What to Predict? Exploring How Sentence Structure Influences Contrast Predictions in Humans and Large Language Models

Shuqi Wang **Xufeng Duan** **Zhenguang Cai**
Department of Linguistics and Modern Languages, CUHK
{shuqiwang, xufeng.duan}@link.cuhk.edu.hk, zhenguangcai@cuhk.edu.hk

Abstract

This study examines how sentence structure shapes contrast predictions in both humans and large language models (LLMs). Using Mandarin ditransitive constructions — double object (DO, "She gave the girl the candy, but not…") vs. prepositional object (PO, "She gave the candy to the girl, but not…") as a testbed, we employed a sentence continuation task involving three human groups (written, spoken, and prosodically normalized spoken stimuli) and three LLMs (GPT-4o, LLaMA-3, and Qwen-2.5). Two principal findings emerged: (1) Although human participants predominantly focused on the theme (e.g., "the candy"), contrast predictions were significantly modulated by sentence structure—particularly in spoken contexts, where the sentence-final element drew more attention. (2) While LLMs showed a similar reliance on structure, they displayed a larger effect size and more closely resembled human spoken data than written data, indicating a stronger emphasis on linear order in generating contrast predictions. By adopting a unified psycholinguistic paradigm, this study advances our understanding of predictive language processing for both humans and LLMs and informs research on human–model alignment in linguistic tasks.

1 Introduction

Predictive processing is fundamental to how both humans and large language models (LLMs) handle language. When people read or listen, they continuously anticipate upcoming words and meanings, facilitating swift integration of new information and maintaining efficient comprehension (Altmann & Kamide, 1999; Christiansen & Chater, 2016; Clark, 2013; Kuperberg & Jaeger, 2016; Pickering & Gambi, 2018). Prediction also underlies language use in LLMs, as these models are explicitly designed to predict the next token in a sequence (Brown et al., 2020; Radford et al., 2018, 2019).

As a linguistic cue, **contrast** plays a key role by guiding attention toward the most distinctive or unexpected element in the context and prompting the prediction of an alternative (Repp, 2010; Rooth, 2016). Contrast often involves opposing or comparing one element to another of the same semantic type (Roberts, 2012). It is often signaled by a negation operator (e.g., "not") or discourse markers (e.g., "but"). Empirical findings indicate that human comprehenders are highly attuned to these cues. Upon encountering contrast markers, they actively anticipate an alternative that stands in contrast to a previously mentioned element (Carlson, 2014; Lowder & Ferreira, 2016).

A crucial question thus arises: which preceding element is being contrasted and which potential alternatives should be predicted? Contrast closely intersects with focus — the most emphasized or central constituent (Calhoun, 2009; Husband & Ferreira, 2016; Lowder & Gordon, 2015; Repp, 2010). Thus, the element chosen for contrast is often the sentence's focus. However, determining focus can become complicated in lengthy or structurally complex sentences. For instance, consider the ditransitive construction "She gave the girl the candy, not…". The focus—and therefore the contrast—could fall on the recipient ("the girl"), the theme ("the candy"), or the verb ("gave"). Depending on which element is in focus, comprehenders might predict contrasting recipients (e.g., "the father," "the boy"), contrasting themes (e.g., "the toy," "the cake"), or contrasting verbs (e.g., "bought," "made"). This is because **ditransitive structure** introduces multiple arguments and allows flexible constituent orders, complicating the task of pinpointing the focal element and thus the likely contrast (Paterson et al., 2007; Shyu, 2010).

This study thus uses ditransitive structure as a testbed for understanding how humans and LLMs

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 244–252
May 3, 2025 ©2025 Association for Computational Linguistics

predict contrasts, and how sentence structure influences these predictions. Two primary constructions of the ditransitive pattern are the double object (DO) construction (e.g., "She bought the girl the candy") and the prepositional object (PO) construction (e.g., "She gave the candy to the daughter"). Critically, in Mandarin Chinese, these two constructions use the same set of segmental materials but in different orders (DO: 她送给了女孩糖果, literally "She gave to girl candy"; PO: 她送了糖果给女孩, literally "She gave candy to girl"), making them ideal for examining how linear arrangement affects focus and contrast.

