A Systematic Assessment of Language Models with Linguistic Minimal Pairs in Chinese

Yikang Liu[1*], Yeting Shen[1], Hongao Zhu[1,5], Lilong Xu[4], Zhiheng Qian[1], Siyuan Song[1,3], Kejia Zhang[1], Jialong Tang[2], Pei Zhang[2], Baosong Yang[2], Rui Wang[1#], Hai Hu[1,6#]

[1]Shanghai Jiao Tong University, China [2]Tongyi Lab, Alibaba Group, China
[3]The University of Texas at Austin, USA [4]University of Glasgow, UK
[5]University of California, San Diego, USA [6]City University of Hong Kong, China
Correspondence: yikangliu@sjtu.edu.cn; wangrui12@sjtu.edu.cn; hu.hai@outlook.com

Abstract

We present ZhoBLiMP, the largest linguistic minimal pair benchmark for Chinese, with over 100 paradigms, ranging from topicalization to the *Ba* construction. We then train from scratch a suite of Chinese language models (LMs) with different tokenizers, parameter sizes, and token volumes, to study the learning curves of LMs on Chinese. To mitigate the biases introduced by unequal lengths of the sentences in a minimal pair, we propose a new metric named sub-linear length normalized log-probabilities (SLLN-LP). Using SLLN-LP as the metric, our results show that ANAPHOR, QUANTIFIERS, and ELLIPSIS in Chinese are difficult for LMs even up to 32B parameters, and that SLLN-LP successfully mitigates biases in ZhoBLiMP, JBLiMP and BLiMP. We conclude that future evaluations should be more carefully designed to consider the intricate relations between linking functions, LMs, and targeted minimal pairs.

1 Introduction

Acceptability judgment is an important empirical method to measure human linguistic competence (Chomsky, 1965; Schütze, 2016), which has also been used to assess the linguistic knowledge of language models (LMs). Much work in this line adopted the minimal pair paradigm (MPP) in evaluating LMs (Linzen et al., 2016; Wilcox et al., 2018; Warstadt et al., 2020; Hu and Levy, 2023; Warstadt et al., 2023). A minimal pair is a pair of sentences with minimal contrast that affects whether the sentence is acceptable or not. A well-trained LM should assign higher probabilities to the acceptable sentences than the unacceptable ones (marked with *):

(1) a. The bureaucrat was bribed deliberately.

b. *The bureaucrat bribes deliberately.[1]

Well-curated and large-scale MPP benchmarks are some of the most widely used benchmarks for assessing LMs' linguistic competence due to their ease of use (Warstadt et al., 2023; AlKhamissi et al., 2025), and have also been widely used to study the mechanisms of language acquisition in LMs. For instance, using benchmarks such as the English BLiMP (Benchmark of Linguistic Minimal Pair, Warstadt et al., 2020), LMs have been found to acquire syntax with just around 100M tokens (Zhang et al., 2021; Warstadt et al., 2023); show similar acquisition order regardless of initialization, architecture, and training data (Choshen et al., 2022); and sometimes over-generalize with U-shaped learning patterns in which models only truly acquire a linguistic phenomenon after an initial dip in performance (Evanson et al., 2023; Haga et al., 2024).

These findings were mostly English-centric, highlighted by the use of the BLiMP. While recent endeavors in other languages—JBLiMP for Japanese (Someya and Oseki, 2023), BLiMP-NL for Dutch (Suijkerbuijk et al., 2025), among others—have facilitated research in more languages (see Table 1), there is still no systematic study of LMs' learning patterns in any non-English language.

For one of the most widely spoken languages, Chinese, there are two existing MPP benchmarks: CLiMP (Xiang et al., 2021) and SLING (Song

*Work done during an internship at Tongyi Lab.
#Corresponding authors.

[1]Taken from Sprouse et al. (2013, pp. 239).

Transactions of the Association for Computational Linguistics, vol. 14, pp. 755–771, 2026. https://doi.org/10.1162/TACL.a.648
Action Editor: Benoit Crabbé. Submission batch: 6/2025; Revision batch: 6/2025; Published 5/2026.
© 2026 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

Benchmark	Language	Size	N
BLiMP (Warstadt et al., 2020)	English	67k	67
SyntaxGym (Hu et al., 2020)	English	NA	39
CLiMP (Xiang et al., 2021)	Chinese	16k	16
SLING (Song et al., 2022)	Chinese	38k	38
JBLiMP (Someya and Oseki, 2023)	Japanese	331	39
LINDSEA (Leong et al., 2023)	Indonesian	380	38
	Tamil	200	20
RuBLiMP (Taktasheva et al., 2024)	Russian	45k	45
BLiMP-NL (Suijkerbuijk et al., 2025)	Dutch	8.4k	84
ZhoBLiMP (Ours)	Chinese	35k	118

Table 1: Comparison of MPP benchmarks for different languages. *Size* refers to the number of minimal pairs in total; *N* refers to the number of linguistic paradigms.

et al., 2022). However, both fall short of including enough linguistic phenomena, with CLiMP only covering 16 and SLING 38. Furthermore, CLiMP uses a lexicon translated from English, which has been noted for generating unnatural sentences (Song et al., 2022). Sentences in SLING are derived from the Penn Chinese Treebank (Xue et al., 2005), which limits its sentence structures to mainly the news domain and also makes it difficult to extend to new paradigms.

On the other hand, an implicit constraint for all BLiMP-style benchmarks is that the two sentences in a minimal pair should have equal lengths, for unbiased calculation of log-probabilities. Yet this is difficult to observe, since in linguistic research for human acceptability judgment it is common for one sentence to have more/fewer words, as shown in (1). Even if the two sentences had the same number of words, the sentences might be tokenized with different numbers of tokens due to subword tokenization. Ueda et al. (2024) argued to filter out pairs of unequal length, but in our opinion, it would be better to design metrics that normalize length-related biases to allow more flexibility in data curation.

The research gaps above are twofold: (1) the relative inadequacy in resources in Chinese, and (2) the challenges posed by length-related bias in evaluation. To fill these gaps, we present ZhoBLiMP, train the Zh-Pythia LM suite, and propose a new linking function to debias model evaluation.

ZhoBLiMP is the largest MPP benchmark for Chinese to date, with 118 paradigms[2] covering 15 linguistic phenomena, totaling 35k minimal pairs (see Table 2 for examples). Compared with CLiMP and SLING, we consult more Chinese linguistic literature to obtain a more comprehensive coverage, particularly phenomena frequently discussed in the field of Chinese syntax, such as classifiers, the *ba* and *bei* constructions. Besides, we loosen the ''equal length'' constraint, without controlling the length at all levels, including word, character, or token, which allows us to include phenomena such as *ellipsis* (§2).

We then train a suite of Transformer-based LMs of different sizes (14M-1.4B parameters) from scratch on Chinese text (10M-3B tokens), with three types of tokenization methods and a careful checkpointing strategy, to study the learning patterns and trajectories of LMs, as well as the effect of tokenization on model performance (§3).

With ZhoBLiMP and the LM suite, we quantify the biases caused by unequal lengths and propose a new linking function—sub-linear length normalized log-probabilities (SLLN-LP)—to mitigate the bias. Through extensive validation on Chinese and preliminary experimentation on English, Dutch and Japanese, we find that SLLN-LP can successfully mitigate the ''unequal length'' problem and improve general accuracy in ZhoBLiMP, BLiMP, and JBLiMP, potentially applicable to other languages as well (§4).

Using SLLN-LP to assess the trained LM suite and the Qwen2.5 series on ZhoBLiMP, we find that an LM of 160M parameters trained on 3B tokens can achieve ∼87% accuracy on par with much larger multilingual LLMs. Notably, there are three phenomena that are challenging for LMs even up to 14B parameters: ANAPHOR, ELLIPSIS, and QUANTIFIERS (§5). We discuss our findings in the context of previous BLiMP benchmarks and provide suggestions for future researchers in §6.

Minimal pairs in ZhoBLiMP, our codebase for data generation and the Zh-Pythia suite are available at `https://github.com/sjtu-compling/ZhoBLiMP`.

