Do Language Models Use Logophoric Cues? Evidence from Mandarin Chinese Long-Distance Reflexive

Yunfang Dong
Independent Researcher
yunfang.dong@outlook.com

Abstract

Resolving anaphora requires integrating syntactic, semantic, and discourse information. Mandarin Chinese offers a particularly revealing case through the reflexive *ziji*, whose interpretation permits long-distance binding licensed by logophoric cues (i.e., cues relevant to discourse perspective). While these cues have been extensively studied in linguistic theory and psycholinguistic experiments, it remains an open question to what extent such cues are captured by computational models. We investigate this question by probing large language models' sensitivity to four logophoric cues known to license long-distance binding of *ziji*: predicate type, perspective marking, discourse topicality, and discourse relation. Using minimal pairs and surprisal-based measures, we assess whether models exhibit systematic biases toward non-local antecedents in logophoric contexts. Across two model families, we find that (i) models exhibit above-chance sensitivity to all four cues; (ii) lexically anchored cues are more robustly captured than discourse-level cues; and (iii) some cues generalize cross-lingually, whereas others appear to depend on language-specific training data. Taken together, these findings provide non-English evidence that large language models capture certain aspects of logophoricity, yet continue to struggle with discourse-level representations that are central to human anaphora resolution.[1]

1 Introduction

Understanding the meaning of natural language sentences requires linking expressions to their referents. While proper names such as *Mary* typically refer directly, anaphoric expressions like pronouns and reflexives derive their interpretation from entities in discourse. The mechanisms governing such dependencies have been extensively studied in linguistic theory, particularly in the domain of anaphora and binding. Under Binding Principle A (Chomsky, 1981), reflexive anaphors must be bound within a local structural domain. This generalization is well supported in languages such as English. However, cross-linguistic evidence reveals systematic exceptions. In several languages, including Mandarin Chinese, Japanese, and Icelandic, certain reflexives permit **long-distance binding**, allowing them to take antecedents outside their local domain. For example, in Mandarin Chinese, the reflexive *ziji* in Example (1a) may refer either to the embedded subject *Lisi* or to the matrix subject *Zhangsan*, as indicated by the co-indexed elements. Figure 1 illustrates both local and long-distance binding of *ziji* in a syntactic tree.

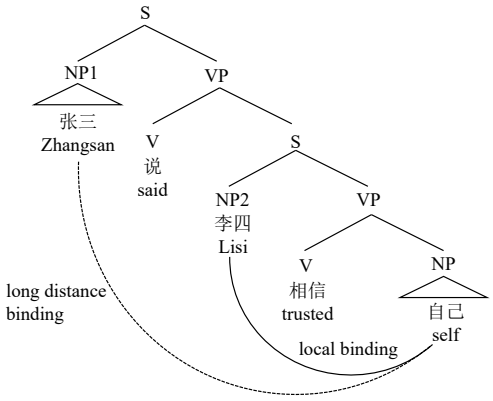

Figure 1: Syntactic tree showing local and long distance binding of Mandarin reflexive *ziji*.

(1) a. 张三$_i$说李四$_j$相信自己$_{i/j}$。
Zhangsan$_i$ say Lisi$_j$ trust self$_{i/j}$
Zhangsan said that Lisi trusted him/himself.
b. 张三$_i$听说李四$_j$相信自己$_{i/j}$。
Zhangsan$_i$ hear Lisi$_j$ trust self$_{i/j}$
Zhangsan heard that Lisi trusted him/himself.

[1]Code and data are available at: https://github.com/yunfang-dong/mandarin-logophoricity-llm

Findings of the Association for Computational Linguistics: ACL 2026, pages 43061–43072

The availability of long-distance interpretations challenges purely structural accounts of binding and has motivated functional approaches that treat *ziji* as a **logophor**—an expression whose interpretation is determined by discourse perspective rather than syntactic locality alone (Clements, 1975). Prior theoretical and experimental studies have identified a range of cues that facilitate logophoric binding including predicate type and discourse topicality (Huang and Liu, 2000; Liu, 2020; Lyu and Kaiser, 2023, 2024). For instance, Example (1b), though only different from Example (1a) in the matrix predicate (i.e., "said" vs. "heard"), can lead to different interpretations of the reflexive *ziji*.

Despite the extensive theoretical literature on logophoricity, it remains unclear whether contemporary large language models (LLMs) are sensitive to such logophoric cues in reflexive resolution. While recent work has demonstrated strong performance of LLMs on coreference-related tasks (Le and Ritter, 2024), existing evaluations of syntactic and discourse phenomena are predominantly English-focused (Dong et al., 2024; Liu et al., 2025) so rarely probe long-distance reflexives which English lacks (Xiang et al., 2021; Song et al., 2022). Therefore, we still lack a principled understanding of whether LLMs capture the discourse representations required to support logophoric binding.

In this paper, we present a controlled, theory-driven evaluation of LLMs on long-distance reflexive interpretation in Mandarin Chinese. We focus on the simplex reflexive *ziji*, which has been extensively studied in previous theoretical and experimental work and provides a particularly informative testbed for examining interactions between perspective, discourse, and binding.[2] To this end, we construct minimal pairs targeting four logophoric cues—predicate type, perspective marking, discourse topicality, and discourse relation—established in the human studies, and assess model preferences using surprisal-based metrics.

