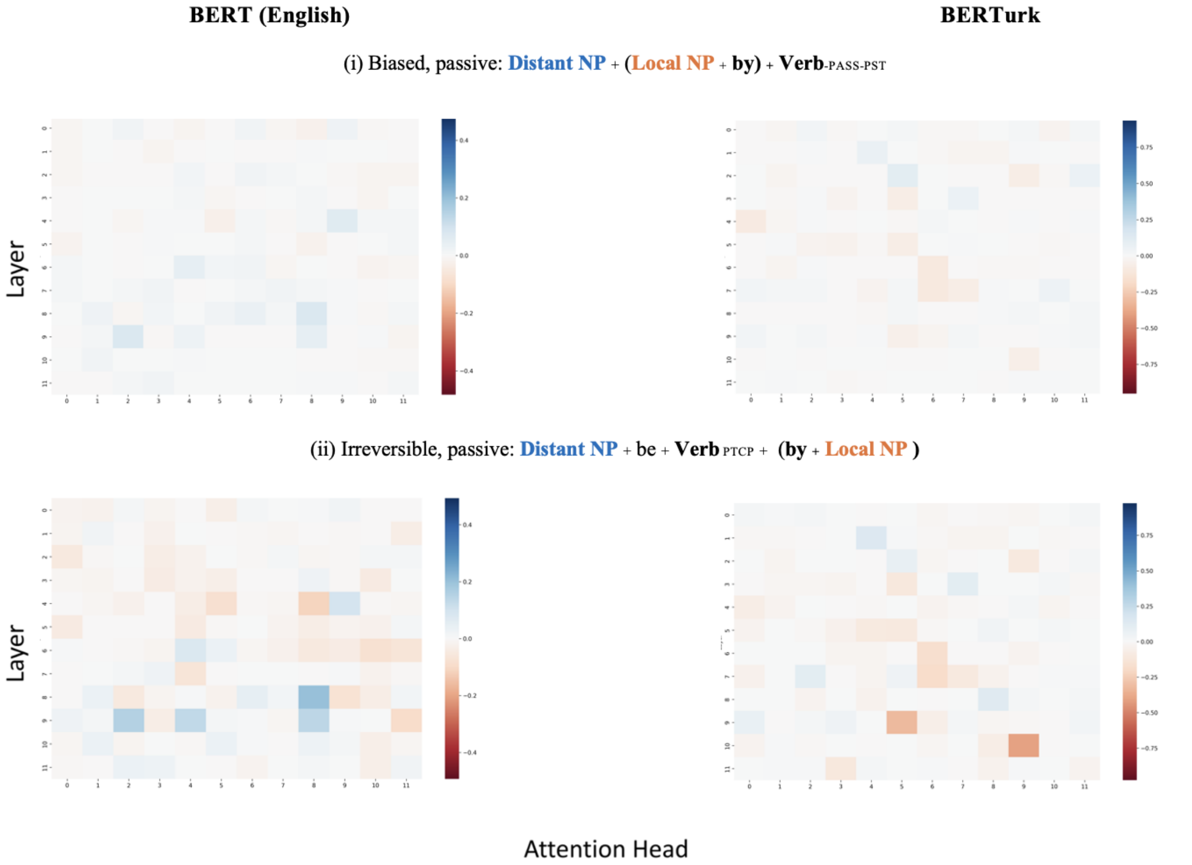


Figure 7: Attention difference between the NPs distant and local to the post-position (in Reversed - Non-reversed word order) for passive constructions, comparing the English BERT and BERTurk (Turkish BERT) (12 layers, 12 heads). Blue regions indicate stronger attention to Distant NP in reversed word order, while red regions indicate stronger attention to Local NP. Top row shows differences for biased passive sentences, bottom row does so for irreversible passive sentences.