Language Models at the Syntax-Semantics Interface: A Case Study of the Long-Distance Binding of Chinese Reflexive *ziji*

Xiulin Yang
Georgetown University
xy236@georgetown.edu

Abstract

This paper explores whether language models can effectively resolve the complex binding patterns of the Mandarin Chinese reflexive *ziji*, which are constrained by both syntactic and semantic factors. We construct a dataset of 240 synthetic sentences using templates and examples from syntactic literature, along with 320 natural sentences from the BCC corpus. Evaluating 21 language models against this dataset and comparing their performance to judgments from native Mandarin speakers, we find that none of the models consistently replicates human-like judgments. The results indicate that existing language models tend to rely heavily on sequential cues, though not always favoring the closest strings, and often overlooking subtle semantic and syntactic constraints. They tend to be more sensitive to noun-related than verb-related semantics.[1]

1 Introduction

Binding is a specific type of co-indexation that "lies at the very heart and soul of human language" (Abbott, 2010). In a sentence, if a noun phrase NP_A binds another noun phrase NP_B , it indicates that both refer to the same entity (Carnie, 2021). In such cases, a pronoun or reflexive that refers back to an NP is called an *anaphor*, and the NP it refers to is termed the *antecedent*.

The impressive performance of Pretrained Language Models (PLMs) in various NLP tasks has raised an important question: **Do these models inherently acquire abstract linguistic knowledge solely from their training on sequences of strings?** To investigate this, many researchers treat language models as psycholinguistic objects. By designing minimal pairs and analyzing the probabilistic outputs from these models, they assess the models' preferences in linguistic judgments. Numerous studies have explored linguistic patterns across different levels using minimal pairs, such as syntax (e.g., Wilcox et al., 2018; Linzen and Baroni, 2021; de Dios-Flores et al., 2023), semantics (e.g., Zhang et al., 2023), and pragmatics (e.g., Davis, 2022). A few have examined the binding phenomenon in English, such as Reflexive Anaphor Licensing (Hu et al., 2020; Warstadt et al., 2020; Lee and Schuster, 2022; Marvin and Linzen, 2018) and the constraints of Principle B in Binding Theory (Davis, 2022). There has also been work on binding in Chinese (Xiang et al., 2021; Song et al., 2022), but these studies typically focus on simple cases like gender/number agreement in local binding of complex reflexive *ta-ziji* (*himself/herself*).

1. 杰克$_i$ 在 玛丽$_j$ 家 写 自己$_{i/*j}$ 的 作业. 。
   Jack be Mary home write self POSS homework .
   Jack is writing his homework at Mary's home.

2. 杰克$_i$ 知道 玛丽$_j$ 在 炫耀 自己$_{*i/j}$ 。
   Jack know Mary be show-off self .
   Jack knows that Mary is showing off herself.

3. 杰克$_i$ 知道 玛丽$_j$ 在 等 自己$_{i/*j}$ 。
   Jack know Mary be wait self .
   Jack knows that Mary is waiting for him.

Figure 1: Examples of binding, the words highlighted in the same color are co-indexed.

However, binding in Chinese reflexives, particularly with the bare form *ziji*, involves more than just gender agreement – it is governed by intricate syntactic, semantic, and pragmatic constraints (see accounts from Pan (1998); Pan and Hu (2003); Lam (2021)). Chomsky's Binding Theory (Chomsky, 1993), especially Principle A, explains English reflexives but fails to generalize to long-distance reflexives like *ziji* in Mandarin and others[2] (see 2.1). As shown in Figure 1, different syntactic structures

[1] Code and data are accessible via https://github.com/xiulinyang/zh-reflexive

[2] e.g., *zibun* in Japanese and *kaki* in Teochew (Cole et al., 2006).

Proceedings of the 31st International Conference on Computational Linguistics, pages 3808–3824
January 19–24, 2025.

and verb types can lead to varied readings.

This study aims to investigate how language models handle the nuanced syntactic and semantic constraints in the complex binding patterns of the reflexive *ziji* in Mandarin Chinese. Specifically, we seek to answer the following research questions:

- Can language models accurately process the intricate binding patterns of *ziji* as humans do?
- What factors contribute to the alignment or discrepancy between human judgments and the predictions made by these models?

We examine various language models, including monolingual, multilingual, masked language models, and autoregressive models, using both synthetic and natural datasets.

Our findings indicate that most models can predict some linguistic constraints but primarily rely on linear preferences rather than fully grasping syntax or semantics. Interestingly, not all models prefer binders closer to *ziji*; for instance, `bert-base-chinese` favors long-distance binders. Furthermore, our experiments show that the models are better at noun-related semantics than verb-related ones.

Our key contributions are as follows: (1) We introduce a unique dataset comprising both synthetic and natural examples, along with human evaluation. To our knowledge, this is the first study to publicly provide human judgments and data in exploring the long-distance binding of *ziji*; (2) we conduct one of the first comprehensive evaluations of multiple language models on the syntax-semantics interface in Chinese, revealing their limitations and investigating the underlying factors contributing to these shortcomings.

2 Background & Related Work

2.1 Chinese Reflexive *ziji*

Chinese reflexive *ziji* has been extensively studied for decades due to its exceptional behaviors violating Principle A in the classic Binding Theory (Chomsky, 1993). According to Principle A, an anaphor must be bound within its governing category, typically the clause in which it appears. For example, in the sentence shown in Example (1), the reflexive *ziji* can only refer to the subject *Mary*, following a pattern known as **local binding**.

(1) 玛丽$_j$ 相信 自己$_j$ 。
Mary$_j$ xiangxin ziji$_j$
Mary trust self

Mary trusted herself.

However, *ziji* exhibits more complex behaviors beyond local binding. Its interpretation is governed by a range of syntactic and semantic constraints that have been extensively documented in the literature (e.g., Tang, 1989; Huang and Tang, 1991; Pan, 2000; Charnavel and Huang, 2018; Lam, 2021; Charnavel and Huang, 2018). These complexities make *ziji* a particularly intriguing case for studying reflexive binding patterns. In this research, we focus on several of these.

Long-distance Binding The antecedent can be bound remotely by the matrix subject in a complex clause (Liejiong, 1993; Huang and Tang, 1991; Tang, 1989). For example, in (2), *ziji* can refer to either the antecedent within the subordinate clause (*Mary*) or beyond the clause (*Jack*).

(2) 杰克$_i$ 知道 玛丽$_j$ 相信 自己$_{i/j}$。
Jack$_i$ zhidao Mary$_j$ xiangxin ziji$_{i/j}$
Jack knew Mary trust self

Jack knew that Mary trusted herself/him.

Blocking Effect In a complex clause, when the first/second person pronoun is inserted between the long-distance binder and reflexive, the long-distance binding is blocked or not allowed (Pan, 2000). Different from (2), in (3), *ziji* instead can only refer to the first person *I* but not *Jack*.

(3) 杰克$_i$知道我$_j$相信自己$_{*i/j}$。
Jack$_i$ zhidao wo$_j$ xiangxin ziji$_{*i/j}$
Jack knew I trust self

Jack knew that I trusted myself.

Animacy Effect The antecedent of *ziji* must be animate (Tang, 1989). In sentence (4), although syntax allows both local binding and long-distance binding, the inanimacy of the subordinate subject makes the local binding impossible.[3]

(4) 杰克$_i$说这本书$_j$欺骗了自己$_{i/*j}$。
Jack$_i$ shuo zhe ben shu$_j$ qipian le ziji$_{i/*j}$
Jack say this CLS book deceive ASP self

Jack said that this book deceived him.