Our results show that LLMs are responsive to all four cues at above-chance levels, but with substantial variation across cue types and model families. Lexically anchored cues are captured more robustly than discourse-level cues. Comparisons across model families, sizes, and training regimes further reveal that some logophoric cues generalize cross-lingually, whereas others benefit from language-specific training data. These findings have implications for downstream tasks that rely on accurate entity tracking and coreference resolution—such as question answering, dialogue understanding, and summarization—especially in non-English contexts. Thus, our work contributes both a linguistically grounded evaluation and practical insights for the design of systems sensitive to discourse-level phenomena.

[2]The compound form *ta-ziji* is typically analyzed as locally bound and is not considered here.

2 Related Work

2.1 Previous Linguistic Accounts of Long-Distance Reflexives

Long-distance reflexives (LDRs) have attracted sustained attention in linguistic theory, and a wide range of proposals have been developed to account for their behavior. Broadly, existing approaches can be divided into **formal** and **functional** accounts. **Formal approaches** attempt to explain LDRs within a purely structural framework. Some analyses accommodate LDRs by parameterizing the notion of the local binding domain (Manzini and Wexler, 1987), while others reduce long-distance binding to successive applications of local binding operations (Tang, 1989; Cole et al., 1990; Cole and Sung, 1994; Cole and Wang, 1996; Huang and Tang, 1991). While influential, such accounts have been argued to struggle with a number of well-documented empirical patterns, including the blocking effect (Xue et al., 1994), number asymmetry (Tang, 1989; Huang and Tang, 1991), and person asymmetry (Pan, 2013).

In contrast, **functional approaches** analyze LDRs as a discourse-dependent phenomenon. Early work by Clements (1972, 1975) introduced the notion of logophoricity, whereby certain pronouns refer to the individual whose perspective—such as speech, thought, or consciousness—is being reported. This perspective-based analysis has been extended to reflexives in languages such as Mandarin Chinese and Japanese (Sells, 1987; Chen, 1992; Yu, 1992, 1996; Pan, 2013; Charnavel, 2021).

A central question in this line of work concerns **what linguistic cues license logophoric interpretations** and how antecedents are selected. Kuno (1987) proposes an Empathy Hierarchy, according to which discourse topics and other prominent entities are more likely to serve as perspective centers. Sells (1987) further distinguishes the roles of Source, Self, and Pivot in logophoric interpretation.

Building on these ideas, Culy (1994) proposes a hierarchy of logophoric licensors (e.g., speech > thought > knowledge > direct perception), while Huang and Liu (2000) characterizes Mandarin *ziji* as sensitive to the "consciousness" of an event participant and identifies specific linguistic environments—such as *causal* versus *temporal* relations—that facilitate long-distance binding.

Importantly, many of these theoretical claims have been substantiated by experimental evidence. Psycholinguistic studies have shown that reflexive interpretation is systematically modulated by predicates, discourse relations, perspective shifts, and discourse prominence (Liu, 2020, 2022; Lyu and Kaiser, 2023, 2024). These experimentally validated cues provide a principled basis for probing LLMs, allowing computational studies to test whether models capture the conditions that license logophoric and long-distance binding.

2.2 Syntactic Evaluation in LLMs

The rapid progress of large language models has sparked extensive interest in their linguistic competence, particularly with respect to core syntactic phenomena. A common evaluation paradigm involves minimal pairs differing only in a targeted linguistic property, with probability-based measures such as *surprisal* used to assess model sensitivity (Hale, 2001). Using this methodology, prior work has examined subject–verb agreement, filler–gap dependencies, negative polarity items, and other core phenomena (Marvin and Linzen, 2018; Wilcox et al., 2018; Hu et al., 2020b).

Binding principles have also been studied computationally. Prior work has investigated English reflexives (Principle A) (Marvin and Linzen, 2018; Hu et al., 2020a) and pronouns (Principle B) (Davis, 2022). For Mandarin Chinese, reflexive resolution has been incorporated into linguistic benchmarks such as CLiMP (Xiang et al., 2021) and SlING (Song et al., 2022). More recently, Yang (2025) provides a fine-grained analysis of Mandarin *ziji*, examining several properties such as blocking effects, animacy, and ambiguity.

However, existing computational studies largely focus on structural availability of antecedents rather than the discourse conditions that license long-distance binding. While recent computational work has begun to investigate discourse-sensitive phenomena, including implicit discourse relation recognition and the interaction between syntactic configurations and discourse relations (Dong et al., 2024; Liu et al., 2025), no prior work has systematically evaluated whether LLMs are sensitive to the logophoric cues—such as predicate type, perspective marking, discourse topicality, and discourse relation—that have been shown in human experiments to promote long-distance reflexive interpretations. Addressing this gap, the present work provides the first computational investigation of how LLMs respond to logophoric environments.

3 Evaluating LLM Sensitivity to Logophoric Cues in Long-Distance Reflexives

3.1 Logophoric Cues

We investigate whether large language models (LLMs) exhibit sensitivity to logophoric cues that license long-distance binding of the Mandarin reflexive *ziji*. We focus on four cues that have been independently motivated in theoretical work and validated in controlled human judgment experiments: Predicate type, perspective marker, discourse topicality, and discourse relation.

3.1.1 Predicate Type

Culy (1994)'s Logophoric Hierarchy describe a hierarchical classification of prediate types varying in their degree to trigger logophoricity:

> speech > thought > knowledge > direct perception

According to Culy (1994), the more higher a predicate is on the hierarchy, the more reliable a situation is reported by the predicate, thus creating a more likely context to induce logophoric interpretations.

