Examining Large Language Models' form-meaning mappings of information structure constructions in Mandarin Chinese

Shihui Li[♣] **Xiaojuan Tan**[◇♡] **Jelke Bloem**[♣♠]

[♣]Institute for Logic, Language and Computation, University of Amsterdam
[◇]Amsterdam Center for Language and Communication, University of Amsterdam
[♡]STL (Savoirs, Textes, Langage) lab, Université de Lille
[♠]Data Science Center, University of Amsterdam
shihui.li@student.uva.nl x.tan@uva.nl j.bloem@uva.nl

Abstract

Construction Grammar (CxG) knowledge in language models has been extensively studied for English, but remains underexplored in other languages. In Mandarin Chinese, the *ba* (把, disposal) and *bei* (被, passive) constructions are widely used for managing information structure. They foreground topical elements (information structure) and encode systematic form-meaning mappings (CxG), particularly with respect to the semantic role of the object. We probe language models' linguistic competence with these constructions using minimal pairs, constructing a new minimal-pair dataset comprising seven paradigms that target both syntactic constraints and verb–construction compatibility. Our results show that it remains a challenge for many models to capture the form-meaning mappings underlying the *ba* construction, although they achieve high accuracy on paradigms driven by surface syntactic cues.

1 Introduction

Construction grammar views linguistic knowledge as pairings between form and meaning. A central question in recent NLP research is whether neural language models (LMs) acquire such construction-level generalizations, beyond surface co-occurrence patterns. While constructional knowledge in LMs has been extensively probed for English (Tayyar Madabushi et al., 2020; Li et al., 2022; Veenboer and Bloem, 2023; Bonial et al., 2025; Mackintosh et al., 2025), analogous evidence for other languages remains comparatively limited. Furthermore, while interesting pairs of form and meaning have been investigated, there has been little attention for constructions that touch upon the 'third component' of grammar — information structure (Leino, 2013).

Information structure constructions, such as the English passive construction, mainly serve to organize how information is presented relative to discourse context, rather than contributing semantically. Mandarin Chinese has constructions that combine both information-structural and semantic functions. Proper use of such a construction requires both a good semantic fit with the verb and satisfying information-structural constraints (Liu and Ambridge, 2021). This may make them more challenging to acquire for large language models.

To address this gap, we focus on two Mandarin Chinese constructions that serve important information-structure functions compared to canonical SVO clauses. In those clauses, the subject typically realizes the Agent role and the object the Patient role. *ba* and *bei* constructions alter surface word order while preserving underlying thematic relations, much like the English passive, but crucially, these structures are not freely interchangeable. The *ba* construction imposes a semantic requirement of disposal or affectedness (Sybesma, 1999): it licenses only bounded, resultative predicates that bring about a clear change of state in the object. Both constructions also have syntactic constraints.

By probing whether language models encode these characteristics of *ba* and *bei* structures, we gain more insight into (i) the capacities of different models to represent form-meaning mappings of Mandarin Chinese, a language typologically distinct from English, and (ii) the models' ability to represent information structure constructions. To do so, we construct a minimal-pair dataset comprising seven paradigms involving *ba* and *bei* constructions.[1] These paradigms include the main syntactic constraints of each construction as well as their alternations with the information-structurally neutral canonical SVO structures. We evaluate four language model families that vary in training lan-

[1] https://github.com/li-shihui/mandarin-information-structure-dataset

guage and model size, including both base models and instruction-tuned models. Base models are evaluated using log-probability-based comparisons, while instruction-tuned models are evaluated using a forced-choice prompting setup.

2 Background

2.1 *Ba* and *bei* constructions in Mandarin

Ba and *bei* are two unique constructions in Mandarin Chinese that reorganize the information structure of the clause by reshaping discourse prominence and perspective. Both constructions position the patient before the verb, which departs from the canonical SVO order to foreground the object in the discourse. The *ba* construction is commonly described as expressing a disposal meaning, specifying how an object is handled (Wang, 1954). It requires a definite object and foregrounds its affectedness (Pinker et al., 1987), which presents the event as one that brings about a clear change of state in that object (Liu and Ambridge, 2021). For example, in 张三把李四救了 (Zhangsan *ba* Lisi saved; "Zhangsan saved Lisi"), the construction highlights that the definite object (Lisi) undergoes a bounded change of state—from being in danger to being safe—thereby emphasizing the concrete result of the action on the patient. In this way, it promotes the object to a topic-like, discourse-prominent position.

The *bei* construction is the canonical Mandarin Chinese passive which, by contrast, promotes the patient to grammatical subject position and aligns it with the primary discourse topic, backgrounding or optionally omitting the agent (Liu and Ambridge, 2021). For example, in 李四被张三打了 (Lisi bei Zhangsan hit; "Lisi was hit by Zhangsan"), Lisi appears as the grammatical subject and discourse topic, while Zhangsan is demoted to a post-*bei* agent phrase that can be omitted (李四被打了). The construction highlights Lisi's affected experience and typically suggests a surprising impact on the patient. Thus, both constructions shape which entity is construed as topical, prominent, and perspectivally central in the event.

Alternations with canonical SVO. Although many SVO sentences can be converted into *ba* and *bei* sentences, such conversions are subject to strict semantic constraints on the predicate. *Ba* and *bei* constructions do not share the same information structure. In *bei*-passives, the post-*bei* Agent NP typically introduces new discourse information, whereas in *ba*-actives, the verbal event itself is more likely to carry the focus of new information (Liu and Ambridge, 2021). With respect to semantic constraints, verbs that do not directly affect the object, such as 发现 (discover) and 知道 (know), cannot appear in *ba* constructions, but they are otherwise permissible in *bei* constructions (Zhang, 2001). Predicates denoting psychological, experiential, existential, or possessive meanings are generally incompatible with *ba* (Wang, 1943). For example, verbs meaning "to know", such as *zhidao* (知道), *liaojie* (了解), and *zhixiao* (知晓), are acceptable in SVO and *bei* constructions but not in *ba* constructions.