2 Creation of ZhoBLiMP

Bearing in mind the gaps in current benchmarks, we adopt the strategy of generation from templates and a vocabulary, to build a controlled and extendable Chinese benchmark of minimal pairs. To do so, we first build a GUI that can generate minimal pairs given grammar templates of a minimal pair and a vocabulary. Then, Chinese

[2] ''Paradigm'' refers to the patterns of *one* minimal pair. Thus, 118 paradigms means 118 such minimal pair patterns.

Phenomenon	N	Acceptable example	Unacceptable example
ANAPHOR	6	她的弟弟讨厌他自己。 *Her little brother hates himself.*	她的弟弟讨厌她自己。 *Her little brother hates herself.*
ARGUMENT STRUC.	7	我预习了教材。 *I previewed the textbook.*	我出现了教材。 *I appeared the textbook.*
BA	13	她把那条鱼放在池塘里。 *She BA that fish put in the pond.*	把那条鱼她放在池塘里。 *BA that fish she put in the pond.*
CLASSIFIER	3	那边站着八位舞者。 *Eight WEI dancers are standing there.*	那边站着八条舞者。 *Eight TIAO dancers are standing there.*
CNTL & RAISING	4	那杯红酒会变质。 *That glass of wine will go bad.*	会那杯红酒变质。 *Will that glass of wine go bad.*
ELLIPSIS	3	你们拉了小提琴，我们也拉了。 *You played the violin, we played too.*	你们笑了一天，我们也笑了。 *You laughed all day, we laughed too.*
FCI LICENSING	5	任何人都可以去。 *Anyone can go.*	任何人可以去。 *Anyone go.*
NOMINAL EXP.	11	他是司机。 *He is a driver.*	他司机。 *He driver.*
NPI LICENSING	9	没有任何人来了。 *Nobody came.*	任何人没有来了。 *Anyone didn't come.*
PASSIVE	12	他被小明打断了鼻子。 *His nose was hit-broken by Xiao Ming.*	他被小明打断了教材。 *His textbook was breathed hit-broken by Xiao Ming.*
QUANTIFIERS	2	没有人吃了超过九块糖果。 *No one ate more than nine candies.*	没有人吃了至少九块糖果。 *No one ate at least nine candies.*
QUESTION	21	你到底喝不喝啤酒？ *You DAODI will drink the beer or not?*	你难道喝不喝啤酒？ *You NANDAO will drink the beer or not?*
RELATIVIZATION	4	我知道赵大爷不想要你笑的原因。 *I know why Zhao doesn't want you to laugh.*	我知道赵大爷不想要谁笑的原因？ *I know why Zhao doesn't want who to laugh?*
TOPICALIZATION	4	我们没盖什么被子。 *We didn't have any quilt.*	我们什么被子没盖。 *We any quilt didn't have.*
VERB PHRASE	14	她没有吃过蛋糕。 *She hasn't eaten a cake.*	她没有吃了蛋糕。 *She hasn't ate a cake.*

Table 2: Overview of the 15 phenomena in ZhoBLiMP, with the number of paradigms (N), and one randomly sampled minimal pair. Each paradigm contains 300 minimal pairs. English translation for illustrative purposes, using a mixture of word-by-word gloss and translation to show the contrast.

linguists manually write the grammar templates for minimal pairs extracted from multiple sources and create a vocabulary with the necessary features. Finally, we hire Chinese native speakers to validate samples in ZhoBLiMP.

2.1 Minimal Pair Generation Platform

Features of the Platform. The platform provides users with a web interface to craft grammar templates with four types of rules listed below to generate minimal pairs (see Figure 1).

- **`Lexical`**: A set of key-value pairs assigning values to certain lexical properties. Lexical items will be searched accordingly in the vocabulary. E.g., `pos:NN` will randomly sample lexical items the part-of-speech of which is `NN`.
- **`Direct`**: A list of string expressions that can be directly used in the composition of sentences rather than searching the vocabulary. E.g., ''自己'' (*ziji*) will directly occupy the position.
- **`(mis)Matched`**: Similar to `Lexical`, but assigns (dis)agreement in one lexical property between two different positions. E.g., `pos:PN mPos:0 mPro:gender`[3] samples an item of PN that agrees in gender with the first item in the template.
- **`Phrase`**: A pre-defined phrase that supports recursion of various depths. `ReflV` yields a

[3]Stands for: POS is pronoun; matched position is 0; matched property is gender.

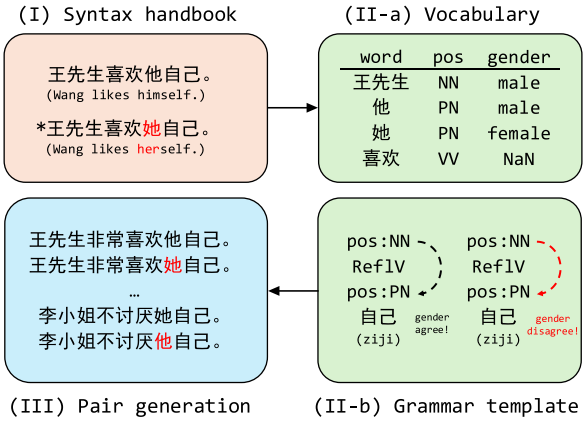

Figure 1: Data generation procedure illustration. Without writing any codes, linguists can easily generate sentence pairs by crafting grammar templates and vocabulary in the green blocks.

phrase such as ‘‘不喜欢’’ or ‘‘非常喜欢’’ (*don't like* or *very like*).

A grammar template consists of a good grammar and a bad one, both containing several rules. With this interface, users can easily generate minimal pairs at scale without any coding.

Vocabulary and Grammar Template. We design lexical properties that can be utilized in the lexicon filtering, and transform linguistic constraints into rule expressions according to lexical properties. As illustrated in the minimal pair:

(2) a. 王先生$_i$ 非常 喜欢 他自己$_i$。
Mr. Wang$_i$ very likes himself$_i$.

b. *王先生$_i$ 非常 喜欢 她自己$_i$。
Mr. Wang$_i$ very likes herself$_i$.

The ungrammaticality in (2) is caused by gender disagreement between 王先生 (Mr. Wang) and 她自己 (herself), which are co-indexed. We can use `(mis)Matched` to create the gender (dis)agreement between the reflexive (`pos:PN` + *ziji*, 她/他+自己) and its antecedent (`pos:NN`, 王先生), where the pronoun is a `Lexical` rule and the *ziji* (*self*) is a `Direct` rule. To complete the sentence structure, we use the phrase `ReflV`, which is pre-defined as a predicate that can take the reflexive pronoun as an object.

2.2 Data Generation for ZhoBLiMP

Sources of the Minimal Pairs. Templates of minimal pairs, which we call paradigms, come from three sources: (1) minimal pairs in a syntax textbook on Chinese—*The Syntax of Chinese* (Huang et al., 2009),[4] (2) BLiMP (Warstadt et al., 2020), and (3) journal articles on Chinese syntax and linguistics. We choose (1) because as a general-purpose textbook, it provides a comprehensive and systematic description of almost all syntactic phenomena in Chinese, with many example minimal pairs, which ensures ZhoBLiMP's coverage. For phenomena not covered in (1), we manually selected linguistics articles (3) that discuss these phenomena, and extract minimal pairs from them. We also select paradigms in the English BLiMP that have counterparts in Chinese for future comparison of the two languages.

Generation from Templates and a Vocab. A group of eight Chinese linguists first extract the minimal pairs from the three sources above. Then grammar rules for each sentence in the minimal pair are created, and the necessary lexical items with their specific features are added to the vocabulary, which are then used to generate 300 unique minimal pairs for each paradigm. We consult linguistic resources on specific phenomena to create a wide coverage vocabulary that can cover the 100+ paradigms we generate. The creation of templates and vocabulary took about two months.

In the end, 118 paradigms are created, which are grouped into 15 linguistic phenomena, with examples listed in Table 2.