As illustrated in Example (1), *ziji* allows both local and non-local antecedents. Because *shuo* ("say"; speech predicate) ranks higher than *ting-shuo* ("hear"; direct perception predicate) on the Logophoric Hierarchy, long-distance binding to the matrix subject is more strongly licensed in (1a) than in (1b), a contrast confirmed experimentally by Lyu and Kaiser (2024).

3.1.2 Perspective Marking

Liu (2020, 2022) showed that internal versus external perspective markings influence logophoric licensing differently. She used the following example:

> (2) a. 据张三$_i$说，这件事伤害了自己$_i$。

According to Zhangsan$_i$, this matter hurt self$_i$.

According to Zhangsan, this matter hurt him.

b. 说到张三$_i$，这件事伤害了自己$_i$。

Speaking of Zhangsan$_i$, this matter hurt self$_i$.

Speaking of Zhangsan, this matter hurt him.

In (a), "According to" demonstrates the report is anchored to Zhangsan's perspective, whereas "Speaking of" in (b) introduces Zhangsan from the external speaker's viewpoint. Liu (2022) showed stronger logophoric binding of *ziji* in (a) in human experiments.

3.1.3 Discourse Topicality

Previous research suggests when the matrix subject is the discourse topic, it is more likely to function as empathy centers and thus license logophoric binding (Kuno and Kaburaki, 1977; Dillon et al., 2014, 2016; Jäger et al., 2015).

(3) a. 张三最近做事效率很低。张三$_i$说李四$_j$批评了自己$_{i/j}$。

Zhangsan recently do task efficiency very low. Zhangsan$_i$ say Lisi$_j$ criticize self$_{i/j}$

Zhangsan has recently been working inefficiently. Zhangsan said Lisi criticized him/himself.

b. 项目最近推进效率很低。张三$_i$说李四$_j$批评了自己$_{i/j}$。

Project recently advance efficiency very low. Zhangsan$_i$ say Lisi$_j$ criticize self$_{i/j}$

The project has recently been progressing inefficiently. Zhangsan said Lisi criticized him/himself.

In (a), as the context sentence introduces Zhangsan as a topic, Zhangsan in the following sentence becomes discourse-old and topical; in (b), Zhangsan is discourse-new. Lyu and Kaiser (2023) found increased long-distance binding in (a) in human experiments.

3.1.4 Discouse Relation

Beyond reportive contexts, logophoricity can arise when an individual is construed as conscious of or emotionally affected by an event. Huang and Liu (2000) proposes that as causal relations involve a higher degree of engagement than, they promote such consciousness more than temporal relations. They use the contrast *yinwei* ("because") and *dang* ("when") as an example which was later empirically supported by Liu (2020).

(4) a. 张三$_i$离开公司，因为李四$_j$批评自己$_{i/j}$。

Zhangsan$_i$ leave the company because Lisi$_j$ criticize self$_{i/j}$.

Zhangsan left the company because Lisi criticized him/himself.

b. 张三$_i$离开公司，当李四$_j$批评自己$_{i/j}$。

Zhangsan$_i$ leave the company when Lisi$_j$ criticize self$_{i/j}$.

Zhangsan left the company when Lisi criticized him/himself.

3.2 Experiment Design

3.2.1 Dataset Construction

To explore language models' sensitivity of logophoric cues in long distance binding, we adopt a minimal-pair design in which sentence pairs differ only in the target logophoric cue. We examine four logophoric cues and adapted the stimuli from prior experimental studies as mentioned in Section 3.1.

Specifically, predicate type is manipulated by varying the matrix predicate (*shuo* "say" vs. *tingshuo* "hear"), while holding the embedded predicate constant. Perspective marking is manipulated by contrasting constructions that encode the matrix subject's internal perspective (*ju NP shuo* "according to NP") with constructions that present the subject from an external perspective (*shuo-dao NP* "speaking of NP"). Discourse topicality is manipulated by preceding the target sentence with a context sentence that either establishes NP1 as discourse-old or leaves it discourse-new. [3] Dis-

[3] Discourse topicality was manipulated by varying the sentence-initial aboutness topic in the preceding context (e.g., *Zhangsan* vs. *the project*), thereby modulating the discourse accessibility of the potential antecedent. Context pairs were constructed to remain comparable in semantic domain, syntactic frame, sentence length. A diverse set of lexical realizations was further used across a large number of minimal pairs to minimize item-specific biases. This controlled design follows prior work such as Lyu and Kaiser (2023).

Cue	Template	Manipulation
Predicate type	NP1 + **V1** + NP2 + (Adv) + V2 + ziji	V1: *shuo* vs. *tingshuo*
Perspective marking	**According to / Speaking of** NP1, NP2 + (Adv) + V2 + ziji	internal vs. external perspective
Discourse topicality	**Context**. NP1 + V1 + NP2 + (Adv) + V2 + ziji	NP1 discourse-old vs. new
Discourse relation	NP1 + VP1. **CONNECTIVE** + NP2 + V2 + ziji	*yinwei* vs. *dang*

Table 1: Summary of experimental templates and manipulated logophoric cues. Within each minimal pair, only the bolded element varies.

course relation is manipulated by contrasting causal and temporal connectives (*yinwei* "because" vs. *dang* "when"). [4]

To scale up the dataset, we abstract these schemas (as shown in Table 1) and prompt GPT-5 mini (OpenAI, 2025) to generate candidate sentences under strict structural constraints.[5] For each cue, substantially more candidate items were generated than ultimately retained. Two linguistically trained annotators independently evaluated all candidates for grammaticality, naturalness, and correct cue manipulation, and any item judged problematic by either annotator was discarded and replaced. This iterative procedure continued until 100 minimal pairs per cue were jointly approved.