[3]Recent studies (e.g., Charnavel and Huang, 2018; Lam, 2021) have challenged the Animacy Effect assumption, but the counter-examples they provide are only limited to specific constructions which do not overlap with those used in our experiments.

Subject Orientation Only the subject or part of the subject (e.g., possessor) can be a possible antecedent of *ziji* (Tang, 1989; Lam, 2021). For example, in (5), only 杰克 *Jack* can be the antecedent of *ziji* because it is the only subject in the clause. The other pronoun 她 (*her*) is the indirect object of the verb 告诉 (*tell*).

(5) 杰克$_i$告诉她$_j$自己$_{i/*j}$的成绩。
Jack$_i$ gaosu ta$_j$ ziji$_{i/*j}$ de chengji
Jack tell her self DE grade

Jack told her his grade.

Verb Orientation Studies have also found that in complex sentences, the meaning of subordinate predicates might disambiguate the possible readings of *ziji* (Qiu, 2015; Schumacher et al., 2011). For instance, the following two examples share the same syntactic structures but have different readings because of the semantics of the subordinate predicate. For (6a), the flatterer typically flatters someone else rather than themselves. By contrast, in (6b), one can only reflect on their own mind, not others.

(6) a. 杰克$_i$说玛丽$_j$巴结了自己$_{i/*j}$。
Jack$_i$ shuo Mary$_j$ bajie le ziji$_{i/*j}$
Jack say Mary flatter ASP self

Jack said that Mary flattered him.

b. 杰克$_i$说玛丽$_j$反省了自己$_{*i/j}$。
Jack$_i$ shuo Mary$_j$ fanxing le ziji$_{*i/j}$
Jack say Mary reflect.on ASP self

Jack said that Mary reflected on herself.

2.2 Probing Linguistic Knowledge in Language Models

To examine linguistic knowledge in language models, many studies have explored syntactic and semantic structures such as subject-verb agreement (Marvin and Linzen, 2018), Negative Polarity Items (Jumelet et al., 2021), and long-distance dependencies (Marvin and Linzen, 2018) across languages (Xiang et al., 2021; de Dios-Flores et al., 2023). Findings show that while models produce syntactically correct output, their performance may be biased by dependency distance or token frequency (Newman et al., 2021; Wei et al., 2021).

Probing methods in NLP serve as crucial tools for deciphering the intricate workings of language models. These methods range from probing tasks, which assess the model's grasp on linguistic properties through external classifiers (e.g., Wu and Dredze, 2019; Levy and Goldberg, 2014; Tenney et al., 2019; Kulmizev et al., 2020), to attention analysis (Voita et al., 2019), aimed at understanding focus patterns within the network. Further, many studies employ concepts from Information Theory such as perplexity and surprisal to investigate the linguistic behaviors of language models by taking them as psycholinguistic objects (Futrell et al., 2019; Wilcox et al., 2023; Oh et al., 2022). Recently, as more LLMs shift towards closed-source, researchers have turned to prompting techniques to explore their knowledge (Katzir, 2023; Ambridge and Blything, 2024; Lamprinidis, 2023; Dentella et al., 2023). In our study, we employ the perplexity-based method for open-source models and prompting for closed-source models.

Most research on language models has focused on languages whose case and agreement systems facilitate controlled experiments (de Dios-Flores et al., 2023). In contrast, less focus has been given to the syntax-semantics interface in morphologically poor languages like Mandarin Chinese. de Dios-Flores et al. (2023) found that language models often mispredict anaphora resolution in Spanish and Galician when antecedents and anaphors are distantly placed. Similar trends have been observed in English studies (Lee and Schuster, 2022). While some research has examined Chinese linguistic knowledge in models like BERT, it has mainly addressed syntax (Zheng and Liu, 2023; Kulmizev et al., 2020; Xiang et al., 2021), leaving the syntax-semantics interface largely unexplored. Our study aims to fill this gap by investigating the binding of *ziji* in Mandarin Chinese.

3 Data

To address potential discrepancies between synthetic and natural data used for training language models, we develop two distinct datasets: one generated automatically via a script or hand-crafted by linguists, and the other collected from the BCC corpus[4] (Xun et al., 2016).

3.1 Synthetic Data

To create synthetic data, we choose the syntactic structure consistent with most psycholinguistic studies on long-distance binding of *ziji* (e.g., Schumacher et al., 2011; Li and Kaiser, 2009), i.e., $NP_1 + V_1 + NP_2 + V_2 + ziji$. This structure is used to test all constraints except for Subject Orien-

[4] https://bcc.blcu.edu.cn/

Binding Pattern	Constraint	Categories	Example	Gold Binding
Ambiguous Long distance binding (AMB LD)	Syntax&Semantics	Fem pronoun first	她$_f$知道他$_m$相信自己$_{f/m}$。 *She$_f$ knows that he$_m$ trusts himself$_m$/her$_f$.*	Ambiguous
		Masc pronoun first	他$_m$知道她$_f$相信自己$_{m/f}$。 *He$_m$ knows that she$_f$ trusts herself$_f$/him$_m$.*	Ambiguous
Verb Orientation (VO)	Semantics	Reflexive	她$_f$知道他$_m$在检讨自己$_m$。 *She$_f$ knows that he$_m$ is reflecting on himself$_m$.*	Local
		No-reflexive	他$_m$知道她$_f$在躲避自己$_m$。 *He$_m$ knows that she$_f$ is escaping him$_m$.*	Remote
Subject Orientation (SO)	Syntax	Fem pronoun first	她$_f$给他$_m$关于自己$_f$的书。 *She$_f$ gave him$_m$ her$_f$ own book.*	Remote
		Masc pronoun first	他$_m$给她$_f$关于自己$_f$的书。 *He$_m$ gave her$_f$ his$_m$ own book.*	Remote
Blocking Effect (BE)	Syntax&Semantics	NA	她$_f$知道我$_m$相信自己$_m$。 *She$_f$ knows that I$_w$ trust myself$_w$.*	Local
Animacy Effect (AE)	Semantics	NA	她$_f$知道这封信$_t$暴露了自己$_f$。 *She$_f$ knows that the letter$_t$ exposes her$_f$.*	Remote

Table 1: Binding patterns and their corresponding examples. Among the subscript following NPs, *m* refers to third person masculine pronoun, *f* refers to third person feminine pronoun, *w* refers to the first-person pronoun, *t* refers to the third person inanimate pronoun. *Remote* means the antecedent is linearly farther away than the incorrect binder or distractor, not strictly *long-distance* binding explained in 2.1.

tation. Given the focus of our experiments, we specifically varied NP_1, NP_2, V_1, and V_2. Due to the absence of morphological inflections in Chinese to mark agreement between antecedents and anaphors, we limited our selection of NP_1s to four single-character pronouns with different semantic features: 他 *(he/him)*, 她 *(she/her)*, 我 *(I/me)*, and 它 *(it)*.

V_1 is always a statement/attitude verb. Regarding the choice of V_2, we leveraged the comprehensive analysis by Qiu (2015), who examined how the semantic properties of verbs influence the binding of *ziji*. After reviewing the Chinese Verb Usage Dictionary, Qiu (2015) categorized verbs into three main types: non-reflexive verbs, which inhibit local binding of *ziji* (e.g., sentence (6a)); reflexive verbs, which prevent long-distance binding of *ziji* (e.g., sentence (6b)); and bidirectional verbs, which allow ambiguous interpretations of *ziji* (e.g., sentence (2)). We select a random subset of verbs from the former two categories in our study to build sentence pairs for Verb Orientation tests.