2.3 Data Validation

We conduct human validation to verify the quality of the generated minimal pairs. We randomly sample 5 minimal pairs from each paradigm, and create a questionnaire asking native speakers which sentence from the minimal pair sounds more natural. All sampled minimal pairs were split into 10 lists, each containing roughly 65 pairs, plus two catch trials that a participant must answer correctly for her data to be valid. Fifty native speakers of Chinese are hired, each completing one list, leading to 5 responses per minimal pair. Participants are rewarded 10 Chinese RMB for the task. This validation shows that the participants have an overall agreement of 93.9% with our gold labels, with the agreement for each phenomenon listed in

[4]We use large-scale native speaker ratings of these minimal pairs reported in Chen et al. (2020) to select those pairs where judgments of linguists converge with those of native speakers.

# params	14M,70M,160M,410M,1.4B
# tokens	10M,100M,1B,3B

Table 3: Number of model parameters and training tokens of Zh-Pythia language models.

Table 6. This shows our benchmark aligns with the intuition of Chinese native speakers.

3 Training Zh-Pythia LM Suite

We train a series of LMs from scratch in a corpus mainly consisting of Chinese texts to investigate the acquisition of Chinese grammar with different configurations: number of model parameters, size of training corpora, and tokenizers. We name this suite of LMs as Zh-Pythia after Pythia (Biderman et al., 2023), borrowing their model architecture and configurations.[5]

For the training corpus, we collect 7000+ books that are primarily in the fields of humanities, including fiction and non-fictional topics on history, psychology, etc. The full corpus takes up 12GB in the txt format. We train LMs from scratch with different amounts of Chinese texts (see Table 3), resulting in 20 different combinations.

We use off-the-shelf Chinese-Llama (*cllama*, Cui et al., 2023) as our main tokenizer. Cui et al. (2023) create a Chinese vocabulary with 20k items and then merge it with the Llama tokenizer (Touvron et al., 2023). We train models with all configurations listed in Table 3 with *cllama*. In addition to that, we train two more tokenizers from scratch on our training corpus. One is a character-level tokenizer. The other is word-level, using characters as the base vocabulary and merging them into multi-character words with the BPE algorithm (Gage, 1994; Sennrich et al., 2016) (see Table 4). We additionally train two more LMs of 160M parameters on 3B tokens using the two tokenizers, respectively.

We set the global batch size at 256×64 tokens for all models and keep the hyperparameters the same. All LMs are trained on three random seeds. Training is done with two A100-80G GPUs.

4 Debiasing Length Normalization in Minimal Pairs

In this section, we first briefly review how existing linking functions are developed, and show that raw or mean-log probabilities of LMs will bias the evaluation when two sentences in a pair are different in length. We then introduce and validate a new function that addresses the unequal length problem by performing sub-linear length normalization.

[5]Models are released at `https://huggingface.co/collections/SJTU-CL/zh-pythia`.

tokenizer	# all	# single-char	# multi-char
cllama	49,953	10,876	6,963
word	20,276	11,116	7,471
char	12,142	11,116	0

Table 4: Tokenizers used for model training: *all* denotes the total size of the vocabulary; *single-char* and *multi-char* only include tokens that contain Chinese characters.

4.1 Baseline Linking Functions

Token probabilities assigned by LMs provide an unsupervised signal for acceptability judgment: Higher probabilities indicate greater acceptability. While LMs themselves are often evaluated on acceptability judgment tasks, their probabilities can also be used to predict human acceptability ratings. Linking functions bridge this gap between LM probabilities and human acceptability ratings, requiring validity on human Likert-scale ratings.

The most basic linking function is raw log-probability (LP). For a tokenized sequence x with $|x|$ tokens:

$$\text{LP}(x) = \sum_{i=1}^{|x|} \log P(x_i|x_{<i}).$$

Lau et al. (2017) demonstrate that normalizing for length (*mean LP*, MLP) and lexical unigram frequency (*syntactic log-odds ratio*, SLOR) correlates better with human ratings:

$$\text{MLP}(x) = \frac{\text{LP}(x)}{|x|},$$
$$\text{SLOR}(x) = \frac{\text{LP}(x) - \text{U}(x)}{|x|},$$

where $\text{U}(x) = \sum_i^{|x|} \log P_u(x_i)$ and $P_u(x_i)$ represent the token frequency in the training corpus. More recently, Lau et al. (2020) proposed Pen LP as an additional measurement under length normalization, based on Wu et al. (2016):

$$\text{PenLP}(x) = \frac{\text{LP}(x)}{((|x|+5)/(5+1))^{\alpha}},$$

	Zh-Pythia-160m-cllama			Zh-Pythia-160m-word			Zh-Pythia-160m-char				
Length	$\mathcal{D}_=$	$\mathcal{D}_>$	$\mathcal{D}_<$	$\mathcal{D}_=$	$\mathcal{D}_>$	$\mathcal{D}_<$	$\mathcal{D}_=$	$\mathcal{D}_>$	$\mathcal{D}_<$	$acc\uparrow$	$\Delta_{acc}\downarrow$
# pairs	20,584	5,456	9,360	22,206	5,802	7,392	23,347	6,198	5,855		
LP	88.20	60.77	95.62	84.13	52.85	98.73	**86.19**	56.55	94.28	83.44	19.74
mean LP	88.20	96.78	66.19	84.13	92.53	54.64	**86.19**	95.96	66.91	82.59	16.26
pen LP	88.20	83.47	87.09	84.13	81.63	87.63	**86.19**	82.97	80.98	85.46	**3.38**
SLOR*	90.01	88.09	79.60	**86.44**	87.65	74.70	84.08	92.74	70.36	84.83	7.94
MORCELA*	**90.12**	82.32	88.48	86.12	83.55	82.90	86.15	91.03	74.88	**86.22**	5.23
SLLN-LP											
$\alpha = 0.1$	88.20	66.07	93.66	84.13	60.09	97.79	**86.19**	61.85	91.78	84.04	15.87
$\alpha = 0.3$	88.20	74.58	91.06	84.13	71.51	94.00	**86.19**	71.45	87.77	84.95	9.21
$\alpha = 0.5$	88.20	81.60	87.55	84.13	81.18	87.66	**86.19**	79.85	82.93	85.31	<u>3.89</u>
$\alpha = 0.7$	88.20	89.11	80.90	84.13	87.79	77.60	**86.19**	87.72	76.46	84.87	4.94
$\alpha = 0.9$	88.20	95.06	71.66	84.13	91.24	62.82	**86.19**	94.13	69.96	83.54	12.67

Table 5: Accuracy (*acc*) and length-related bias (Δ_{acc}) on ZhoBLiMP using different acceptability linking functions. An appropriate function should have a higher *accuracy* and a lower Δ_{acc} across tokenizers. More optimal parameters of MORCELA are searched independently of each LM (*cllama*: $\beta = 0.6, \gamma = 18$; *word*: $\beta = 0.7, \gamma = 12$; *char*: $\beta = 0.1, \gamma = 12$).

where α usually takes the value of 0.8. Additionally, Tjuatja et al. (2025) find that frequency normalization requirements vary across LMs and propose MORCELA with adjustable normalization:

$$\text{MORCELA}(x) = \frac{\text{LP}(x) - \beta \cdot \text{U}(x) + \gamma}{|x|},$$

where tunable parameters β and γ further strengthen the correlation. While length normalization in these functions aims to improve prediction for independent sentences, their application may differ in MPP contexts.