3.2.2 Grammatical Position of *ziji*

In the stimuli above, *ziji* appears as a direct object. However, *ziji* may occur in a range of grammatical positions, including subject, indirect object, and possessor (Huang, 2000). To examine whether the grammatical function of *ziji* modulates logophoric binding, we further vary the grammatical position of *ziji* in the predicate-type condition, yielding four configurations: subject, direct object, indirect object, and possessor (see Appendix A for examples). This results in 400 additional minimal pairs.

In total, the dataset consists of 700 minimal pairs, covering four logophoric cues and one structural manipulation.

3.2.3 Models

We evaluate eight open source autoregressive models from two families with different sizes:[6] Llama-3.2-1B, Llama-3.2-1B-Instruct, Llama-3.2-3B, Llama-3.2-3B-Instruct, llama-3.1-8B, Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen-2.5-1.5B, Qwen-2.5-1.5B-Instruct, Qwen-2.5-3B, Qwen-2.5-3B-Instruct, Qwen-2.5-7B, Qwen-2.5-7B-Instruct (Yang et al., 2024).

All models are trained on multilingual data. The Llama family does not undergo explicit training on Chinese data, whereas the Qwen models are trained with Chinese-language data.

3.2.4 Evaluation Metric

We measure LLM sensitivity to logophoric cues using *surprisal*, defined as the negative log-likelihood of a token x giving its preceding context C.

$$Surprisal = -\log P(x|C)$$

Lower surprisal indicates that the model considers the token more probable as a continuation. Because *ziji* is inherently ambiguous, surprisal at the reflexive itself does not directly reveal the model's sensitivity to logophoric cues. Following Yang (2025), we append a disambiguating paraphrase that explicitly resolves the antecedent, introduced by the connective 也就是说("That is") . Example (5) illustrates the stimuli obtained by appending disambiguating paraphrases to the sentences in Example (1):

(5) a. 张三$_i$说李四$_j$相信自己$_i$。也就是说，张三$_i$说李四$_j$相信张三$_i$。

Zhangsan$_i$ say Lisi$_j$ trust self. That

[4] We note that discourse connectives can convey multiple sense relations. The connective *dang* "when" can sometimes convey a causal relation. However, as corpus evidence from the PDTB-3 Annotation Manual (Webber et al., 2019; Prasad et al., 2017) show that "yingwei" overwhelmingly signals causal relations and "dang" predominantly signals temporal relations and as our evaluation tests relative preference shifts across strongly biased environments rather than assuming categorical interpretation, we keep using the *yinwei* "because" vs. *dang* "when" contrast in Huang and Liu (2000) and Liu (2022).

[5] Note that we strictly controlled for lexical bias by ensuring that all subordinate-clause verbs are bi-directional predicates (following Yang (2025)), which independently allow both local and long-distance interpretations of ziji. This prevents the verb itself from structurally enforcing one interpretation.

[6] We only evaluated open-source large language models, as we could access logits and compute surprisal values only from such models.

is, Zhangsan$_i$ say Lisi$_j$ trust Zhangsan$_i$.

Zhangsan said that Lisi trusted him/himself. That is, Zhangsan said that Lisi trusted Zhangsan.

b. 张三$_i$听说李四$_j$相信自己$_i$。也就是说，张三$_i$听说李四$_j$相信张三$_i$。

Zhangsan$_i$ hear Lisi$_j$ trust self. That is, Zhangsan$_i$ hear Lisi$_j$ trust Zhangsan$_i$.

Zhangsan heard that Lisi trusted him/himself. That is, Zhangsan heard that Lisi trusted Zhangsan.

We compute surprisal at the disambiguating antecedent (*Zhangsan*) to assess whether the model shows higher preference for long-distance binding under high-logophoric conditions, which corresponds to the first condition for each of the four cues described in Section 3.1.[7]

Formally, for each minimal pair, we expect:

$$\begin{aligned} &Surprisal\big(\text{“}Zhangsan\text{”} \mid C_{\text{high-logophoric}}\big) \\ &\qquad < \\ &Surprisal\big(\text{“}Zhangsan\text{”} \mid C_{\text{low-logophoric}}\big) \end{aligned} \tag{1}$$

We then report the proportion of minimal pairs where surprisal is lower in the high-logophoric condition, providing a measure of the model's sensitivity to each logophoric cue.

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[\Delta_i > 0] \tag{2}$$

where Δ_i = Surprisal$(Zhangsan_i \mid C^{(i)}_{\text{low}})$ − Surprisal$(Zhangsan_i \mid C^{(i)}_{\text{high}})$ and N is the number of sentence pairs for each cue (N = 100) .

4 Results

In this section, we present results on how large language models respond to four logophoric cues argued to license long-distance binding of the Mandarin reflexive *ziji*. Each cue is implemented as a minimal pair contrasting a logophoricity-inducing condition with a less logophoric alternative.

Table 2 shows, for each model and cue, the percentage of items in which the model favors nonlocal binding under the high-logophoric condition. From the table, we observe above-chance averaged proportions across all four cues, suggesting LLMs exhibit sensitivity to logophoric environments.

[7]When *ziji* does not appear sentence-finally, we sum surprisal over the entire disambiguating continuation.