Regarding the Subject Orientation constraint, we utilize the following two syntactic structures: $NP_1 + V_1 + NP_2 + ziji + de + NP_3$, where V_1 is a ditransitive verb, NP_2 is the indirect object, and *ziji* serves as the possessor of the entire direct object phrase (i.e., *ziji de NP_3 (one's own NP_3)*); and $NP_1 + PP + V_1 + ziji + de + NP_3$, where a distractor noun is inserted into the PP. In both cases, *ziji* can only be bound by NP_1.

As for the Blocking Effect, psycholinguistic studies have noted that this constraint is not absolute (Lyu and Kaiser, 2021). To minimize potential biases arising from our templates or verb selection, we supplemented our dataset with 40 sentences from existing literature (Li, 2023; Shuai et al., 2013; Pan, 2000; Schumacher et al., 2011; Huang, 2002; Chen, 2009; Liu, 2010; Yang and Wu, 2015; Cole and Sung, 1994).

Additionally, by replacing the first-person pronoun in sentences from the Blocking Effect category, we design 40 sentence pairs with the third-person pronoun to allow ambiguous binding and test language models' structural bias as a baseline. We aim to assess which binding – local or long-distance – language models/humans prefer when both are acceptable.

Overall, our experimental dataset comprises 240 sentences, with each category containing 40 examples. The linguistic patterns, example sentences, and correct binding are detailed in Table 1.

3.2 In-Context Minimal Pairs

We design the synthetic dataset to test the reading of *ziji*. However, we cannot get who *ziji* refers to simply from the sequence because *ziji* itself does not have any morphological cue to indicate its binder. To address this, we develop a method we call *in-context minimal pairs*. We embed the target sentence in a structure like: *If [TARGET SENTENCE], then [INTERPRETATION OF TARGET SENTENCE]*. In the second clause, the semantic feature of the binder is made explicit by using pronouns or complex reflexive (e.g., ta-ziji *himself*). This approach allows us to test language models' preferred reading in a more natural context. For instance, sentence (2) can be reformulated into the following minimal pair. (7a) suggests a local binding of *ziji*, while (7b) suggests a long-distance binding. The minimal pair examples for different binding patterns are detailed in Appendix A.

(7) a. 如果 杰克$_i$ 知道 玛丽$_j$ 相信 自己$_{i/j}$， 那么 玛丽 相信 她自己。
if Jack$_i$ zhidao Mary$_j$ xiangxin ziji$_{i/j}$, namo
if Jack knew Mary trust self, then
Mary xiangxin taziji
Mary trust herself.

If Jack knew that Mary trusted herself/him, then Mary trusted herself.

b. 如果 杰克$_i$ 知道 玛丽$_j$ 相信 自己$_{i/j}$， 那么 玛丽 相信 他。
If Jack$_i$ zhidao Mary$_j$ xiangxin ziji$_{i/j}$, namo
if Jack knew Mary trust self, then
Mary xiangxin ta.
Mary trust him

Jack knew that Mary trusted herself/him, then Mary trusted him.

3.3 Natural Data

Since previous research has indicated that language models significantly underperform humans on reflexive binding tasks (Song et al., 2022), we aim to extend this evaluation to natural data to determine if similar conclusions hold. We manually select 240 natural sentences from the BCC corpus, ensuring they align with the structure of the synthetic data. Additionally, we collect 80 sentences specifically involving local binding in contrast with Song et al. (2022)'s data.

Model	# Params	Training Data Size
bert-base-chinese (Devlin et al., 2019)	110M	300G
chinese-lert-base (Cui et al., 2022a)	102M	20GB
chinese-lert-large (Cui et al., 2022a)	325M	20GB
chinese-pert-base (Cui et al., 2022b)	102M	20GB
chinese-pert-large (Cui et al., 2022b)	325M	20GB
mengzi-bert-base (Zhang et al., 2021b)	103M	300G
mengzi-bert-base-fin (Zhang et al., 2021b)	103M	320G
ernie-1.0-base-zh (Sun et al., 2019)	110M	173M sent.
mBERT (Devlin et al., 2019)	110M	-
XLM-R-base (Conneau et al., 2019)	125M	2.5TB
XLM-R-large (Conneau et al., 2019)	355M	2.5TB
mt5-small(Xue et al., 2020)	300M	0.5TB
mt5-large (Xue et al., 2020)	1.2B	1TB
GPT2 (Zhao et al., 2019)	117M	14GB
GPT2-medium (Zhao et al., 2019)	345M	14GB
GPT2-large (Zhao et al., 2019)	762M	14GB
GPT2-xlarge (Zhao et al., 2023)	1.5B	14GB
GLM-4-9b-chat (GLM et al., 2024)	9B	-
CPM-Generate (Zhang et al., 2021a)	2.6B	100GB
GPT-3.5 (OpenAI, 2023)	NA	NA
GPT-4o (OpenAI et al., 2023)	NA	NA

Table 2: Overview of the models used in the experiments, categorized by architecture: encoder-only models, encoder-decoder models, and decoder-only models. The table also includes the corresponding number of parameters and training data sizes for each model. Multilingual models are highlighted in blue for clarity.

For local binding constructions, we select sentences where 她自己 (*herself*) or 他自己 (*himself*) appears as the direct object. For other binding constructions except for Subject Orientation, we focus on sentences following the $NP_1 + V_1 + NP_2 + V_2 + ziji$ pattern, allowing for additional contextual elements or modifiers. For Subject Orientation, we select sentences that contain a distractor NP with a different gender feature between the antecedent and *ziji*. To minimize potential confounds brought by gender bias in our experiments, we make minimal alterations to ensure gender balance in the dataset.

Embedding natural sentences into an *if ... then* template (or using other similar connectives) often makes them sound unnatural, as these sentences are longer than typical conditional clauses. This makes it difficult, if not impossible, to create natural-sounding in-context minimal pairs. We assume that using complex reflexives like 她/他/我自己 (*her/him/myself*) serves as a useful proxy for testing language models' preferred readings of *ziji*, as the pronoun makes the reference explicit. Thus, we use minimal pairs by replacing *ziji* with *ta-ziji* to clarify its meaning where contextually appropriate.[5] The gender and animacy features of the incorrect candidate are determined by the non-antecedent noun.

4 Evaluation

4.1 Models

Following Song et al. (2022), we evaluated most models used in their study and included additional language models trained on monolingual or multilingual corpora, featuring different architectures and sizes. In total, we examined 21 language models, including the latest `GPT-4o`. The number of parameters and the size of the training data can be found in Table 2.[6]

4.2 (Pseudo-) Perplexity

In line with Song et al. (2022), we evaluate the performance of autoregressive language models using perplexity (PPL) and masked language models using pseudo-perplexity (PPPL) (Salazar et al., 2020). The equations for PPL are defined as follows.

$$\begin{aligned} L &= \frac{1}{M}\sum_{i=1}^{m}\log p(w_i|w_1 \dots w_{i-1}) \\ \mathrm{PPL} &= \exp(-L) \end{aligned} \tag{1}$$

While PPL measures the probability of tokens based solely on preceding context, PPPL calculates the probability of a token using the entire bidirectional context, informed by the pretrained tasks of MLMs.

$$\begin{aligned} w_{\setminus i} &= w_1 \dots w_{i-1}, w_{i+1} \dots w_m \\ \text{pseudo-}L &= \frac{1}{M}\sum_{i=1}^{m}\log p(w_i|w_{\setminus i}) \\ \mathrm{PPPL} &= \exp(-\text{pseudo-}L) \end{aligned} \tag{2}$$

These metrics are based on the average token-level log probability, allowing for a fair evaluation across sentences of varying lengths in our experiments. For example, consider the comparison between 他 (*he*) and 她自己 (*herself*) in example (7). Although both correspond to one word in English, the former has fewer characters. Averaging sentence length helps mitigate the tendency of language models to favor shorter sentences (Song et al., 2022). Additionally, using perplexity as a common standard enables a more effective comparison of different language models' performance. We take the sentence in a sentence pair that has a lower perplexity as the models' preference.