Examples (1a–c) illustrate a typical alternation among SVO (1b), *ba* (1a), and *bei* (1c) sentences.

(1a) Ta *ba* shu na-zou le.
He *ba* book take-away PFV.[2]
'He took the book away.'

(1b) Ta na-zou le shu.
He take-away PFV book.
'He took the book away.'

(1c) Shu *bei* ta na-zou le.
Book *bei* he take-away PFV.
'The book was taken away by him.'

In addition to these information-structural and semantic constraints, both constructions have characteristic syntactic constraints. In both affirmative and negative sentences, aspectual marking (e.g., *le*) is required to denote a bounded, resultative event. In negative sentences, negators (e.g., *meiyou*, *mei*) should appear in a structurally high position preceding *ba* and *bei*, instead of directly preceding verbs as in canonical SVO structures.

Aspect restrictions. Aspectual marking constitutes a typical syntax–semantics interface phenomenon in *ba* and *bei* constructions. Both *ba* and *bei* constructions typically co-occur with the aspectual marker *le*. The *ba* construction is highly transitive and thus favors perfective marking to signal event boundedness (Hopper and Thompson, 1980). More generally, both constructions denote bounded events affecting an object and are often infelicitous without *le* (Yang, 1995).

Negation. In canonical SVO sentences, negators such as *mei(you)* appear before the verb, as in (2b).

[2]PFV refers to the perfective aspect marker *le*.

In contrast, in negative *ba* and *bei* sentences, the negator must precede the construction marker *ba* or *bei* (Zhu, 1982). The *ba* and *bei* counterparts of (2b) are shown in (2a) and (2c), respectively.

(2a) Ta mei-you *ba* shu na-zou.
He NEG *ba* book take-away.
'He didn't take the book away.'

(2b) Ta mei-you na-zou shu.
He NEG take-away book.
'He didn't take the book away.'

(2c) Shu mei-you *bei* ta na-zou.
Book NEG *bei* he take-away.
'The book wasn't taken away by him.'

2.2 CxG Capabilities of Language Models

Minimal pairs (MPs) are widely used to evaluate language models' fine-grained grammatical knowledge (Linzen et al., 2016; Warstadt et al., 2020; Hu et al., 2020; Huang et al., 2025; Pestel et al., 2025; Jumelet et al., 2026). A MP consists of two sentences that differ only in the target grammatical domain. In acceptability evaluation, an effective LM should assign higher probability (or lower perplexity) to the grammatical or semantically valid sentence in each pair.

Several MP-based benchmarks have been proposed for Mandarin, including CLiMP (Xiang et al., 2021), SLING (Song et al., 2022), and ZhoBLiMP (Liu et al., 2026).[3] However, these resources provide limited coverage of *ba*/*bei* constructions and focus primarily on syntactic patterns. No systematic examination of the syntax–semantics mappings within these constructions has been conducted. For example, CLiMP contains only one *ba* paradigm and one passive paradigm, while SLING does not explicitly target these constructions. ZhoBLiMP offers broader coverage (13 *ba* paradigms and 12 passive paradigms), but for the *bei* construction it includes only affirmative sentences and excludes negation. Moreover, these datasets mainly probe word order and part-of-speech patterns, offering limited leverage for testing whether models capture construction-specific meanings such as disposal. In addition, all these datasets use a relatively limited range of verbs and rarely test predicate compatibility with constructions. These limitations motivate the use of paradigms targeting the syntax–semantics interface in LMs, for which conversion among SVO, *ba*, and *bei* constructions is particularly suitable.

Related work on English argument structure constructions suggests that language models can learn abstract constructional templates and associate them with meaning beyond lexical content. Li et al. (2022) show that, within a given construction, even semantically anomalous verbs cluster closer to prototypical verbs than to incongruent ones in embedding space, indicating sensitivity to constructional meaning. However, none of the Chinese MP benchmarks above include minimal pairs consisting of two infelicitous sentences.

This field is highly Anglocentric and we are only aware of three studies that relate to constructional knowledge of LLMs beyond English. First, Bunzeck et al. (2025) experiment with German BabyLMs to investigate whether the constructional profile of child-directed speech is beneficial to acquisition, by manipulating the relative frequencies of different constructions in the training data.

Second, Huang et al. (2025) perform an evaluation using minimal pairs of Chinese Verb-Resultative Complement Constructions. They introduce the ZhVrcMP benchmark of minimal pairs and find that several decoder language models are able to assess the grammaticality of minimal pairs of this construction (based on perplexity) reasonably well. This is the study most closely related to ours – while their construction is lexical-semantic, encoding event structure and change-of-state semantics, it interacts with information structure. For example, it can be combined with the *ba* construction. They do not draw conclusions beyond construction-specific results, apart from the influence of parameter size, and find that Zh-Pythia outperforms Mistral.

Third, building upon Scivetti and Schneider (2025) who investigated the English noun-preposition-noun construction ('day by day') in BERT, Gorzoni et al. (2026) extend this investigation to Italian. They observe evidence for constructional representation also for Italian. Furthermore, they find that multilingual models can perform an identification task but underperform in a disambiguation task. Such potential limitations of constructional generalization in multilingual models can only be identified by investigating non-English languages.

[3] Liu et al. (2026) do not report the paradigm-specific accuracies, so we cannot make a direct comparison.

2.3 Information Structure Constructions

So far, the field of NLP has paid very limited attention to information-structural capabilities of LLMs more broadly and information structure constructions more specifically. An exception is Wu et al. (2025), who studied the production of referring expression by LLMs as influenced by factors such as recency and discourse status. However, this investigation is limited to whether names, pronouns or descriptions are used, rather than any information structure constructions. Stephenson et al. (2022) find that BERT can be tuned to produce contrastive focus in text-to-speech synthesis, but with limited performance. Ozaki et al. (2022) investigate whether LSTMs can learn cleft and topicalization constructions as examples of filler-gap constructions, but these are viewed from a syntactic perspective in the study.