	Pred.	Persp.	Rel.	Top.
L-3.2-1B	0.77*	0.48	0.19*	0.74*
L-3.2-1B-IT	0.75*	0.32*	0.11*	0.51
L-3.2-3B	0.51	0.60	0.72*	0.56
L-3.2-3B-IT	0.81*	0.13*	0.27*	0.41
L-3.1-8B	0.70*	0.61*	0.97*	0.64*
L-3.1-8B-IT	0.86*	0.55	0.72*	0.66*
Q-2.5-1.5B	0.36*	0.89*	0.72*	0.48
Q-2.5-1.5B-IT	0.35*	0.77*	0.80*	0.40
Q-2.5-3B	1.00*	0.84*	0.34*	0.75*
Q-2.5-3B-IT	0.98*	0.93*	0.15*	0.39
Q-2.5-7B	0.96*	0.61	0.79*	0.40
Q-2.5-7B-IT	0.89*	0.64*	0.30	0.45
Average	0.74	0.61	0.51	0.53

Table 2: The proportion of items for which models prefer long distance binding in logophoricity-inducing condition than in less logophoric alternative. Asterisks indicate statistical significance under paired t-tests with Bonferroni correction ($\alpha = .05/48 \approx .001$). Pred. = Predicate type, Persp. = Perspective marking, Rel. = Discourse relation, Top. = Discourse topicality.

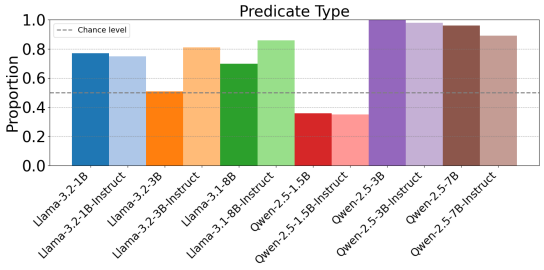

Figure 2: Model performance on predicate type manipulations.

Among the four types of cues, models exhibits the highest sensitivity to predicate type, followed by perspective marking, discourse topicality, and discourse relation. This suggests that models are more sensitive to lexically anchored cues for logophoric licensing than discourse-level cues.

This discrepancy may arise because discourse-level cues require models to track inter-sentential discourse context and maintain a representation of discourse prominence or causal relations, which is more challenging than capturing lexical-semantic cues such as predicate type, whose effects are largely contained within a single sentence.

Below we analyze and discuss each cue in turn.

4.1 Predicate Type

The predicate type manipulation tests whether models distinguish predicates that are known to introduce a logophoric environment from those that are less so.

As shown in Figure 2, most models exhibit

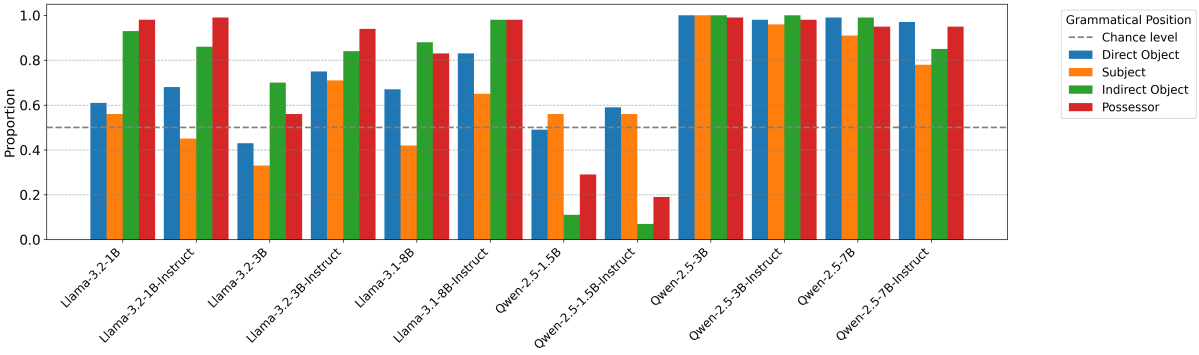

Figure 3: Model performance across different grammatical positions of *ziji*.

above-chance sensitivity to this contrast: reflexives embedded under speech predicates *shuo* ("say") are more likely to be interpreted as long-distance bound than those embedded under perception predicates *tingshuo* ("hear"). This pattern aligns with the Logophor Hierarchy (Culy, 1994) as well as recent experimental findings (Lyu and Kaiser, 2024), according to which speech predicates strongly favor logophoric interpretations in human judgments.

Qwen models exhibit greater discrepancy between model sizes, with 3B and 7B variants approaching categorical preference for speech predicates and 1B models displaying below chance results. This suggests that explicit Chinese training appears to facilitate predicate-based logophoric licensing, though the predicate type is not primarily driven by scale. And instruction tuning slightly reduces sensitivity across Qwen models.

Llama models show consistently high sensitivity to predicate type across sizes and training regimes, despite not being explicitly trained on Chinese data. This indicates that predicate-based logophoric licensing may be supported by cross-linguistically general semantic representations, rather than language-specific patterns.

To further probe the interaction between predicate semantics and structure, we further examined whether the grammatical position of *ziji* within the embedded clause modulates predicate-based sensitivity. As shown in Figure 3, Llama models exhibit stronger preference for long-distance binding when *ziji* appears in indirect object and possessor positions, with reduced sensitivity when *ziji* occupies the embedded subject and the direct object positions. Qwen models exhibit a different pattern. Larger Qwen models (3B and 7B) maintain consistently high sensitivity across syntactic positions. Smaller Qwen models, by contrast, display

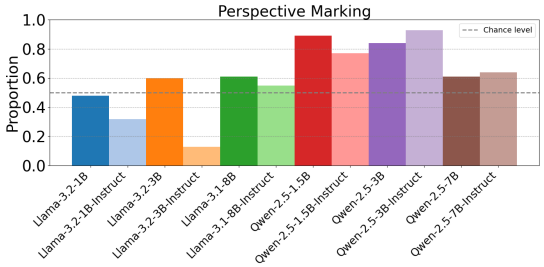

Figure 4: Model performance on perspective marking manipulations.

positional preferences, favoring direct object and subject positions over indirect object and possessor positions. This divergence points to differences in how structural and semantic cues are integrated across model families.