4.3 Evaluation of closed-source LLMs

For closed-source LLMs, i.e., `GPT-3.5-turbo` and `GPT-4o`, we use prompts to ask the model to select the more natural and acceptable sentence. The prompts can be found in Appendix C.

4.4 Human Evaluation

To compare the performance of language models with humans, we recruited 24 native Mandarin speakers as volunteers to complete a cloze-filling task. To minimize bias toward any specific sentence structure, each participant annotated 5 sentences from each category, with sanity check sentences, totaling 70 sentences per person. Each sentence was annotated by three different participants, and the most frequently chosen response was adopted as the final annotation.

To assess the reliability of the annotations, we calculated Fleiss' Kappa for every group of three annotators. The average Fleiss' Kappa score reached 0.81,[7] which indicates an "almost perfect" inter-annotator agreement (Landis and Koch, 1977).

5 Result & Discussion

5.1 Overall Result

The results are summarized in Tables 3 and 4, which present several noteworthy findings that we will discuss in detail in the following subsections. It is important to note that for the two closed-source LLMs, altering the order of the sentences within minimal pairs in the prompt significantly affected the results (see Appendix **??**). Therefore, we report the results with the sentences randomly shuffled in the minimal pairs. We also discuss the limitation of the prompt-based method in the Limitation section. Before diving into the detailed analysis, we would like to highlight a few key observations.

First, none of the models can match human performance in both settings. In the synthetic data setting, `mengzi-bert-base` shows the best performance among all models, and in the natural setting,

[5] See Appendix B for examples.

[6] Note that discrepancies from (Song et al., 2022) may arise from references to different sources about the model information.

[7] Most of the disagreement comes from the ambiguous setting where three native speakers might have different preferences.

	Human	bert-base-chinese	chinese-lert-base	chinese-lert-large	chinese-pert-base	chinese-pert-large	mengzi-bert-base	mengzi-bert-base-fin	ernie-1.0-base-zh	multilingual-bert	xlmr-base	xlmr-large	mt5-small	mt5-large	gpt2-distil	gpt2-medium	gpt2-chinese	gpt2-xlarge	glm-4-9b-chat	CPM-Generate	GPT-3.5	GPT-4o
Blocking	92.5	22.5	10.0	17.5	15.0	5.0	42.5	45.0	15.0	0.0	7.5	17.5	17.5	30.0	100.0	67.5	100.0	90.0	75.0	100.0	62.5	27.5
Animacy	100.0	100.0	100.0	100.0	100.0	75.0	97.5	97.5	100.0	100.0	100.0	100.0	25.0	27.5	100.0	100.0	92.5	100.0	22.5	100.0	100.0	100.0
Verb$_{refl}$	100.0	5.0	7.5	2.5	0.0	37.5	70.0	45.0	2.5	52.5	2.5	20.0	10.0	20.0	97.5	95.0	100.0	100.0	52.5	5.0	55.0	65.0
Verb$_{nonrefl}$	97.5	100.0	100.0	100.0	100.0	75.0	62.5	65.0	100.0	52.5	100.0	95.0	92.5	100.0	7.5	0.0	0.0	0.0	100.0	72.5	100.0	97.5
SO	87.5	37.5	80.0	87.5	65.0	52.5	85.0	50.0	65.0	7.5	70.0	45.0	0.0	22.5	50.0	47.5	52.5	60.0	35.0	47.5	50.0	50.0
Average	95.5	53.0	59.5	61.5	56.0	49.0	71.5	60.5	56.5	42.5	56.0	55.5	29.0	40.0	71.0	62.0	69.0	70.0	57.0	65.0	65.0	68.5
Ambiguous	17.5	15.0	15.0	12.5	5.0	22.5	12.5	5.0	17.5	20.0	10.0	20.0	12.5	87.5	82.5	95.0	87.5	60.0	35.0	50.0	22.5	35.0

Table 3: Accuracy Scores of Predictions on Synthetic Data and **Local Binding Percentage** on the Ambiguous setting (last row). Cells are shaded to reflect performance levels, with darker shades indicating higher accuracy. **Blocking** refers to the blocking effect setting; **Animacy** refers to the animacy effect experiment. **Verb**$_{refl}$ refers to the reflexive subcategory within the Verb Orientation category, while **Verb**$_{nonrefl}$ denotes the non-reflexive category. **SO** indicates Subject Orientation. The two gray-shaded GPT models are highlighted because they are evaluated using prompting rather than perplexity.

	Human	bert-base-chinese	chinese-lert-base	chinese-lert-large	chinese-pert-base	chinese-pert-large	mengzi-bert-base	mengzi-bert-base-fin	ernie-1.0-base-zh	multilingual-bert	xlmr-base	xlmr-large	mt5-small	mt5-large	gpt2-distil	gpt2-medium	gpt2-chinese	gpt2-xlarge	glm-4-9b-chat	CPM-Generate	GPT-3.5	GPT-4o
Blocking	100	50.0	72.5	72.5	55.0	52.5	72.5	72.5	80.0	55.0	45.0	65.0	50.0	90.0	62.5	47.5	57.5	55.0	90.0	72.5	57.5	100.0
Animacy	97.5	97.5	97.5	97.5	100.0	80.0	97.5	100.0	97.5	90.0	82.5	87.5	15.0	97.5	97.5	92.5	97.5	97.5	75.0	100.0	80.0	97.5
Verb$_{refl}$	100	72.5	62.5	80.0	72.5	50.0	65.0	77.5	65.0	70.0	70.0	87.5	57.5	82.5	45.0	42.5	32.5	57.5	92.5	22.5	60.0	100.0
Verb$_{nonrefl}$	100	75.0	72.5	95.0	65.0	85.0	65.0	72.5	72.5	30.0	82.5	77.5	47.5	87.5	55.0	80.0	65.0	82.5	95.0	75.0	90.0	100.0
SO	97.5	72.5	80.0	90.0	57.5	57.5	87.5	80.0	72.5	60.0	70.0	80.0	27.5	62.5	55.0	72.5	65.0	82.5	85.0	47.5	60.0	97.5
Average	99.0	73.5	77.0	87.0	70.0	65.0	77.5	80.5	77.5	61.0	70.0	79.5	39.5	84.0	63.0	67.0	63.5	75.0	87.5	63.5	69.5	99
Local$_m$	100	85.0	87.5	90.0	87.5	10.0	90.0	87.5	87.5	82.5	62.5	90.0	17.5	77.5	85.0	87.5	87.5	97.5	95.0	72.5	62.5	92.5
Local$_f$	100	85.0	85.0	92.5	92.5	87.5	100.0	100.0	95.0	82.5	97.5	92.5	17.5	80.0	97.5	95.0	97.5	95.0	82.5	80.0	57.5	97.5

Table 4: Accuracy Scores of Predictions on Natural Data. Cells are shaded to indicate performance levels, with darker shades representing higher accuracy scores. The last two rows show the results of **local binding** on two gender settings in natural data.