From a CxG perspective, other argument structure constructions have been studied, in BERT (Ramezani et al., 2025) and in decoder models (Li et al., 2022; Bonial et al., 2025), but without discussing information-structural influence. The closest work we are aware of is Fujihara et al.'s (2022) study of Japanese topicalization constructions with GPT2-small, finding that its generalizations are not human-like. We are not aware of any investigations into information structure constructions with more recent models, even for English, even though this area covers focus constructions such as wh-clefts, topic-comment constructions such as topicalization or voice constructions such as passives.

To address these gaps, we construct a new minimal-pair dataset featuring four underexplored paradigms in *ba* and *bei* constructions: (1) the placement of negation and *le* in *ba* constructions, (2) negation placement in *bei* constructions, (3) constructional conversion among *ba*, *bei*, and SVO sentence types, and (4) *ba* constructions with semantically infelicitous verbs. We generate minimal pairs using templates. To examine the effects of model size, training language, and model type (base vs. instruction-tuned), we evaluate four model families.

3 Experimental Setup

3.1 Data

We constructed our minimal-pair dataset based on paradigms adapted from the ZhoBLiMP dataset (Liu et al., 2026). The dataset contains seven paradigms involving the *ba* and *bei* constructions, with 300 minimal pairs for each paradigm. Examples of each paradigm are provided in Table 1. Two paradigms (*ba le* and *ba neg*) were directly adopted from ZhoBLiMP. One paradigm (*bei neg*) was created through deterministic transformation of the ZhoBLiMP *ba neg* items into corresponding *bei* constructions. The remaining paradigms were derived from the ZhoBLiMP *passive_suo* paradigm through sentence structural transformations.

The minimal pairs were validated by a native speaker. Human evaluation data were collected only for the *ba le* and *ba neg* in Liu et al.'s (2026) work. The authors selected five samples from each paradigm and evaluated them with five native speakers, reporting accuracy rates of 0.9091 and 0.8182, respectively. The dataset covers four types of phenomena: two paradigms targeting the *ba* construction, one targeting the *bei* construction, two paradigms involving *ba*/*bei*/SVO conversion, and two paradigms targeting infelicitous *ba* constructions. Our dataset is publicly available at: `https://github.com/li-shihui/mandarin-information-structure-dataset`.

For both felicitous and infelicitous *ba* constructions, we manipulated (i) the presence of the aspectual marker *le* (了) and (ii) the position of the negator *meiyou* (没有) in negative sentences. For *bei* constructions, we manipulated the position of the negator *meiyou*. For consistency, we used the same lexicons in the *ba neg* and *bei neg* paradigms.

For the *ba*/*bei*/SVO conversion paradigms, we selected three stative cognition verbs—*zhidao* (知道), *liaojie* (了解), and *zhixiao* (知晓), which are compatible with *bei* and SVO constructions but infelicitous in *ba* constructions. Using these verbs, we generated corresponding *ba*, *bei*, and SVO sentence variants as in (1a–c). For the infelicitous *ba* minimal pairs, we treated the *ba* sentences in the conversion paradigms as the acceptable sentences (grammatical but semantically incompatible), and derived unacceptable counterparts by removing *le* or placing the negator in an incorrect position.

3.2 Models

To provide a more comprehensive picture of LMs' knowledge of the target constructions, we conducted experiments on four model families differing in model scale and primary training language:

Zh-Pythia (14M-1.4B) (Liu et al., 2026): A series of Chinese language models trained on Chi-

Phenomenon	Paradigm	Acceptable Example	Unacceptable Example
BA	LE	Wangwu ba women piping le。 Wangwu ba us criticize PFV. 'Wangwu criticized us.'	Wangwu ba women piping。 Wangwu ba us criticize. 'Wangwu criticized us.'
	Negation	Women mei-you ba daxiang mazui。 We NEG ba elephant anesthetize. 'We did not anesthetize the elephant.'	Women ba daxiang mei-you mazui。 We ba elephant NEG anesthetize. 'We did not anesthetize the elephant.'
BEI	Negation	Daxiang mei-you bei women mazui。 Elephant NEG bei us anesthetize. 'The elephant was not anesthetized by us.'	Daxiang bei women mei-you mazui。 Elephant bei us NEG anesthetize. 'The elephant was not anesthetized by us.'
BA/BEI/SVO Conversion	BA/BEI	Naxie xiaoxi bei Song nüshi zhidao le。 Those messages bei Ms. Song know PFV. 'Those messages became known to Ms. Song.'	Song nüshi ba naxie xiaoxi zhidao le。 Ms. Song ba those messages know PFV. 'Those messages became known to Ms. Song.'
	BA/SVO	Song nüshi zhidao naxie xiaoxi le。 Ms. Song know those messages PFV. 'Ms. Song learned those messages'	Song nüshi ba naxie xiaoxi zhidao le。 Ms. Song ba those messages know PFV. 'Ms. Song learned those messages'
INFELICITOUS BA	LE	Song nüshi ba naxie xiaoxi zhidao le。 Ms. Song ba those messages know PFV. 'Ms. Song learned those messages.'	Song nüshi ba naxie xiaoxi zhidao。 Ms. Song ba those messages know. 'Ms. Song knows those messages.'
	Negation	Song nüshi mei-you ba naxie xiaoxi zhidao。 Ms. Song NEG ba those messages know. 'Ms. Song did not know those messages.'	Song nüshi ba naxie xiaoxi mei-you zhidao。 Ms. Song ba those messages NEG know. 'Ms. Song did not know those messages.'

Table 1: Examples of seven paradigms in four phenomena in *ba* and *bei* constructions included in the dataset.

nese texts with the GPT-NeoX architecture and a Chinese tokenizer. The model sizes are 14M, 70M, 160M, 410M and 1.4B.