Three main hypotheses address the potential locus of focus and, by extension, the nature of contrast predictions in ditransitive sentences:

1. The **Sentence-Final Hypothesis** posits that the focus tends to fall at the end of the sentence (Xu, 2004; Yan & Calhoun, 2020), predicting that DO sentences would contrast the theme (e.g., "not the candy") and PO sentences would contrast the recipient (e.g., "not the girl").
2. The **Thematic Hierarchy Hypothesis** proposes that focus falls on the element highest in the thematic hierarchy, namely the theme, which is more closely related to the verb (Shyu, 2010). Thus, both DO and PO constructions would yield the focus on the theme, leading to identical contrast predictions.
3. The **Verb-Dominant Hypothesis** claims that the verb or entire verb phrase is focused (Carlson, 2014; Roettger et al., 2021). In this scenario, both DO and PO constructions would lead comprehenders to predict a verb-related contrast (e.g., "not bought" or "not sang a song").

Notably, the latter two hypotheses predict similar outcomes for DO and PO forms, while the first emphasizes a structural effect tied to word order. Because written language is presented at once for readers, its linear-order impact may be weaker than in spoken language, where information unfolds sequentially (Ferreira & Anes, 1994), we employ both written and spoken stimuli in human experiments to determine whether modality modulates the influence of sentence structure on contrast prediction.

Taken together, this work aims to address two core questions:

1. How do humans predict contrasts in Mandarin ditransitive constructions and how does sentence structure modulate these predictions in both written and spoken contexts?
2. How do large language models predict contrasts in the same constructions, and how similar are these predictions to human behavior?

2 Methods

2.1 Design and Materials

We employed a sentence continuation task to examine how humans and LLMs predict and complete contrasts in ditransitive sentences. A total of 42 experimental items were created based on previous studies (Cai et al., 2013, 2022), each consisting of a ditransitive construction followed by a contrast marker. Each item appeared in two conditions: a DO construction (e.g., 她送给了女孩糖果，而不是…; "She gave the girl the candy, but not…") and a PO construction (e.g., 她送了糖果给女孩，而不是…; "She gave the candy to the girl, but not…"). We selected 14 ditransitive verbs (e.g., 买 'buy', 交 'hand', 借 'lend', 卖 'sell', 奖 'award', 带 'bring', 扔 'throw', 抛 'toss', 拿 'take', 捐 'donate', 让 'give away', 还 'return', 送 'send', 递 'pass'), each appearing in three items, yielding 42 experimental preambles.

To reduce participants' focus on contrast markers and maintain variety, we incorporated 90 filler sentences. Each filler contained different structures and a connective (e.g., "because," "so," "then"). This design aimed to ensure that participants engaged with the full range of sentence structures and did not develop a strategy specific to the contrast condition.

We used Microsoft Azure to generate spoken versions of the experimental items. Specifically, we selected a male adult speaker of simplified Mandarin ("Yunyang") at a speed of 0.75 and exported the files at 48 kHz. Two types of spoken stimuli were created: The first one is the original recording from Azure. These versions contained natural variations in sentence-final stress, such that DO sentences ended with a higher pitch and longer duration on the theme, whereas PO sentences ended with a higher pitch and longer duration on the recipient; The second type is the normalized recording, where segments from one condition were replaced with those from the other (counterbalanced between two conditions) and also added white noise to standardize duration. As a result, all segmental and suprasegmental features are the same in both conditions.

Our motivation for including these two types of spoken stimuli was twofold. First, the original version reflected more natural spoken processing, capturing how individuals predicted contrasts in everyday speech contexts. Second, the normalized version controlled for prosodic differences, allowing us to focus on the role of syntactic structure and word order in shaping contrast predictions.

2.2 Human experiments

2.2.1 Participants

A total of 164 native Mandarin speakers participated in this study, divided into three groups based on the type of stimuli they received: 52 for the written stimuli, 57 for the original spoken stimuli, and 55 for the normalized spoken stimuli. Following data screening (e.g., incomplete responses, procedural errors), we excluded some participants' responses. This resulted in 50 participants in the written group (27 females, 23 males; *M* age = 21.5), 50 participants in the original spoken group (17 females, 33 males; *M* age = 21.8), and 48 participants in the normalized spoken group (18 females, 30 males; *M* age = 21.3).