`glm-4-9b-chat` outperforms other models. Multilingual models perform worse than monolingual models.

Second, larger model sizes do not necessarily lead to better performance. In the synthetic data setting, `gpt2-distill` outperforms `gpt2-xlarge` with the same training data. Similarly, `chinese-pert-base` and `XLM-R-base` surpass their larger counterparts in both synthetic and natural settings. The two largest models `GPT-4o` and `GPT-3.5` show limited performance in this task as well.

As shown in Table 5, the difficulty of various constraints is consistent across synthetic and natural data. However, all models perform better in the natural data setting, despite natural language often containing more distractors and longer sentences than synthetic data. This contrasts with the semantic parsing results noted by Yang and Schneider (2024). We hypothesize two possible reasons for this phenomenon: (1) natural data may better reflect the distribution of the training data, suggesting that the models struggle to generalize the underlying abstract rules, and (2) the pretrained data is contaminated with our evaluation set. Most of our examples come from literature, making both hypotheses plausible for models trained on literary works or CommonCrawl. However, `bert-base-chinese`, `ernie-base`, and `mbert` are trained on data from Wikipedia and non-literary domains, indicating that the first hypothesis might be more likely.[8] Additionally, we hypothesize that the second explanation applies to `GPT-4o`, given the significant difference in its performance between the synthetic and natural data.

Binding	Syn Data	Natural Data
Blocking	41.3	65.5
Animacy	87.5	89.4
Verb$_{refl}$	40.2	65.0
Verb$_{nonrefl}$	70.5	74.8
SO	50.5	69.6

Table 5: Average accuracy of different binding phenomena across all evaluated models.

5.2 Language models show linear biases but not all language models prefer local binders

Both Song et al. (2022) and Xiang et al. (2021) observe the models' vulnerability to linearly close distractors. Similar findings have been confirmed

[8]However, a recent study (Misra and Mahowald, 2024) shows that language models can generalize rare phenomena from less rare ones. Validating this hypothesis requires rigorous experimental design, which we leave for future work.

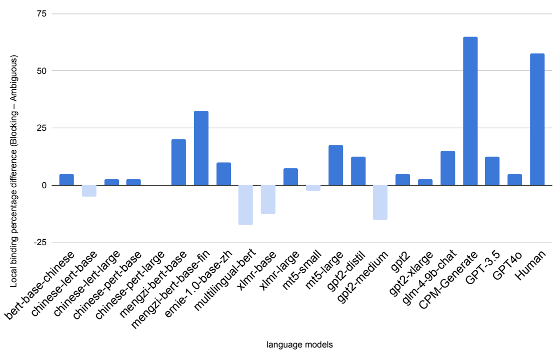

Figure 2: Local binding tendency caused by the blocking effect based on the baseline result.

in other languages, such as English with `GPT-2` (Lee and Schuster, 2022) and Spanish with `mbert` (de Dios-Flores et al., 2023). Our experiments show two key findings: (1) almost all languages have linear biases, yet (2) not all models show a bias toward local binders. In particular, most of the **encoder-only** models prefer **long-distance** binders while **decoder-only** models prefer **local binding** as shown in Table 3.

Regarding our first observation, we find that most models predict the blocking effect not because of the insertion of a first-person pronoun, but due to their linear preference. Since the examples in the Ambiguous Binding category are adapted from the Blocking Effect category by replacing first-person pronouns with third-person pronouns, this setup allows us to compare the results of these two groups and assess the influence of first-person pronouns on model behavior. Specifically, if language models have truly learned the underlying constraints of the blocking effect, they should assign higher probabilities to local binding readings when third-person pronouns are replaced with first-person pronouns. Consequently, we expect stronger local binding preferences in the blocking effect experiment than in the ambiguous binding experiment.

To quantify this, we define the local binding tendency as the difference between the number of local binding cases in the blocking effect and the number of local binding cases in the ambiguous binding category. Our findings reveal that most models – except for `chinese-lert-base`, `xlmr-base`, `chinese-pert-base`, `gpt2-medium`, and `glm4` – exhibit a slightly stronger tendency toward local binding in cases involving first-person pronouns, as illustrated in Figure 2. However, this tendency is generally weak across most models. The notable exception is `CPM-Generate`, which demonstrates a significant increase in its preference for local binding when a first-person pronoun is present, effectively mimicking human behavior in similar contexts. In conclusion, with the exception of `CPM-Generate`, all models appear to make correct predictions in the blocking effect setting primarily due to their linear bias, rather than an understanding of the constraints underlying the blocking effect pattern.

As for the second observation, we find that in the Blocking Effect and Verb Orientation (*reflexive verbs*) settings, where *ziji* should be bound to its local antecedent, most encoder-only models perform poorly. However, their performance improves in the Verb Orientation (*non-reflexive verbs*) and Subject Orientation settings where long-distance binding is expected. In contrast, decoder-only models exhibit near-perfect performance in local-binding settings, such as the Blocking Effect, Animacy Effect, and Verb Orientation with Reflexive Verbs.

5.3 Language Models Are More Sensitive to Semantics of Nouns than Verbs

The animacy effect and two verb orientation experiments investigate whether language models possess the semantic knowledge required to resolve binding. As shown in the table, most models, except for `mt5-small`, perform well in the animacy setting, indicating they encode the knowledge that *ziji* can only refer to an animate NP. This is particularly evident among the decoder-only models, which, despite exhibiting a strong bias toward local binding, can successfully switch to long-distance binding when the local binder is an inanimate noun, achieving nearly perfect accuracy.

This raises another question: is the success of the encoder-only models in the animacy setting due to their knowledge of animacy or their linear bias? To address this, we switch the order of the animate matrix subject and the inanimate subordinate subject, where local binding is the correct interpretation. As shown in Figure 3, models that perform well in the typical animacy setting also excel in the switched experiment. This supports the first hypothesis: encoder-only models do learn animacy knowledge about *ziji*.

In contrast, none of the models perform equally well in the two Verb Orientation experiments which require different binding readings. We observe that models favoring local binding tend to perform poorly in non-reflexive verb scenarios, where

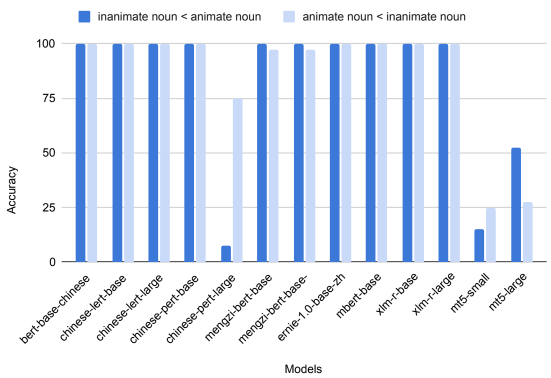

Figure 3: Accuracy of language models across two settings of the animacy effect: (1) matrix subject is animate and subordinate subject is inanimate (animate < inanimate), and (2) matrix subject is inanimate and subordinate subject is animate (inanimate < animate).

the meanings of subordinate predicates necessitate lang-distance binding. Similarly, models preferring long-distance binding excel in non-reflexive verb settings but struggle with reflexive verbs. This pattern is particularly evident in synthetic data.