DeepSeek (Bi et al., 2024; DeepSeek-AI et al., 2025): A large language model developed by the Chinese company DeepSeek AI. It is trained on a large-scale corpus of approximately 2 trillion tokens in English and Chinese. The proportion of Chinese data, and whether the model is instruction-tuned on Chinese are not publicly specified. We evaluated `deepseek-llm-7b-base` and `deepseek-v3.2` API.

LLaMA-3.1-8B (Grattafiori et al., 2024): LLaMA 3.1 is a family of pretrained and instruction-tuned multilingual autoregressive language models developed by Meta. Although trained on data covering multiple languages, it does not include dedicated Chinese adaptation. We evaluated both `llama-3.1-8B-base` and `llama-3.1-8B-Instruct`.

Mistral-7B v0.3: A commercial 7B-parameter English model optimized for instruction tasks, pretrained without Chinese adaptation. We evaluated both `Mistral-7B v0.3` (base model) and `Mistral-7B-Instruct v0.3`.

All models, except for the DeepSeek API, were accessed via Hugging Face (Wolf et al., 2020)[4].

[4] https://huggingface.co

3.3 Evaluation

We adopt different evaluation methods for base and instruction-tuned models. Base models are evaluated using Mean Log Probability (Lau et al., 2017) comparisons over minimal pairs and perplexity, while instruction-finetuned models are evaluated using prompt-based methods. Exact models and their corresponding evaluation methods are summarized in Table 2

Mean Log Probability. Given a sentence $\gamma = (x_1, \ldots, x_T)$, the mean log probability (MLP) under a language model Θ is defined as

$$MLP_\Theta(\gamma) = \frac{1}{T}\sum_{t=1}^{T} \log p(x_t \mid x_{<t}). \quad (1)$$

Minimal-Pair Accuracy. Let p denote a set of minimal pairs (g, u), where g is an acceptable sentence and u is its unacceptable counterpart. A model Θ is considered correct on a pair if it assigns a higher mean log probability to g than to u. We compute the accuracy over each paradigm as

$$S(p) = \frac{1}{|p|}\sum_{(g,u)\in p} \mathbb{I}\big(MLP_\Theta(g) > MLP_\Theta(u)\big), \quad (2)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Model	Log-Probability	Prompt-Based	Main Training Language
Zh-Pythia-14M, 70M, 160M, 410M, 1.4B	✓	–	Chinese
DeepSeek-7B-Base	✓	–	Chinese
DeepSeek API	–	✓	Chinese
LLaMA-3.1-8B-Base	✓	–	English
LLaMA-3.1-8B-Instruct	–	✓	English
Mistral-7B v0.3	✓	–	English
Mistral-7B-Instruct v0.3	–	✓	English

Table 2: Language models evaluated under log-probability and prompt-based settings. Log-probability evaluation compares sentence likelihoods within minimal pairs, while prompt-based evaluation uses forced-choice acceptability judgments.

Perplexity. Perplexity is derived from the mean log probability as

$$PPL_{\Theta}(\gamma) = \exp(-MLP_{\Theta}(\gamma)) \,. \tag{3}$$

Perplexity measures the model's average uncertainty per token. Lower values indicate better language modeling performance.

Relative Perplexity Increase. To quantify how strongly a model penalizes unacceptable sentences, we report the relative perplexity increase from the acceptable sentence to its unacceptable (*) counterpart:

$$\Delta\text{PPL} = \frac{\text{PPL}_{*} - \text{PPL}_{\text{acc}}}{\text{PPL}_{\text{acc}}}. \tag{4}$$

Larger values indicate that the model assigns substantially higher perplexity to the unacceptable variant, and is more sensitive to it.

For **instruction-tuned models**, we used a forced-choice prompting setup in which the model was asked to select the more natural sentence in each minimal pair. The prompt template was:

```
判断下面哪个句子更自然，必须回答A或B。
(Which of the following sentences is more
 natural? You must answer A or B.)
A: {sent1}
B: {sent2}
答案：(Answer:)
```

4 Results

***Ba* construction.** For the *ba* paradigms, we test the basic structural syntactic properties of the construction. Most models achieve ceiling performance on the *ba le* test (Table 3), even higher than the human evaluation (0.9091, see Section 3.1), with the only exceptions being Mistral-base (0.5067) and LLaMA-instruct (0.8400). In contrast, the negation placement test is more challenging. While Chinese-trained models remain strong (all above 0.8467, higher than human evaluation 0.8182), English-centric models show less consistent performances.

***Bei* construction.** This paradigm tests the basic negation syntax of *bei* constructions. Compared with *ba* negation, *bei* negation is generally better and more consistently handled by Chinese-trained models, which suggests that, given sufficient Chinese exposure, models capture negation placement in *bei* constructions more robustly than in *ba* constructions. In contrast, English-centric models show smaller gains or even declines from the *ba* negation test, as well as lower performance ceilings on *bei* negation.

***Ba/bei*/SVO conversion.** The conversion paradigms constitute more demanding tests than previous ones, as they require construction-level knowledge of verb-construction compatibility. Overall, DeepSeek shows the strongest performance on this phenomenon, achieving mean accuracies above 0.95 for both base and instruction-finetuned models, and substantially outperforming the next-best model family, Zh-Pythia. In contrast, English-centric models are considerably less reliable and exhibit substantial inconsistencies across different paradigms and model variants. Their accuracies range from 0.1667 to 0.7500. In particular, their performance is highly unstable across base and instruction models. The most extreme case occurs in the *ba* vs. *bei* paradigm, where the Mistral base model achieves 0.6267 accuracy, but the instruction-finetuned model drops to 0.0567 under prompting.