2.2.2 Procedure

All experiments were conducted online using Qualtrics (Qualtrics, 2024). Participants joined a Zoom session, shared their screen, and began the experiment while the researcher monitored their progress. This arrangement helped mitigate potential issues associated with online data collection, such as inattentiveness or lack of engagement. After providing informed consent, participants read on-screen instructions and examples explaining the sentence continuation task. They were asked to type the first, most natural completion that came to mind for each presented sentence preamble. Each stimulus was presented on a separate Qualtrics page.

Two lists of stimuli were created, with each item appearing in only one condition (DO or PO) in each list. Participants were randomly assigned to one of these lists. Upon completing the task, they provided demographic information and received a payment of 30 RMB in appreciation for their time.

2.3 LLMs Experiments

2.3.1 Models

We employed three LLMs in this study: OpenAI's GPT-4o (OpenAI, 2024), Meta's LLaMA-3 (Meta, 2024), and Alibaba's Qwen-2.5 (Yang et al., 2025). These models were chosen for three main reasons. First, they each represented state-of-the-art performance at the time of the study. Second, they allowed us to compare closed-source (GPT-4o) with open-weight (LLaMA-3 and Qwen-2.5) systems. Finally, we included English-dominant LLMs (GPT-4o and LLaMA-3) alongside a Chinese-dominant LLM (Qwen-2.5), ensuring coverage of different training backgrounds and linguistic emphases.

2.3.2 Procedure

We collected output from the three language models using an R package called "MacBehavior" (Duan et al., 2024), which was specifically developed for behavioral experimentation with large language models. The same stimuli given to human participants were presented to each model under a "one-trial-per-run" configuration. In this setup, each prompt–stimulus pair was input into the model in a new conversation, ensuring that no trial could be influenced by preceding prompts. The prompts mirrored the instructions given to human participants — "Please read the first half of a sentence and fill in the first word or phrase that comes to mind. Make the sentence complete, natural, and reasonable. The first half of the sentence is:". We conducted 50 sessions for each of the three LLMs, so each item received a total of 50 responses. This design roughly matched the sample size in the human experiments.

3 Analyses and results

3.1 Data coding

We employed the same coding scheme for both human and LLM continuations, categorizing responses into three main contrast types: (1) Theme Contrast (T): The continuation contrasts the theme (e.g., "the cake" in "She gave the girl the candy, but not the cake."). (2) Recipient Contrast (R): The continuation contrasts the recipient (e.g., "the boy"