Therefore, we argue that while most language models possess semantic knowledge of (animate) nouns, they struggle to understand the nuances of verb meanings. We assume that this difficulty may arise from the fact that the animacy of nouns is generally easier to distinguish than the reflexiveness of verbs. It is possible that there are more readily available distributional cues of animacy versus reflexiveness of verbs if we accept the assumption that all nouns can be either animate or inanimate while reflexive and non-reflexive verbs are less frequent than ambiguous verbs.

6 Conclusion

In this paper, we evaluated 21 language models across two data settings. Our results reveal that none of the models consistently replicate human-like judgments. We observe that all language models rely heavily on sequential biases, even when tasked with modeling syntactic and semantic cues. Furthermore, most models demonstrate a better understanding of the semantics of nouns compared to verbs. Several intriguing questions remain open. For instance, why do models find it easier to handle natural data, despite its longer sequences and more distractors, than synthetic data? Can language models generalize complex constraints based on more frequent and simpler linguistic phenomena? Why does SO not show a clear linear bias among LMs? We leave these questions for future research and welcome new insights into these areas.

Limitations

We are aware that the prompt-based method for `GPT-3.5` and `GPT-4o` is **not directly comparable** to the perplexity-based approach, as results obtained using meta-linguistic prompts tend to perform worse than those derived from model representations (Hu and Levy, 2023). As we mentioned in Appendix **??**, both models are highly sensitive to the order of the sentence pairs in the prompt, with `GPT-3.5` showing a stronger bias toward Option-A. This observation remains consistent across different prompt designs. Therefore, the results we report may not fully reflect the language capability of these two LMs and our conclusion might not apply to them. We advise readers to interpret the results from these models with caution.

Ethical Considerations

Our project had minimal computational costs since no additional model training was required. For human participants, informed consent was obtained prior to their participation in the questionnaire, and all collected data was anonymized and kept confidential to protect their privacy. Additionally, when creating and collecting sentences for the study, we ensured that the content was free from harmful or offensive material.

Acknowledgements

We would like to express our gratitude to Amir Zeldes, Nathan Schneider, Tatsuya Aoyama, Wesley Scivetti, Yixiao Song, and all members of the NERT lab for their insightful suggestions and support. We are also deeply thankful to the volunteers who participated by completing the questionnaires; their contributions were essential to the success of this study. Finally, we extend our appreciation to the anonymous reviewers for their thoughtful and constructive feedback.

References

Barbara Abbott. 2010. *Reference*, volume 2. Oxford University Press.

Ben Ambridge and Liam Blything. 2024. Large language models are better than theoretical linguists at theoretical linguistics. *Theoretical Linguistics*, 50:33 – 48.

Andrew Carnie. 2021. *Syntax: A generative introduction*. John Wiley & Sons.

Isabelle Charnavel and Yujing Huang. 2018. Inanimate ziji and condition a in mandarin. In *Proceedings of the 35th West Coast conference on formal linguistics*, pages 132–141.

Hsiang-Yun Chen. 2009. Logophoricity and ziji. In *Proceedings of the 21st North American Conference on Chinese Linguistics (NACCL-21)*, volume 2, pages 464–481. Bryant University Smithfield, RI.

Noam Chomsky. 1993. *Lectures on government and binding: The Pisa lectures*. 9. Walter de Gruyter.

Peter Cole, Gabriella Hermon, and C-T James Huang. 2006. Long-distance binding in asian languages. *The Blackwell companion to syntax*, pages 21–84.

Peter Cole and Li-May Sung. 1994. Head movement and long-distance reflexives. *Linguistic Inquiry*, pages 355–406.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. In *Annual Meeting of the Association for Computational Linguistics*.

Yiming Cui, Wanxiang Che, Shijin Wang, and Ting Liu. 2022a. Lert: A linguistically-motivated pre-trained language model. *arXiv preprint arXiv:2211.05344*.

Yiming Cui, Ziqing Yang, and Ting Liu. 2022b. Pert: Pre-training bert with permuted language model.

Forrest Davis. 2022. Incremental processing of principle B: Mismatches between neural models and humans. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 144–156, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Iria de Dios-Flores, Juan Garcia Amboage, and Marcos Garcia. 2023. Dependency resolution at the syntax-semantics interface: psycholinguistic and computational insights on control dependencies. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 203–222, Toronto, Canada. Association for Computational Linguistics.

Vittoria Dentella, Fritz Guenther, Elliot Murphy, Gary Marcus, and Evelina Leivada. 2023. Testing ai on language comprehension tasks reveals insensitivity to underlying meaning.

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

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

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

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

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

C-T James Huang. 2002. Distributivity and reflexivity. *On the formal way to Chinese languages*, pages 21–44.

C-T James Huang and C-C Jane Tang. 1991. 13 the local nature of the long-distance reflexive in chinese.

Jaap Jumelet, Milica Denic, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language models use monotonicity to assess NPI licensing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4958–4969, Online. Association for Computational Linguistics.

Roni Katzir. 2023. Why large language models are poor theories of human linguistic cognition: A reply to piantadosi. *Biolinguistics*, 17:1–12.

Artur Kulmizev, Vinit Ravishankar, Mostafa Abdou, and Joakim Nivre. 2020. Do neural language models show preferences for syntactic formalisms? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4077–4091, Online. Association for Computational Linguistics.

Chit Fung Lam. 2021. A constraint-based approach to anaphoric and logophoric binding in mandarin chinese and cantonese. In *Proceedings of the LFG'21 Conference*, pages 202–222. CSLI Publications, Stanford University.

Sotiris Lamprinidis. 2023. Llm cognitive judgements differ from human. *ArXiv*, abs/2307.11787.

J Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74.

Soo-Hwan Lee and Sebastian Schuster. 2022. Can language models capture syntactic associations without surface cues? a case study of reflexive anaphor licensing in English control constructions. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 206–211, online. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics.

David Cheng-Huan Li and Elsi Kaiser. 2009. Overcoming structural preference: Effects of context on the interpretation of the chinese reflexive ziji. In *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009)*, pages 64–72.

Nini Li. 2023. What makes the blocking effect happen? *Asian Languages and Linguistics*, 4(1):76–100.

Xu Liejiong. 1993. The long-distance binding of ziji. *Journal of Chinese Linguistics*, 21(1):123–142.

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

Lijin Liu. 2010. A pragmatic account of anaphora: The cases of the bare reflexive in chinese. *Journal of Language Teaching & Research*, 1(6).

Jun-Hyun Lyu and Elsi Kaiser. 2021. Unpacking the blocking effect: Syntactic prominence and perspective-taking in antecedent retrieval in mandarin chinese. *Glossa: a journal of general linguistics*.

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.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing aanns. *arXiv preprint arXiv:2403.19827*.

Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. 2021. Refining targeted syntactic evaluation of language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3710–3723, Online. Association for Computational Linguistics.

Byung-Doh Oh, Christian Clark, and William Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5:777963.

OpenAI. 2023. Gpt-3.5. https://platform.openai.com/docs/models/gpt-3-5.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor

Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Haihua Pan. 1998. Closeness, prominence, and binding theory. *Natural Language & Linguistic Theory*, pages 771–815.

Haihua Pan. 2000. Why the blocking effect? In *Long distance reflexives*, pages 279–316. Brill.