4.2 Perspective Marking

The perspective marking manipulation examines whether models distinguish between internal and external perspective markers in logophoric licensing: *ju...shuo* ("According to") vs. *shuo-dao...* ("Speaking of"). Figure 4 shows that Llama models exhibit only moderate sensitivity to this contrast, with smaller models like Llama 1B, 1B Instruct and 3B Instruct falling below chance. Sensitivity increases with model size and the effects of instruction tuning are inconsistent. In contrast, Qwen models demonstrate strong sensitivity to perspective markers across sizes and training variants. This asymmetry suggests that perspective marking may rely more heavily on language-specific constructions that benefit from explicit Chinese training.

4.3 Discourse Topicality

The discourse topicality manipulation tests whether models prefer non-local binding when the matrix subject is discourse-old and continues as the topic

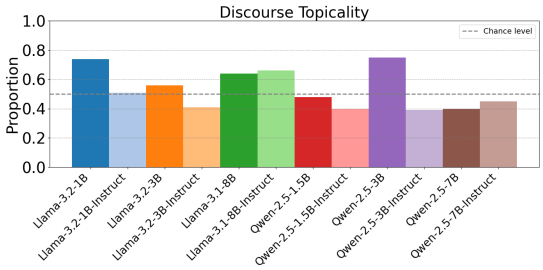

Figure 5: Model performance on discourse topicality manipulations.

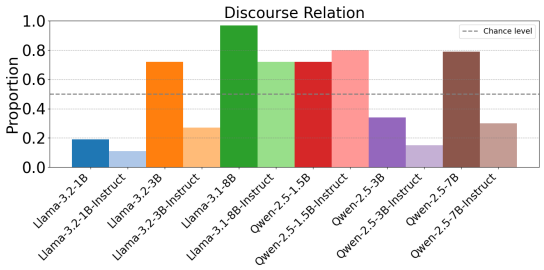

Figure 6: Model performance on discourse relation manipulations.

versus discourse-new. As shown in Figure 5, Both model families show broadly similar and only moderate sensitivity to discourse topicality. Although several models prefer long-distance binding when the antecedent is discourse-prominent, the effect remains weaker and less consistent than for lexically anchored cues such as predicate type and perspective marking in the Qwen models. In contrast, the Llama models show topicality effects that are weaker than predicate type but comparable to, and in some cases stronger than, perspective marking. Notably, the Qwen models do not show stronger topicality effects despite their primary exposure to Chinese training data. Overall, topicality contributes only an above-chance bias toward nonlocal binding rather than a strong determinant of interpretation.

To assess whether the observed differences between high- and low-logophoric conditions were reliable, we conducted paired t-tests across items for each model and cue type, with Bonferroni correction for multiple comparisons ($\alpha = .05/48 \approx .001$). The majority of cue effects remained statistically significant after correction, particularly for lexically anchored cues such as predicate type and perspective marking. Discourse-level cues, especially topicality, showed weaker and less consistent significance patterns across models. These results confirm that the proportion differences reported above reflect systematic contrasts rather than item-level variation.

4.4 Discourse Relation

Finally, the discourse relation manipulation examines whether causal connectives such as *yinwei* ("because") promote long-distance binding more strongly than temporal connectives like *dang* ("when"). Overall, causal relations tend to increase non-local binding preferences, but the effect varies across models.

As shown in Figure 6, Llama models show increasing sensitivity over sizes, with Llama-3.1-8B model achieving near-ceiling performance, while Qwen models flunctuate widely across size. This variability indicates that causal discourse relations are potentially informative for logophoric licensing but are not uniformly captured by models.

5 General Discussion and Conclusion

We present the first computational investigation of how contemporary large language models respond to logophorical cues that license long-distance binding, a phenomenon that has been extensively studied in theoretical and experimental linguistics but remains unexplored in computational work. Focusing on the Mandarin reflexive *ziji*, we examine four logophoric cues -predicate type, perspective marking, discourse topicality, and discourse relation - all of which have been studied independently in human behavioral studies. By adopting a minimal-pair paradigm and a surprisal-based metric, we isolate the contribution of each cue and directly test whether LLMs encode sensitivity to logophoric licensing conditions.

Our results demonstrate that LLMs exhibit above-chance sensitivity to all four logophoric cues, aligning with previous experimental findings (Liu, 2020; Lyu and Kaiser, 2023, 2024) and providing evidence that these models can utilize discourse and perspective information beyond purely syntactic constraints when resolving long-distance anaphora. These cues have been independently established in prior psycholinguistic research as reliable licensing conditions for long-distance interpretations of *ziji*. The present results provide converging computational evidence that models are sensitive to the same logophoric environments identified in earlier human studies.