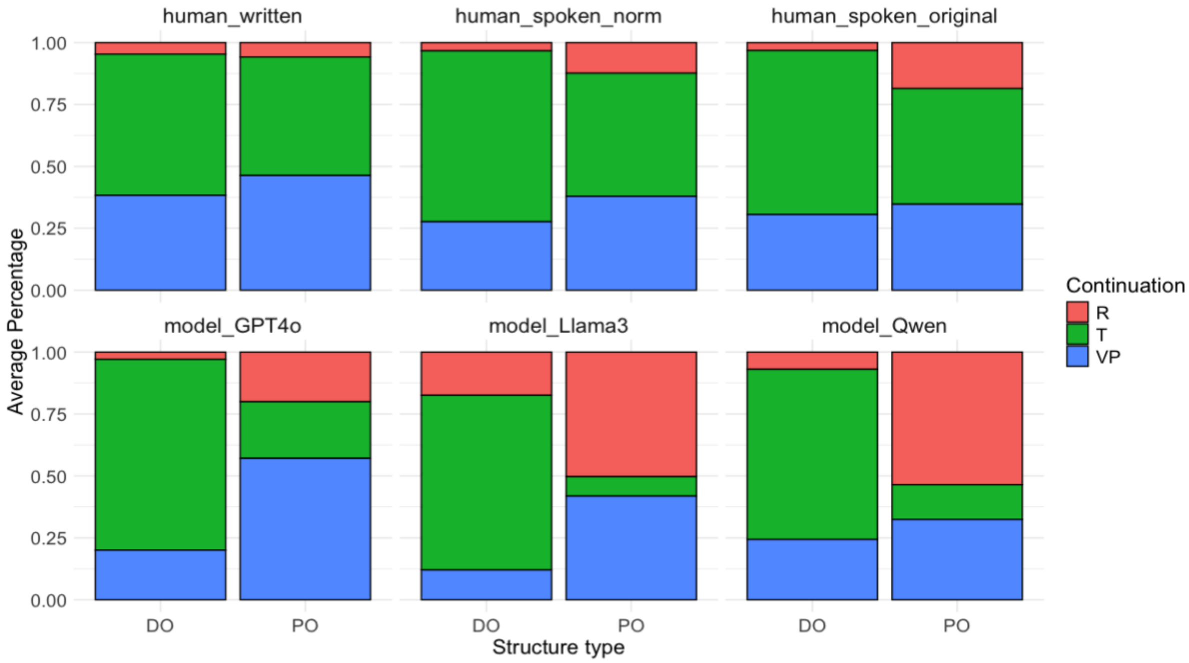

Figure 1 Average percentages of three types of continuations for ditransitive structures in human participants (top panel) and LLMs (bottom panel)

in “She gave the girl the candy, but not the boy.”). (3) Verb or Verb Phrase Contrast (VP): The continuation contrasts the verb or verb phrase (e.g., “buy” or “sing a song” in “She gave the girl the candy, but didn’t buy her one / sing a song to her.”).

3.2 Statistical analysis

We adopted a three-step approach to analyze our data. First, we examined which continuation type was most prevalent across Mandarin ditransitive sentences. To this end, we performed a series of t-tests comparing the mean frequencies of these three contrast types. Second, to determine whether sentence structure significantly influenced contrast predictions, we conducted a linear mixed-effects model analysis. We began by aggregating responses by item, calculating the percentage of each continuation type for each item. We then performed a by-item analysis with structure type (DO vs. PO), continuation type (T, R and VP; With T as reference level), and group (humans vs. LLMs; Pairwise comparisons were conducted between each human modality and model, resulting in a total of nine comparisons) as fixed effects, and item as a random effect. Finally, to further assess how closely the LLM predictions aligned with human continuations (both written and spoken), we calculated Pearson correlations between the LLMs’ aggregated prediction patterns and those in the human experiments. This approach allowed us to gauge the degree of similarity in contrast prediction patterns across the different groups.

3.3 Results

3.3.1 Human results

First, we performed the T-test to investigate which continuation type was the dominant. Across all three participant groups (written, original spoken, and normalized spoken), theme contrasts emerged as the most frequent continuation type, followed by verb contrasts and then recipient contrasts (see Figure 1). Specifically, in the written stimuli group, participants produced more theme ($M = 0.52$) than verb ($M = 0.42$) contrasts, $t(4183.5) = -6.56$, $p < .001$, while verb contrasts also exceeded recipient ($M = 0.05$, $t(2911.3) = 31.28$, $p < .001$). Similarly, in the original spoken stimuli group, theme ($M = 0.56$) contrasts were more frequent than verb ($M = 0.3274$) contrasts, $t(4169.4) = -15.88$, $p < .001$, and verb contrasts again exceeded recipient ($M = 0.11$), $t(3626.8) = 17.83$, $p < .001$. The normalized spoken stimuli showed the same pattern: theme ($M = 0.60$) contrasts dominated verb ($M = 0.32$) contrasts, $t(4006.3) = -18.82$, $p < .001$, which in turn were more frequent than recipient ($M = 0.08$) contrasts, $t(3219.7) = 19.92$, $p < .001$. Overall, these findings support the Thematic Hierarchy Hypothesis, suggesting that the theme is consistently viewed as the primary focal element for contrast in ditransitive constructions.

Second, the structure can modulate the contrast predictions in both written and spoken modalities. That is, in all three groups, the difference between recipient contrast and theme contrast was larger in the PO condition than in the DO condition (written stimuli group: $\beta = 0.10$, $SE = 0.02$, $t(164) = 5.82$, $p < .001$; original spoken stimuli group: $\beta = 0.35$, $SE = 0.02$, $t(164) = 15.25$, $p < .001$; normalized spoken stimuli group: $\beta = 0.28$, $SE = 0.03$, $t(205) = 8.49$, $p < .001$), as shown in the top panel of Figure 1. This pattern suggests a sentence-final bias in the focus locus and, consequently, in participants' contrast predictions—partially supporting the Sentence-Final Hypothesis.