4.2 Goal of Debiasing Length Normalization

Our goal here is to find a linking function f that can properly perform length normalization without over-penalizing longer or shorter sentences. We categorize a benchmark $\mathcal{D}$ consisting of minimal pairs (g, u), where g and u refer to grammatical and ungrammatical sentences, into three parts:

$$\mathcal{D}_< = (g, u) \in \mathcal{D} : |g| < |u|$$
$$\mathcal{D}_= = (g, u) \in \mathcal{D} : |g| = |u|$$
$$\mathcal{D}_> = (g, u) \in \mathcal{D} : |g| > |u|$$

where $\mathcal{D}_<$ consists of pairs, the grammatical sentence of which is shorter than the ungrammatical counterpart, while the grammatical sentences are longer than the ungrammatical ones in $\mathcal{D}_>$, and $\mathcal{D}_=$ is more standard minimal pairs of the same length. We then estimate the linguistic knowledge of LMs by the accuracy of correctly selecting the good sentence from a minimal pair using the linking function f given a dataset $\mathcal{D}$:

$$\text{acc}(\mathcal{D}; f) = \frac{1}{|\mathcal{D}|} \sum_{(g,u)\in\mathcal{D}} \mathbb{I}(f(g) > f(u)).$$

An ideal f will generate similar model predictions no matter how the length varies between two sentences in a pair. We take the accuracy on split $\mathcal{D}_=$ as reference, as these pairs are not affected by length normalization. If the linking function is robust to the length difference in a pair, then the performance on $\mathcal{D}_<$ and $\mathcal{D}_>$ should be close to the reference accuracy. Thus, we formalize the goal of debiasing length normalization as Δ_{acc}, and we want to find the f that minimizes the value:

$$\Delta_{acc} = \frac{1}{2} \sum_{\mathcal{D}\in\{\mathcal{D}_<,\mathcal{D}_>\}} |\text{acc}(\mathcal{D}; f) - \text{acc}(\mathcal{D}_=; f)|.$$

4.3 SLLN-LP: Sublinear Length Normalized Log-probabilities

We observe that LP favors shorter sentences without normalization, while MLP sometimes over-normalizes longer sentences. For instance, Zh-Pythia-160m achieves only 60.77% of accuracy on $\mathcal{D}_>$ but 95.62% on $\mathcal{D}_<$ when using LP. In contrast, the accuracy on $\mathcal{D}_>$ improves to 96.78% but the accuracy on $\mathcal{D}_<$ drops to 66.19% (see Table 5), while Pen LP has a more balanced performance. We observe that the degrees of length normalization might be the solution. Thus, we propose to use a sublinear function, in between LP and MLP, to normalize length:

$$\text{SLLN-LP}(x; \alpha) = \frac{\text{LP}(x)}{|x|^\alpha}, \text{where } \alpha \in (0, 1).$$

In SLLN-LP, the parameter α determines the degree of length normalization. Higher values of α yield more aggressive normalization. The boundary cases of $\alpha = 0$ and $\alpha = 1$ reduce SLLN-LP to LP and MLP, respectively. With this parameter, we aim to find how much we should perform length normalization to mitigate the noise brought by pairs of unequal lengths.

We exclude frequency normalization in SLLN-LP for two reasons: (1) Frequency should not significantly impact forced-choice since minimal pairs mostly contain identical words. (2) We aim to isolate length normalization from frequency normalization. Without frequency normalization, LM log-probabilities in the numerators remain negative. Frequency normalization can alter these signs, which changes the monotonicity of length normalization.[6]

4.4 Validating SLLN-LP Across Tokenizers

In this and the next subsections, we comprehensively analyze the impact of length biases in Chinese, English, Japanese, and Dutch in more controlled settings.

Different tokenization methods can result in varying sentence lengths. We first validate SLLN-LP's performance across different tokenizers. We conduct experiments using three models: (1) Zh-Pythia-160m-cllama, (2) Zh-Pythia-160m-word, and (3) Zh-Pythia-160m-char. These models are trained with identical token volumes, with the tokenizer serving as the sole variable. Using these differently tokenized models, we evaluate model performance acceptability judgments through various linking functions (LP, mean LP, pen LP, SLOR, MORCELA, and SLLN-LP) on ZhoBLiMP.

Improper length normalization leads to bias in MPP evaluation. As shown in Table 5, approximately 30–40% of pairs in ZhoBLiMP have unequal lengths, a proportion that remains consistent across different tokenizers. These unequal-length pairs are particularly sensitive to the choice of linking functions. As highlighted by the underlined cells in Table 5, improper normalization in linking functions can either underestimate or overestimate model performance, with accuracy varying significantly among $\mathcal{D}_{=}$, $\mathcal{D}_{+}$, and $\mathcal{D}_{-}$.

[6]Signs of over 90% of sentences are altered using SLOR.

LP tends to assign higher log-probabilities to shorter sentences regardless of their grammaticality, resulting in poor performance on $\mathcal{D}_{>}$ but inflated performance on $\mathcal{D}_{<}$. Conversely, mean LP over-normalizes for length in MPP, favoring longer sentences due to their larger denominators, which leads to higher acceptability ratings. These opposing biases explain the substantial Δ_{acc} observed in both LP and mean LP methods.

Frequency normalized functions can mitigate the bias, but not always. The two starred functions in Table 5 incorporate token frequency normalization. The results show that unigram frequency normalization substantially reduces Δ_{acc} to below 10 (specifically, 7.94 and 5.23) while improving overall average accuracy. For LMs trained with word-level tokenizers, this normalization also enhances accuracy on $\mathcal{D}_{=}$, where sentence pairs have equal lengths. However, LMs trained with character-level tokenizers still exhibit high length-difference bias and show accuracy drops on $\mathcal{D}_{=}$, producing results comparable to mean LP. These findings suggest that frequency normalization effects are model-dependent, as evidenced by the variation in optimal frequency control parameters (β) of MORCELA across different models.

SLLN-LP normalizes length more effectively on ZhoBLiMP. Our proposed SLLN-LP with $\alpha = 0.5$ and pen LP demonstrate better length normalization, achieving the Δ_{acc} of 3.89 and 3.38, respectively. The accuracy obtained using these two functions is just slightly lower than that of MORCELA, which has more tunable parameters.

As α in SLLN-LP increases from 0.1 to 0.9, we observe that Δ_{acc} initially decreases before increasing again. Correspondingly, *accuracy* improves as the bias is gradually mitigated through more optimal values of the controlling exponent α.

These results suggest that the length-related bias in ZhoBLiMP can be mitigated by adjusting the degree of length normalization. Pen LP and SLLN-LP are both sub-linear and achieve equivalent performance, but we will use SLLN-LP below for its simpler conceptualization.

4.5 Finding the Optimal α of SLLN-LP Across Scales and Languages

The scale of LMs can influence the optimal degree of normalization. Tjuatja et al. (2025) demonstrate

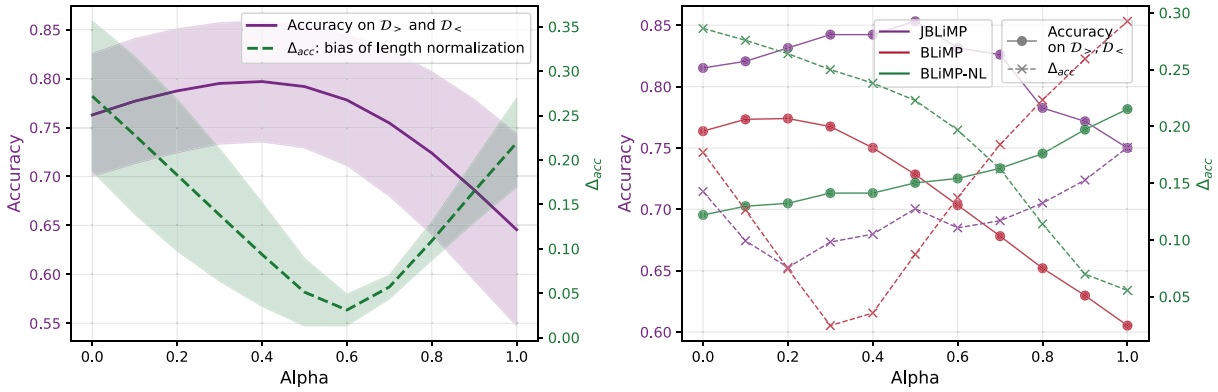

Figure 2: Effectiveness of debiasing length normalization using different α. We report two metrics: (1) average *accuracy* on $\mathcal{D}_+$ and $\mathcal{D}_-$ (when minimal pairs have unequal lengths), and (2) Δ_{acc}. *left* is the average results of 20 Zh-Pythia LMs with the shaded area denoting the standard deviation; *right* is the results of MPP benchmarks in other languages, including JBLiMP, BLiMP, and BLiMP-NL.

that the optimal degree of frequency normalization varies across scales. We investigate whether the length normalization varies across scales as well and examine SLLN-LP's applicability to languages other than Chinese. We analyze model predictions using LP, SLLN-LP, and mean LP (with $\alpha \in [0, 1]$), examining how *accuracy* and Δ_{acc} respond to different α values.