Haihua Pan and Jianhua Hu. 2003. Prominence and locality in the binding of mandarin complex reflexive "ta-ziji"(s/he-self). *Journal of Chinese Linguistics Monograph*, 19:152–70.

Mingbo Qiu. 2015. *Constraint of Orientation of Verb on Intra-sentential Anaphora of Third Peronsal Pronoun and Reflexives in Chinese*. Fudan University Press.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Petra B Schumacher, Walter Bisang, and Linlin Sun. 2011. Perspective in the processing of the chinese reflexive ziji: Erp evidence. In *Anaphora Processing and Applications: 8th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2011, Faro, Portugal, October 6-7, 2011. Revised Selected Papers 8*, pages 119–131. Springer.

Lan Shuai, Tao Gong, and Yicheng Wu. 2013. Who is who? interpretation of multiple occurrences of the chinese reflexive: evidence from real-time sentence processing. *PloS one*, 8(9):e73226.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. SLING: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Chih-Chen Jane Tang. 1989. Chinese reflexives. *Natural Language & Linguistic Theory*, 7(1):93–121.

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.

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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, 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.

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. Frequency effects on syntactic rule learning in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of* the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

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

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

Liang Xu, Xuanwei Zhang, and Qianqian Dong. 2020. Cluecorpus2020: A large-scale chinese corpus for pre-training language model. *Preprint*, arXiv:2003.01355.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. In *North American Chapter of the Association for Computational Linguistics*.

Endong Xun, Gaoqi Rao, Xiaoyue Xiao, and Jiaojiao Zang. 2016. The construction of the bcc corpus in the age of big data. *Corpus Linguistics*, 3(1):93–109.

Xiaolong Yang and Yicheng Wu. 2015. Whether or not multiple occurrences of ziji can take distinct antecedents: A deictic perspective. *Journal of Pragmatics*, 87:142–155.

Xiulin Yang and Nathan Schneider. 2024. The relative clauses AMR parsers hate most. In *Proceedings of the Fifth International Workshop on Designing Meaning Representations @ LREC-COLING 2024*, pages 151–161, Torino, Italia. ELRA and ICCL.

Yuhan Zhang, Edward Gibson, and Forrest Davis. 2023. Can language models be tricked by language illusions? easier with syntax, harder with semantics. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 1–14, Singapore. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, Yusheng Su, Haozhe Ji, Jian Guan, et al. 2021a. Cpm: A large-scale generative chinese pre-trained language model. *AI Open*, 2:93–99.

Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. 2021b. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese. *Preprint*, arXiv:2110.06696.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019*, page 241.

Zhe Zhao, Yudong Li, Cheng Hou, Jing Zhao, et al. 2023. Tencentpretrain: A scalable and flexible toolkit for pre-training models of different modalities. *ACL 2023*, page 217.

Jianyu Zheng and Ying Liu. 2023. What does chinese bert learn about syntactic knowledge? *PeerJ Computer Science*, 9.

A In-Context Minimal Pair Templates for Different Binding Patterns

Constructions	In-context Minimal Pair Template
Blocking Effect	***Original:*** 她$_f$知道我$_i$相信自己$_i$。 *She$_f$ knows that I$_i$ trust myself$_i$.* ***Within Template*** 如果她知道我相信自己，那么我相信我自己/*她。 *If she$_f$ knows that I trust self, then I trust myself/*her.*
Animacy Effect	***Original:*** 他$_m$说这本书$_t$改变了自己$_m$。 *He$_m$ said the book changed him$_m$.* ***Within Template*** 如果他说这本书改变了自己，那么这本书改变了他/*它自己。 *If he said that this book changed self, then this book changed him/*itself.*
Subject Orientation	***Original:*** 他$_m$给她$_f$关于自己$_f$的书。 *He$_m$ gave her his own book.* ***Within Template*** 如果他给她关于自己的书，那么书是他/*她的。 *If he gave her a book about self, then the book is about him/*her.*
Verb Orientation	***Original:*** 她$_f$知道他$_m$巴结自己$_{f/m}$。 *She$_f$ knows that he$_m$ flatter her$_f$/her$_f$.* ***Within Template:*** 如果她知道他巴结自己，那么他巴结她/*他自己。 *If she knows that he flattered self, he flattered her/*himself.*

Table 6: In-context minimal pair templates corresponding to various binding constructions.

B Minimal Pair Examples for Natural Data

(8) Blocking Effect

a. Original Sent:
她会第一个承认我真是有自己的一套习惯。
ta$_i$ hui diyige chengren wo zhenshi you
she would first admit I really have
ziji de yitao xiguan
self DE one habit.

She would be the first to admit that I really have my own habit.

b. Minimal pair sent I:
她会第一个承认我真是有我自己的一套习惯。
ta$_i$ hui diyige chengren wo zhenshi you
she would first admit I really have
woziji de yitao xiguan
self DE one habit.

She would be the first to admit that I really have my own habit.

c. Minimal pair sent II:
* 她会第一个承认我真是有她自己的一套习惯。
ta$_i$ hui diyige chengren wo zhenshi you
she would first admit I really have
woziji de yitao xiguan
self DE one habit.

**She would be the first to admit that I really have her own habit.*

(9) Animacy Effect

a. Original Sent:
...因为他还不懂得瘟疫在威胁着自己。
ta$_i$ ... yinwei ta hai bu
... because he yet NEG understand
dongde wenyi zai weixie zhe ziji
plague is threaten ASP self.

... because he still doesn't understand that the plague is threatening him.

b. Minimal pair sent I:
...因为他还不懂得瘟疫在威胁着他自己。
... yinwei ta hai bu dongde wenyi
... because he yet NEG understand plague
zai weixie zhe ta-ziji
is threaten ASP himself.

... because he still doesn't understand that the plague is threatening him.

c. Minimal pair sent II:
*...因为他还不懂得瘟疫在威胁着它自己。
ta$_i$ ... yinwei ta hai bu
... because he yet NEG understand
dongde wenyi zai weixie zhe ta-ziji
plague is threaten ASP itself.

**... because he still doesn't understand that the plague is threatening itself.*

(10) Verb Orientation Reflexive Verb

a. Original Sent:
她会想，他在炫耀自己高人一等的教育。
ta hui xiang, ta zai xuanyao ziji
she will think, he is boast self
gaorenyideng de jiaoyu
superior DE education

She would think, he is boasting about his superior education.

b. Minimal pair sent I:
她会想，他在炫耀他自己高人一等的教

育。

ta hui xiang, ta zai xuanyao ta-ziji
she will think, he is boast he-self

gaorenyideng de jiaoyu
superior DE education

She would think, he is boasting about his own superior education.

c. Minimal pair sent II:
*她会想，他在炫耀她自己高人一等的教育。

ta hui xiang, ta zai xuanyao ta-ziji
she will think, he is boast she-self

gaorenyideng de jiaoyu
superior DE education

**She would think, he is boasting about her own superior education.*

(11) Verb Orientation Non-reflexive Verb

a. Original Sent:
少女一下子注意到，少年正在目不转睛地望着自己。

shaonv yixiazi zhuyidao, shaonian
girl suddenly notice, boy

zhengzai mubuzhuanjing-de wang zhe
PROG fixedly-ADV gaze ASP

ziji
self

The girl suddenly noticed that the boy was staring fixedly at her.