Third, the modulation effect of structure is larger for spoken language than for written language (written vs. original spoken: $\beta = 0.25$, $SE = 0.03$, $t(410) = 7.68$, $p < .001$; written vs. normalized spoken: $\beta = 0.18$, $SE = 0.04$, $t(451) = 4.59$, $p < .001$), as shown in the top panel of Figure 1. These findings indicate that spoken language amplifies the impact of structural differences (DO vs. PO) on how listeners predict contrast, whereas this effect is comparatively reduced in written language. Moreover, the non-significant difference between original and normalized spoken data (original spoken vs. normalized spoken: $\beta = 0.07$, $SE = 0.04$, $t(451) = 1.688$, $p = .092$) suggests that prosody alone may not fully explain the stronger structure effect in speech; rather, linear-order presentation may heighten the prominence of sentence-final elements in spoken modalities.

3.3.2 Model Results

Similar to the human data, t-tests showed that theme contrast was the primary continuation type for all three models (GPT-4o, LLaMA-3, and Qwen-2.5). However, the relative ranking of verb phrase and recipient contrasts differed across models. In GPT-4o, theme contrast ($M = 0.50$) is greater than verb phrase contrast ($M = 0.39$), $t(4137.4) = -7.11$, $p < .001$, and verb phrase contrast exceeded recipient contrast ($M = 0.11$), $t(3548.1) = 21.67$, $p < .001$. In LLaMA-3, theme contrast ($M = 0.40$) surpassed recipient contrast ($M = 0.33$), $t(3733.6) = 4.49$, $p < .001$, and recipient exceeded verb phrase(M = 0.27), $t(3727.8) = -4.00$, p < .001. Finally, in Qwen-2.5, theme ($M = 0.42$) remained significantly higher than verb ($M = 0.28$), $t(4159.4) = -9.22$, $p < .001$, whereas the difference between verb and recipient ($M = 0.30$) was non-significant, $t(4192.2) = -1.46$, $p = .14$.

Similarly, the structure of the ditransitive sentences modulated contrast predictions in all three models. The difference between recipient contrast and theme contrast was significantly larger under PO constructions than under DO constructions for all three models (GPT-4o: $\beta = 0.71$, $SE = 0.06$, $t(164) = 12.01$, p < .001; LLaMA-3: $\beta = 0.96$, $SE = 0.06$, $t(164) = 17.10$, p < .001; Qwen-2.5: $\beta = 1.00$, $SE = 0.09$, $t(164) = 11.42$, p < .001), indicating that the models generated more theme contrasts in DO (than in PO) and more recipient contrasts in PO (than in DO). This aligns with the human pattern of sentence-final bias on contrast predictions.

3.3.3 Comparing Humans and Models

Having established that sentence structure influenced contrast predictions for both humans and models, we next examined whether the magnitude of this influence differed between the two groups. Across all comparisons, the three models exhibited a larger structural effect than their human counterparts, regardless of whether the human data were drawn from the written, original spoken, or normalized spoken conditions (GPT-4o vs. human: $\beta = 0.61$, $SE = 0.07$, $t(451) = 8.62$, $p < .001$ (written); $\beta = 0.36$, $SE = 0.07$, $t(451) = 4.85$, $p < .001$ (original spoken); $\beta = 0.43$, $SE = 0.08$, $t(451) = 5.50$, $p < .001$ (normalized spoken). LLaMA-3 vs. human: $\beta = 0.85$, $SE = 0.07$, $t(410) = 12.07$, $p < .001$ (written); $\beta = 0.61$, $SE = 0.07$, $t(410) = 8.60$, $p < .001$ (original spoken); $\beta = 0.67$, $SE = 0.07$, $t(410) = 9.10$, $p < .001$ (normalized spoken). Qwen-2.5 vs. human: $\beta = 0.90$, $SE = 0.10$, $t(410) = 8.83$, $p < .001$ (written); $\beta = 0.65$, $SE = 0.10$, $t(410) = 6.30$, $p < .001$ (original spoken); $\beta = 0.90$, $SE = 0.10$, $t(410) = 8.83$, $p < .001$ (normalized spoken)). These results indicate that all three models were more sensitive to structural differences between the DO and PO constructions than human participants. In other words, although humans and LLMs both adjust their contrast predictions based on sentence structure, the magnitude of this adjustment is notably larger in LLMs.