A further notable result is that for the relative perplexity in the *ba* vs. SVO test, all models except DeepSeek show negative or near-zero values (Appendix A). This indicates a systematic preference for structurally well-formed but semantically incompatible *ba* sentences over basic SVO sentences. Overall, these results suggest that the models struggle to capture the incompatibility between *ba* constructions and non-disposal verbs, and thus do not

Model	Size	*ba*			*bei*	*ba/bei*/SVO			infelicitous *ba*		
		ba le	ba neg	mean	bei neg	ba vs bei	ba vs svo	mean	ba infel le	ba infel neg	mean
LLaMA	8B	**1.0000**	0.8967	0.9484	0.8133	0.4700	0.1667	0.3184	**1.0000**	0.9733	0.9867
LLaMA-I	*8B*	*0.8400*	*0.8900*	*0.8650*	*0.7633*	*0.2933*	*0.7400*	*0.5167*	***1.0000***	*0.8733*	*0.9367*
Mistral	7B	**1.0000**	0.7167	0.8584	0.7833	0.6267	0.5300	0.5784	**1.0000**	0.8767	0.9384
Mistral-I	*7B*	*0.5067*	*0.9233*	*0.7150*	*0.8667*	*0.0567*	*0.7500*	*0.4034*	*0.1700*	*0.2967*	*0.2334*
DeepSeek	7B	**1.0000**	0.8967	0.9484	0.9967	**0.9967**	0.9467	0.9717	0.8167	0.9600	0.8884
DeepSeek-I	–	***1.0000***	***0.9967***	*0.9984*	***1.0000***	*0.9233*	***1.0000***	*0.9617*	*0.9733*	***1.0000***	*0.9867*
Zh-Pythia	14M	**1.0000**	0.8467	0.9234	0.8200	0.5733	0.1633	0.3683	**1.0000**	0.9067	0.9534
Zh-Pythia	70M	**1.0000**	0.8500	0.9250	0.9800	0.7233	0.2233	0.4733	**1.0000**	0.9433	0.9717
Zh-Pythia	160M	**1.0000**	0.9400	0.9700	0.9767	0.7767	0.3233	0.5500	**1.0000**	0.9933	0.9967
Zh-Pythia	410M	**1.0000**	0.9467	0.9734	0.9900	0.7867	0.2867	0.5367	**1.0000**	**1.0000**	1.0000
Zh-Pythia	1.4B	**1.0000**	0.9400	0.9700	0.9733	0.7800	0.3300	0.5550	**1.0000**	**1.0000**	1.0000

Table 3: Accuracy for each model across minimal-pair test sets. Italics denote prompt-based evaluation, otherwise log-probability evaluation. Model abbreviations: LLaMA = `LLaMA-3.1-8B`; *LLaMA-I* = `LLaMA-3.1-8B-Instruct`; Mistral = `Mistral-7B v0.3`; *Mistral-I*= `Mistral-7B-Instruct v0.3`; DeepSeek = `deepseek-llm-7b-base`; *DeepSeek-I*= `DeepSeek-v3.2` API.

Model	Size	*ba*		*bei*	*ba/bei*/SVO		infelicitous *ba*	
		ba le	ba neg	bei neg	ba vs bei	ba vs svo	ba infel le	ba infel neg
LLaMA	8B	339.02 / 1042.57	830.36 / 1211.66	1468.17 / 1896.35	506.57 / 506.92	753.60 / 506.92	506.92 / 1533.42	640.60 / 943.55
Mistral	7B	140.80 / 249.90	114.72 / 130.63	138.68 / 160.33	62.03 / 69.85	67.66 / 69.85	69.85 / 117.43	73.08 / 93.69
DeepSeek	7B	732.20 / 4195.14	1756.22 / 2946.24	828.25 / 3818.57	350.65 / 2124.42	486.47 / 2124.42	2124.42 / 3504.95	1655.30 / 3311.07
ZH-Pythia	14M	344.86 / 1168.73	673.58 / 927.10	484.44 / 602.72	170.86 / 172.53	246.34 / 172.53	172.53 / 373.15	256.78 / 321.56
ZH-Pythia	70M	397.61 / 1522.34	821.93 / 1076.10	384.67 / 792.91	137.32 / 178.34	230.25 / 178.34	178.34 / 483.19	286.30 / 445.81
ZH-Pythia	160M	390.81 / 1601.32	729.77 / 1371.87	425.68 / 842.07	146.59 / 193.22	232.14 / 193.22	193.22 / 673.98	292.73 / 607.99
ZH-Pythia	410M	347.07 / 1255.64	738.25 / 1261.99	432.98 / 967.59	137.05 / 192.96	226.14 / 192.96	192.96 / 624.42	280.02 / 694.09
ZH-Pythia	1.4B	350.49 / 1174.00	721.55 / 1155.81	424.74 / 758.66	135.43 / 195.12	229.74 / 195.12	195.12 / 626.78	288.64 / 730.78

Table 4: Good (acceptable) vs. bad (unacceptable) perplexity (good / bad) for each base model across minimal-pair test sets. LLaMA = `LLaMA-3.1-8B`; Mistral = `Mistral-7B v0.3`; DeepSeek = `deepseek-llm-7b-base`.

reliably encode the form-meaning mappings underlying the *ba* construction.

Infelicitous *ba*-construction. The infelicitous *ba* paradigms test whether models treat a formally well-formed *ba* sentence as acceptable when the verb class renders the construction semantically infelicitous. Overall, most models achieve high accuracies on these paradigms (Table 3), in some cases even higher than those for felicitous sentences. Instruction-finetuned Mistral is the only exception, with very low scores throughout (0.1700 on *ba infel le* and 0.2967 on *ba infel neg*).

This pattern suggests that the presence of the aspectual marker *le* remains a salient cue even when the underlying construction is semantically odd. There are two possible interpretations. On the one hand, this may indicate that aspectual marking and negation word-order constraints are robustly represented even under constructional infelicity. On the other hand, the high accuracies may reflect overgeneralization based on surface syntactic templates rather than genuine form-meaning mappings.