b. Minimal pair sent I:
少女一下子注意到，少年正在目不转睛地望着自己。

shaonv yixiazi zhuyidao, shaonian
girl suddenly notice, boy

zhengzai mubuzhuanjing-de wang zhe
PROG fixedly-ADV gaze ASP

ziji
self

The girl suddenly noticed that the boy was staring fixedly at her.

c. Minimal pair sent II:
*少女一下子注意到，少年正在目不转睛地望着他自己。

shaonv yixiazi zhuyidao, shaonian
girl suddenly notice, boy

zhengzai mubuzhuanjing-de wang zhe
PROG fixedly-ADV gaze ASP

ta-ziji
he-self

**The girl suddenly noticed that the boy was staring fixedly at himself.*

(12) Subject Orientation

a. Original Sent:
王小姐带着马伯乐就到自己房里来。

Wang xiaojie daizhe Ma Bole jiu dao
Miss Wang bring Ma Bole then arrive

ziji fang li lai
self room in come

Miss Wang brought Ma Bole and then went to her own room.

b. Minimal pair sent I:
王小姐带着马伯乐就到她自己房里来。

Wang xiaojie daizhe Ma Bole jiu dao
Miss Wang bring Ma Bole then arrive

ta ziji fang li lai
she self room in come

Miss Wang brought Ma Bole and then went to her own room.

c. Minimal pair sent II:
*王小姐带着马伯乐就到他自己房里来。

Wang xiaojie daizhe Ma Bole jiu dao
Miss Wang bring Ma Bole then arrive

ta ziji fang li lai
he self room in come

**Miss Wang brought Ma Bole and then went to his own room.*

C Prompts for GPT-3.5 and GPT-4o

Prompt A 下面两个句子哪个能自然，更容易接受？在这里，更自然指的是一个句子听起来符合母语者日常的语言使用习惯，读起来顺畅且易于理解。请只输出A或者B。A: 句子1 B: 句子2。 *Which of the following two sentences sounds more natural and is easier to accept? Here, "more natural" refers to a sentence that aligns with the everyday language use of native speakers, reads smoothly, and is easy to understand. Please output only "A" or "B". A: sentence 1 B: sentence 2.*

Prompt B 下面两个句子哪个能自然？请只输出A或者B, 然后给出解释。A: 句子1 B: 句子2。 *Which of the following sentences sounds more natural? Please output only "A" or "B"* ***and then give me your explanations****. A: sentence 1 B: sentence 2.*

Prompt C 下面两个句子哪个能自然，更容易接受？在这里，更自然指的是一个句子听起来符合母语者日常的语言使用习惯，读起来顺畅且易于理解。请只输出A或者B, 然后给出解释。A: 句子1 B: 句子2。 *Which of the following sentences sounds more natural and is easier to accept? Here, "more natural" refers to a sentence that aligns with the everyday language use of native speakers, reads smoothly, and is easy to understand. Please output only "A" or "B"* ***and then give me your explanations****. A: sentence 1 B: sentence 2.*

D Performance of GPT-3.5 and GPT-4o with Varying Positions of the Correct Sentence in Prompts: Option A, Option B, or Mixed

This section presents the experimental results of `GPT-3.5` and `GPT-4o` tested by altering the order of sentence pairs, where the correct sentence is either always placed in Option A, always in Option B, or randomly shuffled between the two. Due to this bias, `GPT-3.5` does not perform well in the mixed setting either, because half the correct sentences in the sentence pairs are put in Option B.

As we can see, `GPT-3.5` shows a clear preference for Option A. When all correct sentences are placed in Option A in the prompt, `GPT-3.5` achieves perfect accuracy. However, when the correct sentences are all placed in Option B, its performance declines to the lowest accuracy.

Similarly, `GPT-4o` struggles to make consistent judgments when the order of the two sentences is switched, displaying a bias toward Option B instead.

The detailed results can be found in Table 7 and Table 8.

E Training data distribution of evaluated language models

The training data of the language models we test is listed in Table 9.

	GPT-4o			GPT-3.5		
Prompt	**M**	**A**	**B**	**M**	**A**	**B**
Blocking	27.5	12.5	45.0	62.5	90.0	30.0
Animacy	100.0	100.0	100.0	100.0	100.0	92.5
Verb$_{refl}$	65.0	37.5	97.5	55.0	100.0	12.5
Verb$_{nonrefl}$	100.0	97.5	100.0	57.5	100.0	20.0
SO	50.0	30.0	77.5	50.0	100.0	0.0
Average	68.5	55.5	84.0	65.0	98.0	31.0

Table 7: Performance of GPT-4o and GPT-3.5 across different order settings of the minimal pairs in the synthetic data setting. M: correct options have mixed orders; A: correction options are always option-A; B: correction options are always option-B

	GPT-4o			GPT-3.5		
Prompt	**M**	**A**	**B**	**Mixed**	**A**	**B**
Blocking	100.0	97.5	95.0	57.5	90.0	25.0
Animacy	97.5	100.0	95.0	80.0	97.5	72.5
Verb$_{refl}$	100.0	100.0	100.0	60.0	97.5	22.5
Verb$_{nonrefl}$	100.0	100.0	100.0	90.0	100.0	95.0
SO	97.5	92.5	92.5	60.0	97.5	12.5
Average	99.0	98.0	96.5	69.5	96.5	45.5

Table 8: Performance of GPT-4o and GPT-3.5 across different order settings of the minimal pairs in the natural data setting.

Model	Training Data Domain
bert-base-chinese (Devlin et al., 2019)	Chinese Wikipedia
chinese-lert-base (Cui et al., 2022a)	Chinese Wikipedia, encyclopedia, news, and question answering web
chinese-lert-large (Cui et al., 2022a)	Chinese Wikipedia, encyclopedia, news, and question answering web
chinese-pert-base (Cui et al., 2022b)	Chinese Wikipedia, encyclopedia, news, and question answering web
chinese-pert-large (Cui et al., 2022b)	Chinese Wikipedia, encyclopedia, news, and question answering web
mengzi-bert-base (Zhang et al., 2021b)	Chinese Wikipedia, Chinese News, and Common Crawl
mengzi-bert-base-fin (Zhang et al., 2021b)	Chinese Wikipedia, Chinese News, and Common Crawl, Finance data
ernie-1.0-base-zh (Sun et al., 2019)	Chinese Wikepedia, Baidu Baike, Baidu news and Baidu Tieba
mBERT (Devlin et al., 2019)	Top 100 languages with the largest Wikipedias
XLM-R-base (Conneau et al., 2019)	CommonCrawl
XLM-R-large (Conneau et al., 2019)	CommonCrawl
mt5-small (Xue et al., 2020)	CommonCrawl
mt5-large (Xue et al., 2020)	CommonCrawl
GPT2 (Zhao et al., 2019)	CLUECorpus-small (from Common Crawl) (Xu et al., 2020)
GPT2-medium (Zhao et al., 2019)	CLUECorpus-small (from Common Crawl) (Xu et al., 2020)
GPT2-large (Zhao et al., 2019)	CLUECorpus-small (from Common Crawl) (Xu et al., 2020)
GPT2-xlarge (Zhao et al., 2023)	CLUECorpus-small (from Common Crawl) (Xu et al., 2020)
GLM-4-9b-chat (GLM et al., 2024)	NA
CPM-Generate (Zhang et al., 2021a)	Encyclopedia, Webpage, Story, News, Dialog
GPT-3.5 (OpenAI, 2023)	NA
GPT-4o (OpenAI et al., 2023)	NA

Table 9: Models, training data and information source