To assess the similarity of contrast prediction patterns between humans and LLMs, we computed Pearson correlations for each model and each type

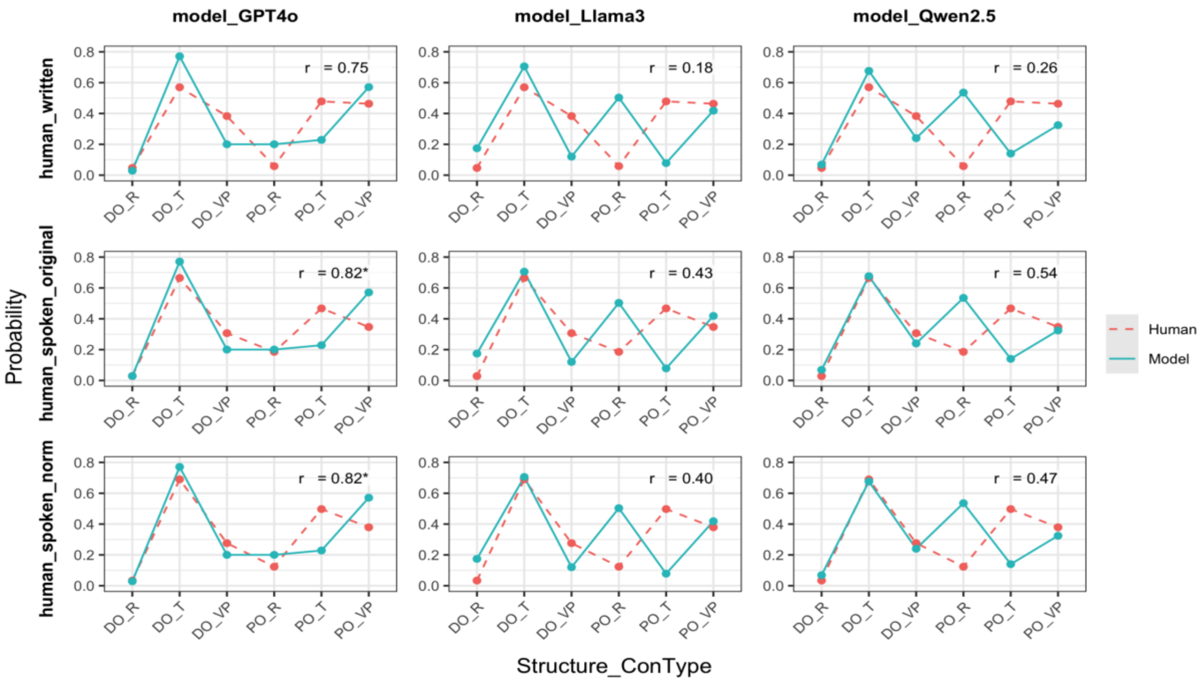

Figure 2 Human-model correlations in contrast predictions across structures and modalities. Each subplot compares a model's prediction probabilities (blue solid line) with human responses (red dashed line) under one of the three modality conditions (written, original spoken, or normalized spoken). The x axis combines structure type (DO, PO) and continuation type (R, T, VP), yielding six categories. The y axis indicates the probability of each category. Pearson correlation coefficients (r) show how closely each model align with human data.

of human data. As shown in Figure 2, two main findings emerged. First, GPT-4o showed the highest correlation with human data across all modalities, suggesting it was more human-like in its contrast predictions compared to LLaMA-3 or Qwen2.5. Second, all three models correlated more strongly with the spoken human data than with the written data, indicating that these models align better with the structure-incremental nature of speech (GPT-4o vs. human: r = 0.75, p = .083 (written); r = 0.82, p = <.05 (original spoken); r = 0.82, p = <.05 (normalized spoken). LLaMA-3 vs. human: r = 0.18, p = .734 (written); r = 0.43, p = .390 (original spoken); r = 0.39, p = .439 (normalized spoken) Qwen-2.5 vs. human: r = 0.26, p = .625 (written); r = 0.54, p = .263 (original spoken); r = 0.47, p = .351 (normalized spoken)).

4 Discussion

The present study investigated how humans and LLMs predict contrasts in Mandarin ditransitive constructions, focusing on whether sentence structure modulates these predictions and whether written or spoken modality influences the size of this structural effect. Our data revealed two primary results. First, for humans, although theme contrast was the most frequent continuation overall, sentence structure significantly modulated contrast predictions, with a stronger effect in spoken language than in written language. Second, LLMs showed an even stronger structure effect than humans, particularly GPT-4o, which most closely mirrored human data.