Beyond accuracy, we further examine the models' behavior using perplexity scores (Table 4). Comparing the mean perplexities of grammatical *ba* sentences with felicitous verbs (*ba* mean good) and infelicitous ones (infelicitous *ba* mean good), all models except DeepSeek show lower perplexity for infelicitous sentences, suggesting that they accept semantically incompatible verbs more readily within a well-formed *ba* structure. In contrast, the perplexity of DeepSeek-7B base is substantially higher in infelicitous sentences (1889.86 vs. 1244.21), which shows the model has great sensitivity to the incompatibility of cognition verbs with the *ba* construction.

4.1 Model Results

Among the base models (Table 3), DeepSeek-7B is consistently the strongest system, reaching an accuracy of over 0.9 in most paradigms. LLaMA-base performs well in the syntax tests (e.g. *ba le* 1.0000), but is close to chance on conversion (mean 0.3184), indicating limited constructional knowledge. Zh-Pythia shows a clear scaling trend, especially in the conversion mean accuracy, which increases from 0.3683 (14M) to 0.5550 (1.4B).

However, even at 1.4B, Zh-Pythia remains far below DeepSeek on *ba vs. svo* (0.3300 vs. 0.9467). On the simpler *ba* and *bei* paradigms, Zh-Pythia reaches near-ceiling performance already at 160M (e.g., *ba neg* 0.9400, *bei neg* 0.9767).

The instruction-tuned models largely perform worse than their base models. Mistral shows the largest performance drop relative to its base model, most notably on the *ba* and "infelicitous *ba*" paradigms (e.g., *ba infel le* from 1.0000 to 0.1700), suggesting strong sensitivity to prompt format and instruction-following behavior. In contrast, for LLaMA, prompting partially improves conversion behavior on *ba vs. svo* (0.1667 vs. 0.7400), but it still remains unstable across paradigms.

5 Discussion and Open Questions

Overall, models performed well on paradigms where minimal pairs could be distinguished by syntactic markers, but struggled in the conversion paradigm requiring construction-level knowledge of verb compatibility and in the semantically infelicitous paradigms.

Only DeepSeek exhibits consistent performance on verb–construction compatibility in *ba* constructions, while also performing well on both conversion paradigms across base and instruction-tuned models. Furthermore, it correctly assigns higher perplexity to infelicitous *ba* sentences than to their felicitous counterparts. In contrast, all other models achieve accuracies below 0.35 in at least one conversion paradigm and consistently assign lower perplexity to infelicitous sentences.

Chinese-trained models (DeepSeek and Zh-Pythia) consistently outperform English-centric models on the conversion paradigms, while the gap is smaller on paradigms solvable with lexical cues (e.g., the *le* marker and word order). On syntactic tests of *ba* and *bei*, models trained on both languages generally achieve accuracies above 0.85; however, on conversion tasks, DeepSeek models remain above 0.9, whereas English-centric models sometimes drop below 0.2. These patterns suggest that semantic compatibility constraints of the *ba* construction are not robustly encoded, and that Chinese exposure is especially important for acquiring higher-level constraints that depend on verb–construction compatibility beyond surface word-order templates marked by lexical cues.

The base Mistral model tends to yield lower absolute values than other models (Table 4), which aligns with Huang et al.'s (2025) observations for the Chinese verb-resultative construction for Mistral and Zh-Pythia. This may be due to different tokenization, with Mistral having a rather small vocabulary. We emphasize perplexity differences within the same model across paradigms.

5.1 Prompting or probability

Researchers have found that prompting is not always a reliable substitute for log-probability comparisons in minimal pair evaluation. Hu and Levy (2023) and Hu et al. (2024) show that prompt-based methods often underperform log-probability scoring for syntactic judgments, especially when models are uncertain or the prompt introduces irrelevant discourse context. In line with their results, we observe cases where prompt-based forced choice yields substantially lower accuracy than log-probability evaluation (especially, Mistral on *ba le* and *ba infel le*; see Table 3). One possible explanation is that prompting introduces additional pragmatic or instructional context that interferes with the model's underlying grammatical preferences.

However, in our experiments, there are also exceptions where prompting improves performance. For example, on the *ba vs. svo* task, LLaMA's accuracy improves from 0.1667 (base) to 0.7400 (instruction-tuned); Mistral rises from 0.53 to 0.75. Future controlled studies should explore how the minimal-pair paradigm interacts with prompt design to give rise to such observations.

5.2 The interaction of form, meaning and information structure

Zh-Pythia performs moderately on *ba vs. bei* but much worse on *ba vs. svo* (Table 3). This imbalance is also reflected in perplexity: for all Zh-Pythia sizes, the relative perplexity increase on *ba vs. svo* is negative (Appendix A), meaning that the infelicitous *ba* variant is preferred over the grammatical and information-structurally neutral SVO alternative consistently.

This raises several questions. First, is this imbalance due to both *ba* and *bei* being overt constructions, while SVO lacks surface marking? The models may exhibit a bias toward constructions with lexically fixed markers, assigning them lower perplexities regardless of grammaticality. As shown in Table 4, the SVO variant consistently receives the highest perplexity among the three sentence types, despite being the only grammatical one.

This suggests that the models may prioritize surface cues over deeper syntactic compatibility.

A further question is whether the model encodes verb–construction compatibility at all, or instead relies primarily on form familiarity as an acceptability heuristic. This calls into question whether a construction grammar-like representation is learned.

Overall, this result suggests that even Chinese monolingual decoder models may lack fine-grained constructional representations for distinguishing between semantically similar information structure constructions. Further studies could disentangle these factors by comparing information-structural constructions that lack surface marking with ones that have them, as well as comparing lexical-semantic constructions with lexically fixed elements compared to those without.