While prior linguistic work has typically investigated these cues in isolation, our unified evalua-

tion makes their relative strengths directly comparable. We found that lexically anchored cues—most notably predicate type—are well captured across model families, whereas discourse-level cues such as topicality and discourse relations yield weaker effects overall. This asymmetry is particularly pronounced in the Qwen models, which show consistently stronger sensitivity to lexically anchored cues than to discourse-level factors. We hypothesize that this pattern may arise because discourse-level cues require maintaining and updating inter-sentential discourse representations, which may be more challenging for current models than encoding lexically grounded, sentence-internal information. This motivates future comparative research on how different types of logophoric cues interact in shaping long-distance anaphora.

Comparing Llama and Qwen model families further reveals that logophoric sensitivity is shaped by both cross-linguistic generalization and language-specific experience. Some cues, such as predicate type, appear to rely on knowledge that transfers robustly across languages: Llama models, despite lacking explicit Chinese training, show strong and stable sensitivity to this cue. Other cues, such as perspective marking, benefit substantially from Chinese-specific training data, as evidenced by Qwen models' stronger effects. Still others, including discourse topicality, show broadly comparable effects across model families rather than a clear advantage associated with additional Chinese training. This divergence underscores that logophoric licensing is not a monolithic phenomenon but instead draws on multiple sources of information -lexical semantics, syntactic position, and discourse structure - that models integrate with different degrees of reliability.

Beyond differences across cue types, we also examined whether logophoric sensitivity varies systematically with model size and instruction tuning. We find no consistent increase in logophoric sensitivity with larger model size or instruction tuning; in many cases, these factors are even associated with reduced sensitivity. Instruction-tuned models differ from their base counterparts in being further optimized for helpfulness and task-following behavior through supervised fine-tuning and preference alignment, which may bias them toward simpler, more direct interpretations favoring local binding and suppress more nuanced long-distance logophoric licensing. With respect to model size, the observation that larger models do not consistently exhibit greater sensitivity to logophoric cues than smaller ones suggests that such sensitivity is not simply a byproduct of increased parameter count, but may instead depend more strongly on the linguistic diversity and structure of the pre-training corpus. This interpretation is consistent with the findings of Hale and Stanojević (2024), who show that scaling alone does not reliably improve models' ability to capture syntactic universals, suggesting that sensitivity to structurally conditioned dependencies may depend on properties of training data rather than model size per se.

While much prior work has focused primarily on English, where long-distance reflexive binding is highly constrained and logophoric effects are relatively limited, Mandarin *ziji*, by contrast, provides a particularly rich testing ground for theories of perspective, discourse, and binding. By grounding our evaluation in this typologically distinct system, we show that probing models on languages with different binding properties can surface aspects of discourse and perspective sensitivity that remain largely invisible in English-centric evaluations. Our study contributes to a more nuanced understanding of how neural language models internalize perspective-sensitive meaning and opens the door to future computational cross-linguistic and representational studies of logophoricity and long-distance binding.

Limitations

While our study reveals LLMs' sensitivity to logophoric cues, certain limitations remain, pointing to directions for future research.

Mandarin Chinese reflexive *ziji* provides a well-established testbed for evaluating model sensitivity to logophoric cues in long-distance binding. Our results reveal nuanced patterns across model families, sizes, and training variants, suggesting promising directions for future cross-linguistic investigation in languages such as Japanese, Icelandic and Dutch.

We focus on the simplex reflexive *ziji*. Mandarin also contains the morphologically marked compound reflexive *ta-ziji*, which parallels English reflexives and is generally locally bound, though its binding properties remain debated. Future computational studies could explore whether models capture potential long-distance interpretations of *ta-ziji*.

Our evaluation uses controlled, template-based stimuli, which allow precise manipulation of individual logophoric cues. Extending analyses to naturalistic corpora could complement these results and provide additional insight into how models generalize to real-world scenarios.

We examine four experimentally validated logophoric cues. Theoretical work suggests additional cues, such as motion verbs 来 ("come") and 去 ("go"), may also influence long-distance binding. Exploring a broader set of cues could further illuminate the interaction of syntax and discourse in large language models.

Furthermore, our stimuli primarily used third-person antecedents to maintain controlled comparisons of logophoric cues. Mandarin *ziji* exhibits a "blocking effect", where long-distance binding is typically blocked if a first- or second-person pronoun intervenes. Future research could test how person-based constraints interact with other logophoric cues.

Acknowledgements

We thank Bonnie Webber for insightful comments on earlier drafts of this paper that substantially improved the paper and suggested several promising directions for future research; and Xixian Liao for helpful discussions and valuable feedback on earlier versions of the manuscript.

References

Isabelle Charnavel. 2021. Logophoricity, perspective, and reflexives. *Annual Review of Linguistics*, 7(1):131–155.

Ping Chen. 1992. The reflexive ziji in chinese: Functional vs. formalist approaches. *Research on chinese linguistics in Hong Kong*, pages 1–36.

Noam Chomsky. 1981. *Lectures on Government and Binding: The Pisa Lectures*. Number 9 in Studies in Generative Grammar. Foris Publications, Dordrecht, Holland.

George N Clements. 1975. The logophoric pronoun in ewe: Its role in discourse. *Journal of West African Languages*, 10(2).

George Nickerson Clements. 1972. *The verbal syntax of Ewe*. University of London, School of Oriental and African Studies (United Kingdom).

Peter Cole, Gabriella Hermon, and Li-May Sung. 1990. Principles and parameters of long-distance reflexives. *Linguistic Inquiry*, pages 1–22.

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

Peter Cole and Chengchi Wang. 1996. Antecedents and blockers of long-distance reflexives: The case of chinese ziji. *Linguistic Inquiry*, pages 357–390.

Christopher Culy. 1994. Aspects of logophoric marking. *Linguistics*, 32(6).

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.