First, our data addressed a key theoretical linguistic question: which element in a Mandarin ditransitive sentence is in focus, thereby prompting contrast-based predictions? Human data showed that in both DO and PO constructions, the theme was consistently the focal element. This outcome aligns with the Thematic Hierarchy Hypothesis, which argues that the theme, closely tied to the verb, tends to be the default focus in ditransitive structures sentence (Shyu, 2010).

Critically, sentence structure also modulated how human participants predicted contrast in ditransitive structures. In the PO construction, there were more recipient contrasts predictions than in DO construction, because in PO construction (e.g., "she gave the candy to the daughter"), the recipient appears at the end of the sentence and thus draws more attention and induce more predictions that stand contrast with it. This finding partially supports the Sentence-Final Hypothesis (Xu, 2004;

Yan & Calhoun, 2020), which posits that focus naturally gravitates toward the last element in the sentence. Nevertheless, theme contrasts remained dominant across both DO and PO constructions, suggesting that linear order competes with overarching thematic structure in directing attention.

We further observed a difference between written and spoken modalities. In spoken language, participants exhibited a more pronounced effect of word order: sentence-final constituents in PO constructions attracted more recipient contrasts than in DO constructions. This enhanced contrast may stem from the incremental nature of speech (Ferreira & Anes, 1994), as listeners cannot revisit earlier segments and thus rely heavily on each new chunk of information. Interestingly, normalizing prosody did not attenuate the structural effect (i.e., no significant difference between original spoken group and normalized spoken group). While intonation can highlight final elements in Mandarin, our findings suggest that linear order alone can drive substantial focus-based predictions, emphasizing the importance of modality in shaping how comprehenders allocate attention.

Turning to LLMs, we found that each model exhibited a larger structural effect than any of the human groups. Similar to human participants, the models produced more contrasts on sentence-final arguments, but the magnitude of this tendency was amplified. Two factors may underlie this difference. First, transformer-based LLMs use positional embeddings to encode token order (Vaswani et al., 2017), which makes recently processed tokens more salient. This feature can mimic, yet also exaggerate, spoken-language emphasis on final constituents. Second, LLMs are trained with a next-token prediction objective on large text corpora, which could favor the final parts of a sequence, as the model aims to reduce prediction loss by paying attention to the most recent context. Our correlation analysis further revealed that all three LLMs resembled spoken human data more closely than written data, suggesting that next-token prediction architectures may align more naturally with the incremental processing profile of speech.

Together, these findings contribute to broader discussions about predictive processing in language. Although both humans and neural language models depend on anticipatory mechanisms (Brown et al., 2020; Pickering & Gambi, 2018), their respective mechanisms may diverge in how strongly they weight syntactic position over other linguistic cues. Our results also highlight that the models' predictive behavior bears closer resemblance to the incremental unfolding of speech than to the flexible reading patterns of silent text comprehension (Christiansen & Chater, 2016). Future research could employ more fine-grained methods (e.g., eye-tracking) and analyses (e.g., attention-weight examinations of LLMs) to investigate why humans and models display these similarities and discrepancies.

5 Conclusions

The current study employed a sentence continuation task to examine how humans and LLMs predict contrast in ditransitive sentences. Two main findings emerged: (1) theme contrasts were dominant for human participants, but sentence structure significantly modulated these contrasts—especially in spoken contexts; (2) LLMs showed stronger structural effects than humans, with GPT-4o aligning most closely with human data. This study highlights the interplay between syntactic structure and modality in guiding predictions in human language processing and offers a clearer lens into how humans and LLMs differ in their weighting sentence structure. By comparing the two groups in a straightforward task, this work offers practical insights for refining language models and yields theoretical implications for understanding predictive language processing across modalities.

Limitations

The first limitation is our reliance on the sentence continuation paradigm, which is offline and intermingles comprehension with production. Incorporating more online and time-sensitive methods like eye-tracking or neuroimaging methods could provide a clearer picture of when and how focus-based contrast predictions arise.

Moreover, although we sampled three prominent LLMs, the rapid evolution of language models suggests that further comparative studies would be valuable—particularly among systems trained mainly on Chinese text vs. models relying heavily on English corpora.