5.3 Verb–construction compatibility

Most models perform as well as or even better on the two "infelicitous *ba*" paradigms than on the corresponding natural paradigms, with many reaching near-ceiling accuracy (Table 3). This suggests that models preserve the form of the *ba* construction even when the sentence is semantically implausible. A related observation is reported by Li et al. (2022), who show that language models associate argument structure constructions (ASCs) with meaning even in semantically nonsensical sentences. Specifically, they find that verbs embedded in infelicitous sentences still cluster closer in embedding space to prototypical verbs associated with a given construction than to incongruent ones, although distances are generally larger than in natural sentences. One possible explanation is that when semantic constraints are poorly represented in the model, models fall back on associations with lexically fixed elements of a construction, as these are represented in the lowest layers with a (sub)token and do not require abstraction.

Together, these findings suggest that decoder language models can struggle to represent verb–construction compatibility in certain cases, as we show this for the case of information structure constructions where the choice of construction depends more on information-structural considerations than on constructional meaning. There also appears to be an issue in representing verb–construction compatibility of argument structure constructions more broadly.

Exactly this phenomenon played a role in debates on theoretical contributions of large language models. In an essay titled "Large language models are better than theoretical linguists at theoretical linguistics", Ambridge and Blything (2024), who are linguists, used the impressive performance of large language models on the task of predicting the acceptability of argument structure constructions with felicitous and infelicitous verbs as evidence for their claim. While this essay was criticized in various ways, none raised the observation that this "LLM theory" of linguistics does not yet extend far beyond English. Our results show that it currently struggles to extend even to a widely spoken language such as Mandarin Chinese.

6 Conclusion

We introduce a new minimal-pair test set comprising seven paradigms designed to probe language models' knowledge of Mandarin Chinese *ba* and *bei* information structure constructions. Our dataset focuses on form-meaning mappings by testing verb–construction compatibility through minimal pairs across *ba*/*bei*/SVO alternations. We additionally investigate models' responses to infelicitous minimal pairs to assess whether form–meaning mappings are overgeneralized.

Our results show that many models perform well on paradigms involving information structure constructions with local syntactic cues (e.g., aspect and negation), but struggle with conversion paradigms requiring deeper construction-level understanding. Chinese-trained models, particularly DeepSeek, consistently outperform English-centric models.

Overall, our findings highlight the challenges that language models face in capturing the syntax–semantics interface phenomena of the *ba* construction. It is challenging to cover many languages while also engaging with language-specific constructions, but in future work, more comprehensive evaluations are needed to assess how language models represent semantic restrictions on information structure constructions in other languages beyond Mandarin. Various other languages share similar information structure constructions with Mandarin, e.g., the topicalization function to highlight affectedness among *bei* constructions, the English passive or the Indonesian passive (Liu and Ambridge, 2021). More generally, research on LLM constructional representations is needed for a broader range of typologically distinct languages.

7 Limitations

We did not collect human annotations in this study because DeepSeek consistently achieved near-ceiling performance across all tasks, allowing it to serve as a reference point for evaluating other models. Additionally, our most challenging paradigm, the *ba*/*bei*/SVO conversion task, uses only a small set of verbs ("知道", "了解", "知晓") that are widely accepted as incompatible with *ba*. However, if this experiment were to be extended with a more diverse range of verbs and more complex constructions in sentences from natural language data, human grammaticality judgments would become important for validation.

Another limitation of our work is the restricted coverage of syntactic phenomena at the syntax–semantics interface, as well as the limited lexical diversity of the dataset. A more comprehensive investigation could broaden both the range of constructions and the diversity of verb classes examined, especially verbs that are incompatible with either *ba* or *bei*, in order to assess to what extent our findings of potential overgeneralization of constructional knowledge hold across semantic domains, verb lemma frequency classes, and other potentially relevant factors.

Lastly, while we investigated the form-meaning mappings of information structure constructions in LLMs and whether semantic restrictions were correctly applied, we did not investigate the models' competence at handling information-structural aspects of these constructions. This would involve testing whether LLMs replicate observed variation that depends on factors in the discourse context, e.g. assigning a higher probability to the fronting of discourse-new objects compared to discourse-given objects. Such experiments were already carried out for Japanese topicalization with GPT2-small (Fujihara et al., 2022), for Spanish and Portuguese topicalized infinitives with GPT-3.5 (Gerhalter, 2024), as well as for the English dative alternation with a range of 7B open models (Tur et al., 2025) and BabyLMs (Yao et al., 2025), though other factors besides information structure are more prominent here. The GPT2 Japanese results suggest a discrepancy with human preferences while the recent English work suggests alignment, though information-structural factors such as givenness are not explicitly investigated in the English work. It would be good to investigate this with recent models on another topic-prominent language such as Mandarin Chinese, or even in English.

8 Acknowledgements

We want to acknowledge the Young Scientists Fund of the National Social Science Fund (Grant No. 24CYY102).

References

Ben Ambridge and Liam Blything. 2024. Large language models are better than theoretical linguists at theoretical linguistics. *Theoretical Linguistics*, 50(1-2):33–48.

DeepSeek-AI: Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, and 67 others. 2024. Deepseek LLM: Scaling open-source language models with longtermism. *Preprint*, arXiv:2401.02954.

Claire Bonial, Taylor Pellegrin, Melissa Torgbi, and Harish Tayyar Madabushi. 2025. From form to function: A constructional NLI benchmark. In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 172–179, Düsseldorf, Germany. Association for Computational Linguistics.

Bastian Bunzeck, Daniel Duran, and Sina Zarrieß. 2025. Do construction distributions shape formal language learning in German BabyLMs? In *The SIGNLL Conference on Computational Natural Language Learning*.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Riki Fujihara, Tatsuki Kuribayashi, Kaori Abe, Ryoko Tokuhisa, and Kentaro Inui. 2022. Topicalization in language models: A case study on Japanese. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 851–862.

Katharina Gerhalter. 2024. How do DeepL and ChatGPT process information structure and pragmatics? an exploratory case study on topicalized infinitives in Spanish (and Portuguese). *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses*, 1(1).