- *Scale analysis*: We evaluate 20 Zh-Pythia LMs, using *cllama* tokenizer, on ZhoBLiMP.
- *Cross-linguistic analysis*: We evaluate GPT2-small-Dutch (de Vries and Nissim, 2021) on BLiMP-NL (Suijkerbuijk et al., 2025), Japanese-GPT2-small (Sawada et al., 2024) on JBLiMP (Someya and Oseki, 2023), and Pythia-160m (Biderman et al., 2023) on BLiMP (Warstadt et al., 2020).

The left of Figure 2 shows consistent trends in both *accuracy* and Δ_{acc} across scales, with optimal α between 0.4 and 0.6, yielding the highest *accuracy* and the lowest Δ_{acc} on ZhoBLiMP.

Unequal lengths in minimal pairs are also common in benchmarks other than ZhoBLiMP: 55.59% of JBLiMP (Japanese) pairs, 14.37% of BLiMP (English) pairs, and 20.01% of BLiMP-NL (Dutch) pairs exhibit unequal lengths. The length normalization parameter α influences both *accuracy* and Δ_{acc}, though trends vary across languages (see the right plot of Figure 2). Japanese shows similarity to Chinese, with *accuracy* first increasing then decreasing, while Δ_{acc} follows an inverse pattern (yet the trend is not stable due to the comparatively small sample size of JBLiMP). For English, the optimal α falls between 0.2 and 0.4, while for Dutch it is $\alpha = 1.0$. The results suggest that SLLN-LP is applicable to JBLiMP and English as well, while mean LP provides better length normalization for BLiMP-NL.

Through comprehensive validation of SLLN-LP in Chinese across tokenizers and model scales, we observe that setting $\alpha = 0.5$ consistently yields robust performance and debiased length normalization. We hypothesize that this optimal degree of length normalization is primarily determined by data- or language-specific characteristics rather than model architecture or size, which seems to be different from the frequency normalization (Tjuatja et al., 2025).

Given the above observations, we will use SLLN-LP ($\alpha = 0.5$) to assess LMs on ZhoBLiMP for the following reasons: (1) SLLN-LP effectively mitigates the evaluation bias arising from length differences, demonstrating greater generalizability across tokenizers compared to SLOR and MORCELA. (2) SLLN-LP also maintains a stable optimal α value (0.5) across model scales. This consistency allows a fair comparison among the Zh-Pythia checkpoints using a uniform α.

5 Assessing LMs with ZhoBLiMP

With ZhoBLiMP, Zh-Pythia LMs, and debiased linking functions, we are ready to investigate how LMs acquire Chinese syntax. We present the assessment in three parts: (1) performance on ZhoBLiMP of five Zh-Pythia LMs trained on 3B tokens and two off-the-shelf LLMs, Qwen2.5-7B and Qwen2.5-14B (Qwen, 2024), (2) analysis on performance variation against training FLOPs,

	Zh-Pythia					Qwen2.5		Gap	Human
Phenomenon	14M	70M	160M	410M	1.4B	7B	14B		
BA	90.37	91.19	95.91	96.39	**96.65**	92.28	91.03	−0.50	96.15
ANAPHOR	46.15	40.72	44.28	56.67	63.20	59.22	**64.72**	20.87	85.60
ARG. STRUCTURE	76.71	85.27	87.08	88.44	**88.95**	88.38	87.00	6.97	95.92
CLASSIFIER	58.37	70.30	84.33	90.26	**94.11**	91.67	89.33	−0.06	94.05
CONTROL RAISING	86.67	92.67	97.36	97.86	**98.28**	91.25	94.92	−3.81	94.46
ELLIPSIS	32.81	41.78	45.33	47.89	49.37	**54.44**	54.11	37.70	92.14
FCI LICENSING	98.02	96.33	98.16	97.98	**98.29**	88.87	90.93	0.28	98.57
NOMINAL EXP.	91.77	96.21	96.40	**97.06**	96.39	92.27	93.76	−4.79	92.27
NPI LICENSING	59.65	68.53	76.52	81.32	**84.58**	73.78	74.89	8.75	93.33
PASSIVE	62.20	66.11	72.04	77.47	78.36	78.17	**79.81**	15.19	95.00
QUANTIFIERS	54.11	53.94	52.67	50.28	53.89	56.50	**68.33**	28.10	96.43
QUESTION	85.08	94.81	96.80	97.35	**97.62**	94.76	93.76	−0.14	97.48
RELATIVIZATION	96.81	99.50	98.50	99.33	**99.53**	97.92	97.50	−9.35	90.18
TOPICALIZATION	96.69	98.36	98.39	**99.17**	99.06	96.67	94.33	−1.67	97.50
VERB PHRASE	91.13	95.06	95.45	**95.52**	95.10	93.48	93.29	−1.64	93.88
$\text{OVERALL}_{\text{SLLN-LP}}$	79.12	83.94	87.01	89.10	**89.99**	86.68	87.18	4.62	94.61
$\text{OVERALL}_{\text{LP}}$	78.44	83.14	85.57	87.16	**88.40**	86.38	86.87	6.21	94.61
$\text{OVERALL}_{\text{MLP}}$	72.77	79.00	82.78	84.83	**85.38**	82.68	83.05	9.23	94.61

Table 6: Accuracy of Zh-pythia and Qwen2.5 models on ZhoBLiMP, broken down into different linguistic phenomena. Rows at the bottom compare overall results for different linking functions.

and (3) learning curves of Chinese grammar by examining hundreds of intermediate checkpoints.

5.1 Overall Performance

The results of the five Zh-Pythia LMs and two pre-trained Qwen2.5 LLMs are presented in Table 6. We first observe that with SLLN-LP, all models achieve higher accuracy, and therefore, the following analysis is based on SLLN-LP.

Model performance ranges from 79% to almost 90%. Even the smallest model, Zh-Pythia-14M, achieves 79.12% accuracy, approximately 30 percentage points above the chance level. Within the Zh-Pythia suite, performance consistently improves with increasing model size. Notably, Zh-Pythia-1.4B achieves the highest performance, outperforming the substantially larger Qwen2.5-7B and Qwen2.5-14B by 2-3 percentage points.

While most of the linguistic knowledge (covered in ZhoBLiMP) can be easily acquired, there are still linguistic phenomena that are challenging even for state-of-the-art LLMs like Qwen2.5. For ANAPHOR, ELLIPSIS, PASSIVE, and QUANTIFIERS, differences greater than 15 points between the best model and humans are observed. Interestingly, for these four phenomena, Qwen2.5 outperforms Zh-Pythia, suggesting that for more difficult phenomena, a larger training data size and model size are helpful.

5.2 Effect of Scaling

To analyze how scaling affects model performance across different linguistic phenomena in ZhoBLiMP, we plot performance against training FLOPs for the Zh-Pythia models (see Figure 3). Training FLOPs are calculated using the formula $C = 6ND$ (Hoffmann et al., 2022), where C represents training FLOPs, N denotes parameter count, and D indicates training token volume.

For models with 14-160M parameters, overall performance improvements plateau with increased training FLOPs. However, models with 410M and 1.4B parameters continue to show performance gains with additional training tokens, in many phenomena such as BA, ARGUMENT STRUCTURE, NPI, suggesting that larger Zh-Pythia models may not have reached their full potential and could benefit from extended training.