Brian Dillon, Wing-Yee Chow, Matthew Wagers, Taomei Guo, Fengqin Liu, and Colin Phillips. 2014. The structure-sensitivity of memory access: evidence from mandarin chinese. *Frontiers in psychology*, 5:1025.

Brian Dillon, Wing-Yee Chow, and Ming Xiang. 2016. The relationship between anaphor features and antecedent retrieval: Comparing mandarin ziji and taziji. *Frontiers in psychology*, 6:1966.

Yunfang Dong, Xixian Liao, and Bonnie Webber. 2024. Syntactic preposing and discourse relations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2790–2802.

Aaron Grattafiori and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*. Includes discussion of LLaMA 3 family models.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*.

John T. Hale and Miloš Stanojević. 2024. Do LLMs learn a true syntactic universal? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17106–17119, Miami, Florida, USA. Association for Computational Linguistics.

Jennifer Hu, Sherry Yong Chen, and Roger Levy. 2020a. 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.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020b. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1725–1744.

C-T James Huang and C-S Luther Liu. 2000. Logophoricity, attitudes, and ziji at the interface. In *Long distance reflexives*, pages 141–195. Brill.

C-T James Huang and C-C Jane Tang. 1991. The local nature of the long-distance reflexive in chinese. *Long distance anaphora*, pages 263–282.

Yan Huang. 2000. *Anaphora: A cross-linguistic approach*. Oxford University Press.

Lena A Jäger, Felix Engelmann, and Shravan Vasishth. 2015. Retrieval interference in reflexive processing: experimental evidence from mandarin, and computational modeling. *Frontiers in psychology*, 6:617.

Susumu Kuno. 1987. *Functional syntax: Anaphora, discourse and empathy.* University of Chicago Press.

Susumu Kuno and Etsuko Kaburaki. 1977. Empathy and syntax. *Linguistic Inquiry*, pages 627–672.

Nghia T Le and Alan Ritter. 2024. Are language models robust coreference resolvers? In *First Conference on Language Modeling*.

Meinan Liu, Yunfang Dong, Xixian Liao, and Bonnie Webber. 2025. Multi-token mask-filling and implicit discourse relations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12546–12560, Suzhou, China. Association for Computational Linguistics.

Yingtong Liu. 2020. Logophoricity and mandarin exempt reflexives. *University of Pennsylvania Working Papers in Linguistics*, 26.

Yingtong Liu. 2022. *Evidence from behavioral experiments: Information theory and discourse-based accounts of long-distance dependencies*. Harvard University.

Jun Lyu and Elsi Kaiser. 2023. Multiple constraints modulate the processing of chinese reflexives in discourse. *Glossa Psycholinguistics*, 2(1).

Jun Lyu and Elsi Kaiser. 2024. Logophoricity and the processing of chinese reflexives. *Journal of East Asian Linguistics*, 33(4):559–597.

M Rita Manzini and Kenneth Wexler. 1987. Parameters, binding theory, and learnability. *Linguistic Inquiry*, pages 413–444.

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.

OpenAI. 2025. Gpt-5-mini. `https://openai.com`. Accessed: 2026-01.

Haihua Pan. 2013. *Constraints on reflexivization in Mandarin Chinese*. Routledge.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2017. The penn discourse treebank: An annotated corpus of discourse relations. In *Handbook of linguistic annotation*, pages 1197–1217. Springer.

Peter Sells. 1987. Aspects of logophoricity. *Linguistic Inquiry*, 18(3):445–479.

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.

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

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. *The Penn Discourse Treebank 3.0 Annotation Manual*. Available from the Linguistics Data Consortium, https://catalog.ldc.upenn.edu/docs/LDC2019T05/.

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.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina von der Wense. 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.

Ping Xue, Carl Pollard, and Ivan A Sag. 1994. A new perspective on chinese ziji. In *the Proceedings of the Thirteenth West Coast Conference on Formal Linguistics*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Xiulin Yang. 2025. Language models at the syntax-semantics interface: A case study of the long-distance binding of chinese reflexive ziji. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3808–3824.

William XF Yu. 1992. Challenging chinese reflexive data. *The Linguistic Review*, 9(3).

Xian Fu Yu. 1996. *A study of Chinese reflexives*. University of London, School of Oriental and African Studies (United Kingdom).

A Examples of Mandarin reflexive *ziji* at different grammatical positions

Examples of Mandarin reflexive *ziji* at different grammatical positions: subject, direct object, indirect object, and possessor.

(1) **Subject position**

张三$_i$说李四$_j$觉得自己$_{i/j}$是一个天才。

Zhangsan$_i$ say Lisi$_j$ think self$_{i/j}$ is a genius.

Zhangsan said that Lisi thought that he/himself was a genius.

(2) **Direct object (embedded clause)**

张三$_i$说李四$_j$恨自己$_{i/j}$。

Zhangsan$_i$ say Lisi$_j$ hate self$_{i/j}$.

Zhangsan said that Lisi hated him/himself.

(3) **Indirect object**

张三$_i$说李四$_j$给自己$_{i/j}$一本书。

Zhangsan$_i$ say Lisi$_j$ give self$_{i/j}$ a book.

Zhangsan said that Lisi gave him/himself a book.

(4) **Possessor** [8]

张三$_i$说李四$_j$拿了自己$_{i/j}$的照片。

Zhangsan$_i$ say Lisi$_j$ take self$_{i/j}$'s photo.

Zhangsan said that Lisi took his/his own photo.

[8] We focus on the possessor of a direct object in the present study.