Greta Gorzoni, Ludovica Pannitto, and Francesca Masini. 2026. 'Layer su Layer': Identifying and disambiguating the Italian NPN construction in BERT's

family. In *Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 203–220.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Paul J Hopper and Sandra A Thompson. 1980. Transitivity in grammar and discourse. *Language*, 56(2):251–299.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. 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, Online. 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.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

Xinyao Huang, Yue Pan, Stefan Hartmann, and Yang Yanning. 2025. Assessing minimal pairs of Chinese verb-resultative complement constructions: Insights from language models. In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 144–150, Düsseldorf, Germany. Association for Computational Linguistics.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026. MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Transactions of the Association for Computational Linguistics*, 14:193–216.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Jaakko Leino. 2013. Information structure. In *The Oxford Handbook of Construction Grammar*. Oxford University Press.

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural reality of argument structure constructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7410–7423, Dublin, Ireland. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Li Liu and Ben Ambridge. 2021. Balancing information-structure and semantic constraints on construction choice: building a computational model of passive and passive-like constructions in Mandarin Chinese. *Cognitive Linguistics*, 32(3):349–388.

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, and 1 others. 2026. A systematic assessment of language models with linguistic minimal pairs in Chinese.

Tom Mackintosh, Harish Tayyar Madabushi, and Claire Bonial. 2025. Evaluating CxG generalisation in LLMs via construction-based NLI fine tuning. In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 180–189, Düsseldorf, Germany. Association for Computational Linguistics.

Satoru Ozaki, Dan Yurovsky, and Lori Levin. 2022. How well do LSTM language models learn filler-gap dependencies? In *Proceedings of the Society for Computation in Linguistics 2022*, pages 76–88.

Julia Pestel, Jelke Bloem, and Raquel G Alhama. 2025. Evaluating Dutch speakers and large language models on Standard Dutch: a grammatical challenge set based on the Algemene Nederlandse Spraakkunst. *Computational Linguistics in the Netherlands Journal*, 14:555–582.

Steven Pinker, David S Lebeaux, and Loren Ann Frost. 1987. Productivity and constraints in the acquisition of the passive. *Cognition*, 26(3):195–267.

Pegah Ramezani, Achim Schilling, and Patrick Krauss. 2025. Analysis of argument structure constructions in the large language model BERT. *Frontiers in Artificial Intelligence*, 8:1477246.

Wesley Scivetti and Nathan Schneider. 2025. Construction identification and disambiguation using BERT: A case study of NPN. *Preprint*, arXiv:2503.18751.

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. Association for Computational Linguistics.

Brooke Stephenson, Laurent Besacier, Laurent Girin, and Thomas Hueber. 2022. BERT, can HE predict

contrastive focus? predicting and controlling prominence in neural TTS using a language model. In *Interspeech 2022-23rd Annual Conference of the International Speech Communication Association*, pages 3383–3387. ISCA.

Rint Sybesma. 1999. The ba-construction. In *The Mandarin VP*, pages 131–181. Springer.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ada Tur, Gaurav Kamath, and Siva Reddy. 2025. Language models largely exhibit human-like constituent ordering preferences. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2498–2521.

Tim Veenboer and Jelke Bloem. 2023. Using collostructional analysis to evaluate BERT's representation of linguistic constructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12937–12951.

Li Wang. 1943. 中国现代语法. The Commercial Press, Beijing. [Modern Chinese grammar]. Reprinted in the Chinese Grammar Series (1985 edition).

Li Wang. 1954. 中国语法理论. Zhonghua Shuju, Beijing. [Theory of Chinese grammar].

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Chengzhao Wu, Guanyi Chen, Fahime Same, and Tingting He. 2025. Analysing reference production of large language models. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 182–194.

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.

Suying Yang. 1995. Ba and bei constructions in Chinese. *Journal of the Chinese Language Teachers Association*, 30(3):1–36.

Qing Yao, Kanishka Misra, Leonie Weissweiler, and Kyle Mahowald. 2025. Both direct and indirect evidence contribute to dative alternation preferences in language models. In *Second Conference on Language Modeling*.

Bojiang Zhang. 2001. The symmetry and asymmetry in Bei and Ba constructions. *Zhongguo Yuwen*, (6):519–524.

Dexi Zhu. 1982. 语法讲义. The Commercial Press, Beijing. [Lectures on grammar].

A Relative perplexity differences

Model	**Size**	***ba***			***bei***	***ba*/*bei*/SVO**			**infelicitous *ba***		
		ba le	**ba neg**	**mean**	**bei neg**	**ba vs bei**	**ba vs svo**	**mean**	**ba infel le**	**ba infel neg**	**mean**
LLaMA	8B	2.07	0.46	1.27	0.29	0.00	-0.33	-0.17	2.03	0.47	1.25
Mistral	7B	0.78	0.14	0.46	0.16	0.13	0.03	0.08	0.68	0.28	0.48
DeepSeek	7B	4.73	0.68	2.71	3.61	5.06	3.37	4.22	0.65	1.00	0.83
ZH-Pythia	14M	2.39	0.38	1.39	0.24	0.01	-0.30	-0.15	1.16	0.25	0.71
ZH-Pythia	70M	2.83	0.31	1.57	1.06	0.30	-0.23	0.04	1.71	0.56	1.14
ZH-Pythia	160M	3.10	0.88	1.99	0.98	0.32	-0.17	0.08	2.49	1.08	1.79
ZH-Pythia	410M	2.62	0.71	1.67	1.24	0.41	-0.15	0.13	2.24	1.48	1.86
ZH-Pythia	1.4B	2.35	0.60	1.48	0.79	0.44	-0.15	0.15	2.21	1.53	1.87

Table 5: Relative perplexity increase from good (acceptable) to bad (unacceptable) sentences, computed as $(\text{PPL}_{ungram} - \text{PPL}_{gram})/\text{PPL}_{gram}$. LLaMA = LLaMA-3.1-8B; Mistral = Mistral-7B v0.3; DeepSeek = deepseek-llm-7b-base.