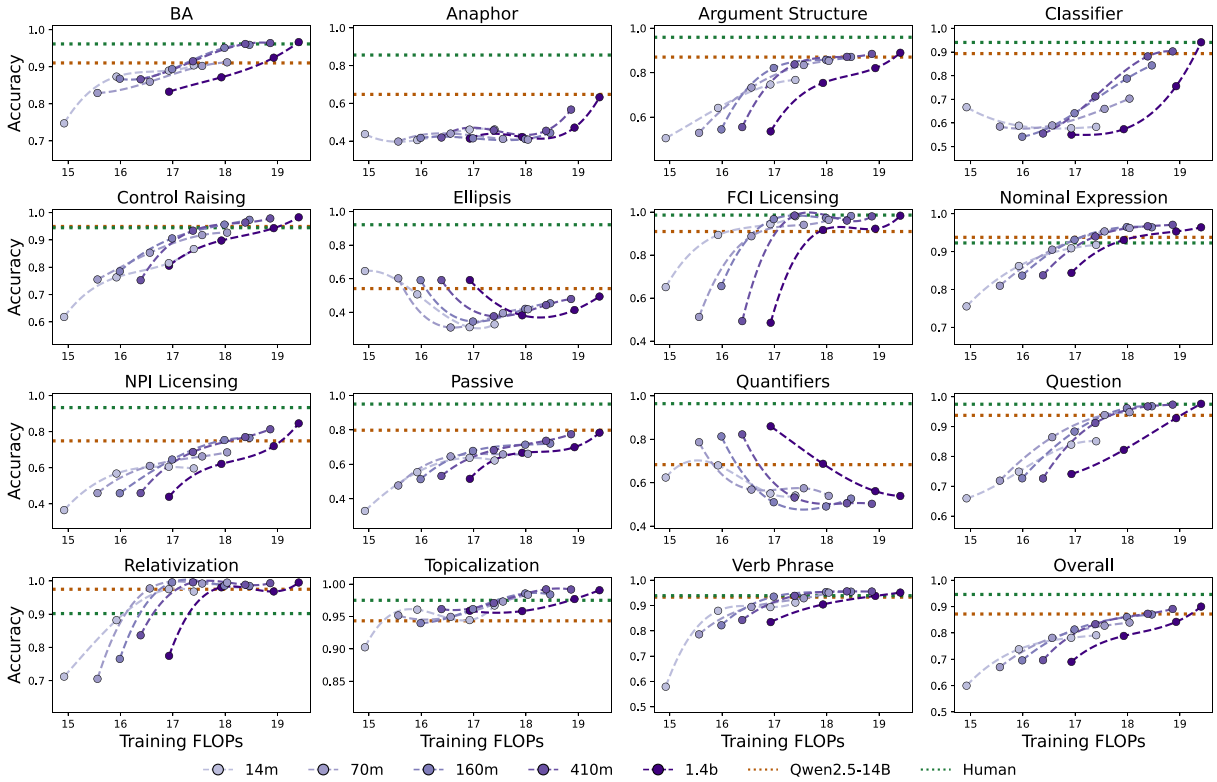

Figure 3: Phenomenon-specific accuracy on ZhoBLiMP plotted against training FLOPs (log scale). Each point represents a distinct Zh-Pythia LM, with models of identical parameter sizes shown in the same color. Points connected by dotted lines represent models of the same size, where higher training FLOPs indicate larger volumes of training tokens. We also plot the performance of Qwen2.5-14B and human as references.

Learning curve patterns vary across phenomena. Some phenomena, including FCI LICENSING, RELATIVIZATION, TOPICALIZATION, and VERB PHRASE, are acquired rapidly within 10^{16}-10^{17} FLOPs. Others, such as BA, CLASSIFIER, NOMINAL EXPRESSION, and QUESTION, show more gradual improvement, with performance plateauing around 10^{18} FLOPs.

The three challenging phenomena–ANAPHOR, ELLIPSIS, and QUANTIFIERS–exhibit different patterns. ANAPHOR remains near chance level until reaching 10^{18} FLOPs. ELLIPSIS shows initial performance degradation followed by improvement with increased scale. QUANTIFIERS, however, demonstrates consistent performance deterioration as model scale increases.

5.3 Learning Curves

To investigate whether LMs acquire syntax gradually or abruptly, we analyze 705 intermediate checkpoints (47 checkpoints $\times$ 5 parameter sizes $\times$ 3 seeds) from our Zh-Pythia models. Figure 4 displays the learning curves for each phenomenon. Most phenomena show minimal improvement during the first 100M training tokens, followed by a sharp performance increase. Performance typically saturates around 1B tokens, consistent across model scales and showing low variance.

Notably, U-shaped learning curves are observed across multiple phenomena, such as FCI LICENSING and TOPICALIZATION. These curves show initial performance decline around 100M tokens before subsequent improvement. While U-shaped learning has been documented in learning the past tense in English, where LMs might over-generalize the rules of inflection to irregular verbs (Rumelhart and McClelland, 1986; Plunkett and Marchman, 1991; Haga et al., 2024). Future research can investigate what causes the U-shaped learning in Chinese.

6 Discussion

This work makes two primary contributions: (1) advancing our understanding of Chinese syntax acquisition through the development of the ZhoBLiMP benchmark and Zh-Pythia LMs, and (2) identifying unequal length bias in minimal

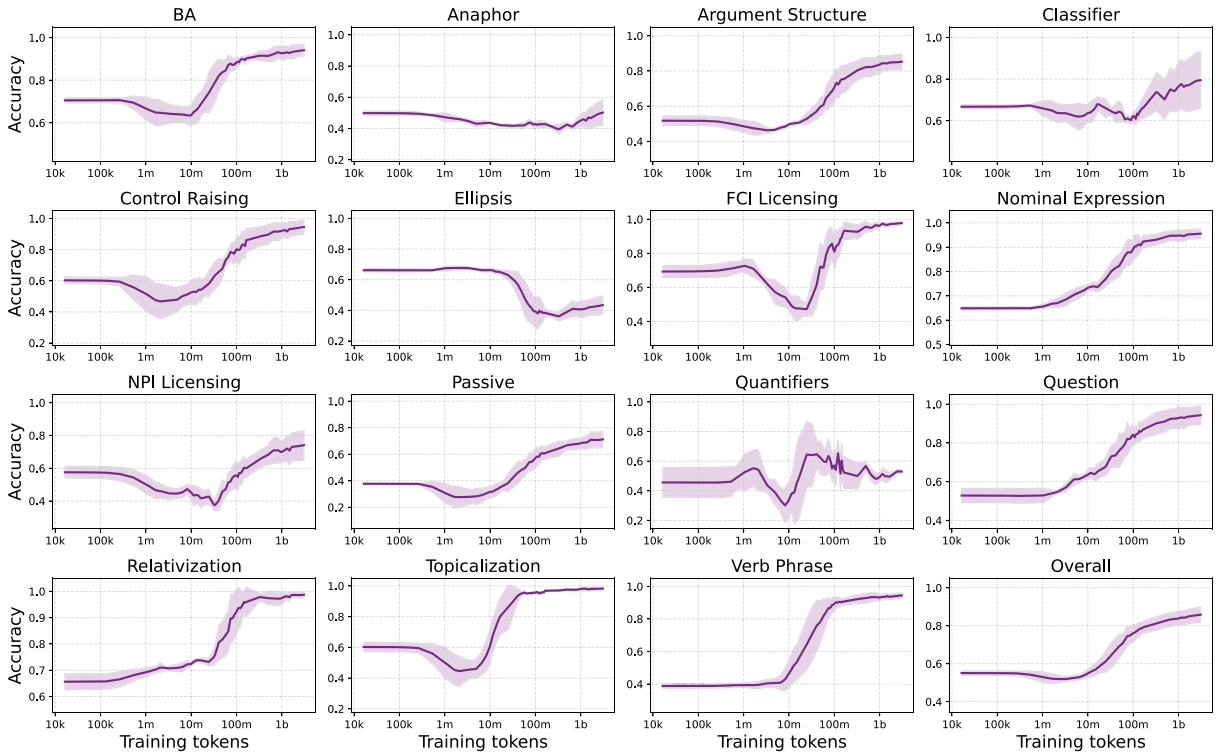

Figure 4: Learning curves across training tokens (calculated from training steps). We analyze 47 intermediate checkpoints from Zh-Pythia models trained on 3B tokens. The plot shows average accuracy across 15 LMs (5 model sizes $\times$ 3 seeds) with interpolation smoothing. The shaded area represents the standard deviation among checkpoints at equivalent training volumes.

pairs and introducing SLLN-LP as a mitigation strategy. We now summarize our findings and discuss their implications.