References

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Cai, Z. G., Pickering, M. J., & Sturt, P. (2013). Processing verb-phrase ellipsis in Mandarin Chinese: Evidence against the syntactic account. *Language and Cognitive Processes*, *28*(6), 810–828. https://doi.org/10.1080/01690965.2012.665932

Cai, Z. G., Zhao, N., & Pickering, M. J. (2022). How do people interpret implausible sentences? *Cognition*, *225*, 105101. https://doi.org/10.1016/j.cognition.2022.105101

Calhoun, S. (2009). What makes a word contrastive? Prosodic, semantic and pragmatic perspectives. *Where Prosody Meets Pragmatics: Research at the Interface*, *8*, 53–78.

Carlson, K. (2014). Predicting contrast in sentences with and without focus marking. *Lingua*, *150*, 78–91. https://www.sciencedirect.com/science/article/pii/S0024384114001624

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62. https://doi.org/10.1017/S0140525X1500031X

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Duan, X., Li, S., & Cai, Z. G. (2024). MacBehaviour: An R package for behavioural experimentation on large language models. *Behavior Research Methods*, *57*(1), 19. https://doi.org/10.3758/s13428-024-02524-y

Ferreira, F., & Anes, M. (1994). Why study spoken language? In M. A. Gernsbacher (Ed.), *Handbook of Psycholinguistics* (pp. 32–56). Academic Press. https://psycnet.apa.org/record/1994-97824-002

Husband, E. M., & Ferreira, F. (2016). The role of selection in the comprehension of focus alternatives. *Language, Cognition and Neuroscience*, *31*(2), 217–235. https://doi.org/10.1080/23273798.2015.1083113

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Lowder, M. W., & Ferreira, F. (2016). Prediction in the processing of repair disfluencies: Evidence from the visual-world paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(9), 1400–1416. https://doi.org/10.1037/xlm0000256

Lowder, M. W., & Gordon, P. C. (2015). Focus takes time: Structural effects on reading. *Psychonomic Bulletin & Review*, *22*(6), 1733–1738. https://doi.org/10.3758/s13423-015-0843-2

Meta. (2024, April 18). *Introducing Meta Llama 3: The most capable openly available LLM to date*. https://ai.meta.com/blog/meta-llama-3/

OpenAI. (2024, May 13). *GPT-4o system card*. https://openai.com/index/gpt-4o-system-card/

Paterson, K. B., Liversedge, S. P., Filik, R., Juhasz, B. J., White, S. J., & Rayner, K. (2007). Focus Identification during Sentence Comprehension: Evidence from Eye Movements. *Quarterly Journal of Experimental Psychology*, *60*(10), 1423–1445. https://doi.org/10.1080/17470210601100563

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, *144*(10), 1002–1044. https://doi.org/10.1037/bul0000158

Qualtrics (Versions 09-2024). (2024). [Computer software]. Qualtrics.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf

Repp, S. (2010). Defining ‘contrast’as an information-structural notion in grammar. *Lingua*, *120*(6), 1333–1345. https://doi.org/10.1016/j.lingua.2009.04.006

Roberts, C. (2012). Information Structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, *5*, 6:1-69. https://doi.org/10.3765/sp.5.6

Roettger, T. B., Franke, M., & Cole, J. (2021). Positional biases in predictive processing of intonation. *Language, Cognition and Neuroscience*, *36*(3), 342–370. https://doi.org/10.1080/23273798.2020.1853185

Rooth, M. (2016). Alternative Semantics. In C. Féry & S. Ishihara (Eds.), *The Oxford Handbook of Information Structure* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199642670.013.19

Shyu, S.-I. (2010). Focus interpretation of zhi 'only' associated arguments in Mandarin triadic constructions. *Linguistics*, *48*(3). https://doi.org/10.1515/ling.2010.021

Xu, L. (2004). Manifestation of informational focus. *Lingua*, *114*(3), 277–299. https://doi.org/10.1016/S0024-3841(03)00031-7

Yan, M., & Calhoun, S. (2020). Rejecting false alternatives in Chinese and English: The interaction of prosody, clefting, and default focus position. *Laboratory Phonology*, *11*(1). https://doi.org/10.5334/labphon.255

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., … Qiu, Z. (2025). *Qwen2.5 Technical Report* (No. arXiv:2412.15115). arXiv. https://doi.org/10.48550/arXiv.2412.15115