6.1 Acquisition of Chinese Syntax

Our assessment reveals that while Chinese syntax can be acquired by small LMs with limited training data, three phenomena remain particularly challenging: ANAPHOR, ELLIPSIS, and QUANTIFIERS. Performance on these phenomena lags approximately 20 percentage points behind human and exhibits distinct learning patterns across model scales and training steps. Unlike other phenomena, these three do not show consistent improvement with increased parameter size or training data in the Zh-Pythia suite (see Figures 3 and 4). However, the larger Qwen2.5 models achieve better (though still modest) accuracy (see Table 6), suggesting that these phenomena may require greater model capacity and training volume.

Comparing these results with those in English, we find that QUANTIFIERS are challenging in both languages, possibly due to the need for pragmatic and discourse information beyond syntax for quantifier understanding (Cohen and Krifka, 2014; Sperlich, 2019; Cremers et al., 2022).

On the other hand, ANAPHOR is especially challenging in Chinese but easy in English: Qwen2.5-14B achieves only 64.72% accuracy on Anaphor in ZhoBLiMP, whereas even GPT-2 reaches 99% on English BLiMP (Warstadt et al., 2020). The lower performance on Chinese anaphora is likely due to the linguistic property of the reflexive pronouns in Chinese: unlike English, Chinese has both 他自己/她自己 (ta-ziji, *himself/herself*) and 自己 (ziji, *self*), with the latter being more frequent. However, in making the minimal pairs, it is only feasible to use the former, as they are gender marked and can be used to control the acceptability of sentences. The status and nature of Chinese anaphora is still debated in theoretical linguistics (Xue et al., 1994; Yu, 2000; Zhu and Chai, 2025). We believe a more detailed and comprehensive comparison between the two types of reflexive pronouns is needed, along the lines of some recent work assessing LMs' interpretation of Chinese reflexive ziji (*self*) (Yang, 2025).

With these findings, we conjecture that the low performance and distinctive learning patterns in these three phenomena stem from their dependence on pragmatic and discourse information, as documented in the Chinese linguistics literature (Huang, 1994; Yuan, 1998; Wu and Tao, 2018; Chen and Hu, 2025), distinguishing them from purely syntactic phenomena.

6.2 Training Small LMs to Study Syntax Acquisition

Our Zh-Pythia-160M model, trained from scratch on just 1-3B tokens, achieves 87.01% accuracy, comparable to the much larger Qwen2.5-7B and Qwen2.5-14B. This demonstrates that relatively small LMs can effectively acquire syntactic knowledge, with further scaling yielding limited improvements–only about 3 points of difference between Zh-Pythia-160M and Zh-Pythia-1.4B (see Table 6).

Beyond similar end performance, smaller models exhibit learning patterns comparable to their larger counterparts across various training volumes (see Figure 3) and intermediate checkpoints (see Figure 4). This suggests that studying syntax acquisition patterns in smaller LMs could yield insights similar to those from much larger models.

Our findings align with previous research on English that the effect of scaling is limited for syntax acquisition (Hu et al., 2020; Zhang et al., 2021; Warstadt et al., 2023). While these studies used models of approximately 100M parameters trained on 100M English words, our Chinese syntax acquisition required an estimated 1-3B Chinese tokens. A key distinction lies in training duration: our models use single-epoch training, whereas English studies typically employ 20 or more epochs. This multiple-epoch approach's effectiveness is corroborated by Wilcox et al. (2025), who demonstrate that repeated exposure to the same 100M words improves BLiMP performance.

Despite these differences, both approaches remain within academic research budgets. We therefore encourage future research on other languages to employ small LMs ($\sim$100M parameters) with small training data (100M-1B tokens) when budgets are limited, as these models achieve adequate performance on MPP benchmarks while exhibiting informative learning patterns for analysis.

6.3 Minimal Pairs, LMs, and Linking Functions

We argue that when performing targeted syntactic evaluation, one should consider minimal pairs, LMs, and linking functions holistically. In previous work (Warstadt et al., 2020; Taktasheva et al., 2024; Suijkerbuijk et al., 2025), emphasis has been placed on the collection and human validation of minimal pairs. The validity and reliability of the linking functions used in evaluation are rarely discussed, as the same function is used uniformly across all models.

Our findings demonstrate that inappropriate linking functions can introduce evaluation bias. For example, using LP as the linking function leads to unbalanced performance in Zh-Pythia-160m across ZhoBLiMP splits ($\mathcal{D}=$: 88.2; $\mathcal{D}+$: 60.7; $\mathcal{D}_-$: 95.6) (see Table 5). This length-related bias extends beyond ZhoBLiMP to benchmarks in other languages, including BLiMP, JBLiMP and BLiMP-NL. Functions with more sophisticated normalization, such as SLLN-LP and MORCELA, can mitigate this bias on ZhoBLiMP, reducing Δ_{acc} to 3.89 and 5.23 while improving accuracy to 86.22 and 85.31. These functions show better handling of unequal-length pairs. However, function effectiveness varies across different LMs–SLOR and MORCELA perform well with *cllama* and *word* tokenizers but poorly with *char* tokenizer (see Table 5).

When developing MPP benchmarks, we therefore recommend validating both data quality and the effectiveness of linking functions. This validation should address two key aspects: (1) the compatibility between linking functions and minimal pairs, and (2) the compatibility between linking functions and the target language models.

7 Conclusion

In this paper, we introduced ZhoBLiMP, the most comprehensive Chinese MPP benchmark, featuring 35k minimal pairs across 15 linguistic phenomena. We trained the Zh-Pythia LM suite—22 models and 230 intermediate checkpoints varying in tokenizers, parameters, and training volumes—and proposed SLLN-LP, a novel linking function for length normalization in MPP. SLLN-LP improves accuracy and reduces ''unequal length'' bias compared to LP and MLP, while matching the performance of SLOR and MORCELA without requiring unigram frequency information.

Our findings reveal two key insights: (1) While small LMs readily acquire most Chinese syntactic phenomena, they struggle with ANAPHOR, ELLIPSIS, and QUANTIFIERS, probably due to their unique linguistic properties in Chinese or a reliance on discourse and pragmatic information; (2) Future benchmark development should validate linking functions alongside data quality to ensure fair model evaluation.

Limitations

One limitation is that we did not ask human annotators to perform a Likert-scale rating of the sentences, in addition to the force-choice human evaluation we did. Future work can add a Likert-scale rating experiment and compare results of the two task formats.

Despite our great effort to ensure the quality of the paradigms in ZhoBLiMP, a very small portion of the acceptable sentences may still sound unnatural to some speakers, likely due to the nature of templates-and-vocabulary-based generation in BLiMP-style corpora in general. For these rare cases, we ensured that the unacceptable sentences are (much) worse than the acceptable ones. As it is a forced-choice task, this will ensure the soundness of the results. Future work can explore other methods of minimal pair generation, such as employing an LLM, as demonstrated in Suijkerbuijk et al. (2025).

As pointed out by a reviewer, ZhoBLiMP also contains some semantically implausible sentences rather than strictly syntactic ill-formed sentences. Future research can try to tease them apart and model them independently.

Acknowledgments

We want to thank the anonymous reviewers and our action editor for their valuable comments and suggestions. This work is funded by a Humanities and Social Sciences grant from the Chinese Ministry of Education (no. 22YJC740020) awarded to Hai Hu, a Shanghai Pujiang Program grant (no. 22PJC063) awarded to Hai Hu, and a General Program of National Natural Science Foundation of China grant (no. 62176153) awarded to Rui Wang.

References

Badr AlKhamissi, Greta Tuckute, Yingtian Tang, Taha Osama A. Binhuraib, Antoine Bosselut, and Martin Schrimpf. 2025. From language to cognition: How LLMs outgrow the human language network. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24332–24350, Suzhou, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1237

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 2397–2430.

Yifan Chen and Xuhui Hu. 2025. Pragmatic ellipsis and its pragmatic consequences: Chinese *ye shi* at the syntax-pragmatics interface. *East Asian Pragmatics*, 9(3):441–462. https://doi.org/10.3138/eap.24006

Zhong Chen, Yuhang Xu, and Zhiguo Xie. 2020. Assessing introspective linguistic judgments quantitatively: The case of *The Syntax of Chinese*. *Journal of East Asian Linguistics*, 29(3):311–336. https://doi.org/10.1007/s10831-020-09210-y

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. The MIT Press, Cambridge, MA. https://doi.org/10.21236/AD0616323

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. The grammar-learning trajectories of neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8281–8297, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.568

Ariel Cohen and Manfred Krifka. 2014. Superlative quantifiers and meta-speech acts. *Linguistics and Philosophy*, 37:41–90. https://doi.org/10.1007/s10988-014-9144-x

Alexandre Cremers, Liz Coppock, Jakub Dotlačil, and Floris Roelofsen. 2022. Ignorance implicatures of modified numerals. *Linguistics and Philosophy*, 45:683–740. https://doi.org/10.1007/s10988-021-09336-9

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for Chinese LLaMA and Alpaca. *arXiv preprint arXiv:2304.08177v3*. https://doi.org/10.48550/arXiv.2304.08177

Wietse de Vries and Malvina Nissim. 2021. As good as new. How to successfully recycle English GPT-2 to make models for other languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 836–846, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.74

Linnea Evanson, Yair Lakretz, and Jean Rémi King. 2023. Language acquisition: Do children and language models follow similar learning stages? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12205–12218, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.773

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Akari Haga, Saku Sugawara, Akiyo Fukatsu, Miyu Oba, Hiroki Ouchi, Taro Watanabe, and Yohei Oseki. 2024. Modeling overregularization in children with small language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14532–14550, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.865

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556v1*. https://doi.org/10.48550/arXiv.2203.15556

Hai Hu, Aini Li, Yina Patterson, Jiahui Huang, and Chien-Jer Charles Lin. 2025. Bilingual influences and sources of variability in acceptability judgments: A case study of Chinese. *Lingua*, 318:103911. https://doi.org/10.1016/j.lingua.2025.103911

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. https://doi.org/10.18653/v1/2020.acl-main.158

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. https://doi.org/10.18653/v1/2023.emnlp-main.306

C.-T. James Huang, Y.-H. Audrey Li, and Li Yafei. 2009. *The Syntax of Chinese*. Cambridge University Press.

Yan Huang. 1994. *The Syntax and Pragmatics of Anaphora: A Study with Special Reference to Chinese*. Cambridge Studies in Linguistics. Cambridge University Press. https://doi.org/10.1017/CBO9780511554292

Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. How furiously can colorless green ideas sleep? Sentence acceptability in context. *Transactions of the Association for Computational Linguistics*, 8:296–310. https://doi.org/10.1162/tacl_a_00315

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. https://doi.org/10.1111/cogs.12414, PubMed: 27732744

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. BHASA: A holistic Southeast Asian linguistic and cultural evaluation suite for large language models. *arXiv preprint arXiv:2309.06085v2*. https://doi.org/10.48550/arXiv.2309.06085

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. https://doi.org/10.1162/tacl_a_00115

Kim Plunkett and Virginia Marchman. 1991. U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition*, 38(1):43–102. https://doi.org/10.1016/0010-0277(91)90022-V, PubMed: 2015756

Team Qwen. 2024. Qwen2.5: A party of foundation models.

David E. Rumelhart and James L. McClelland. 1986. On learning the past tenses of English verbs. In *Parallel Distributed Processing*, volume 2. MIT Press. https://doi.org/10.7551/mitpress/5236.003.0008

Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui, Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, and Koh Mitsuda. 2024. Release of pre-trained models for the Japanese language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13898–13905, Torino, Italia. ELRA and ICCL. https://doi.org/10.63317/5muru25vohoc

Carson T. Schütze. 2016. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Number 2 in Classics in Linguistics. Language Science Press, Berlin. https://doi.org/10.26530/OAPEN_603356

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1162

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-eacl.117

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. SLING: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.305

Darcy Sperlich. 2019. Syntactic and pragmatic theories of Chinese reflexives. *Lingua*, 221:22–36. https://doi.org/10.1016/j.lingua.2019.02.002

Jon Sprouse, Carson T. Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua*, 134:219–248. https://doi.org/10.1016/j.lingua.2013.07.002

Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. BLiMP-NL: A corpus of Dutch minimal pairs and acceptability judgments for language model evaluation. *Computational Linguistics*, pages 1–35. https://doi.org/10.1162/COLI_a_00559

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. RuBLiMP: Russian benchmark of linguistic minimal pairs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9268–9299, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.522

Lindia Tjuatja, Graham Neubig, Tal Linzen, and Sophie Hao. 2025. What goes into a LM acceptability judgment? Rethinking the impact of frequency and length. 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 2173–2186, Albuquerque, New Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.109

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971v1*. https://doi.org/10.48550/arXiv.2302.13971

Naoya Ueda, Masato Mita, Teruaki Oka, and Mamoru Komachi. 2024. Token-length bias in minimal-pair paradigm datasets. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16224–16236, Torino, Italia. ELRA and ICCL. https://doi.org/10.63317/38xk2sffxs5p

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.conll-babylm.1

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. https://doi.org/10.1162/tacl_a_00321

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5423

Ethan Gotlieb Wilcox, Michael Y. Hu, Aaron Mueller, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Adina Williams, Ryan Cotterell, and Tal Linzen. 2025. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, 144:104650. https://doi.org/10.1016/j.jml.2025.104650

Haiping Wu and Hongyin Tao. 2018. Expressing (inter)subjectivity with universal quantification: A pragmatic account of Plural NP + dou expressions in Mandarin Chinese. *Journal of Pragmatics*, 128:1–21. https://doi.org/10.1016/j.pragma.2018.02.003

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144v2*. https://doi.org/10.48550/arXiv.1609.08144

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, EACL 2021, Online, April 19 – 23, 2021*, pages 2784–2790. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.242

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238. https://doi.org/10.1017/S135132490400364X

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*.

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, Abu Dhabi, UAE. Association for Computational Linguistics.

William Xian-fu Yu. 2000. *Chinese Reflexives*. Peeters Publishers.

Boping Yuan. 1998. Interpretation of binding and orientation of the Chinese reflexive ziji by English and Japanese speakers. *Second Language Research*, 14(4):324–340. https://doi.org/10.1191/026765898670904111

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.90

Ruoxuan Zhu and Xingsan Chai. 2025. Who is ziji or ta-ziji? An ERP study on the processing mechanism of Chinese bare and compound reflexives. *Journal of Neurolinguistics*, 75:101267. https://doi.org/10.1016/j.jneuroling.2025.101267

A Evaluate Linking Functions Against Human Likert-scale Ratings

Having established SLLN-LP's effectiveness for length normalization on ZhoBLiMP and potentially other languages' minimal pairs, we evaluate its ability to predict human gradient acceptability ratings in Chinese. We utilize human ratings from Chen et al. (2020) and Hu et al. (2025), which differ from ZhoBLiMP in using expert-crafted minimal pairs from linguistic journals and textbooks. These studies collected Likert scale ratings for individual sentences, with acceptability measured as z-scored averages across multiple annotators.

We analyze correlations in two dimensions: (1) between absolute LM scores $f(x)$ and human ratings of x, and (2) between score differences and rating differences across grammatical-ungrammatical sentence pairs. Correlation strength (Pearson's r) indicates each function's predictive power (see Table 7).

	Pointwise rating (r)			Pairwise Δ (r)		
Func.	cllama	word	char	cllama	word	char
LP	32.44	25.58	24.84	44.61	35.72	36.63
MLP	41.79	28.58	31.12	39.51	30.77	33.96
SLOR	46.24	32.63	32.28	48.82	38.89	38.21
MCL	56.35	47.76	46.76	55.19	46.43	47.24
SLLN	46.69	38.68	39.05	50.94	42.97	43.97

Table 7: Correlation of human acceptability ratings and different linking functions from Zh-Pythia-160m, trained with different tokenizers. LP: log-probabilities; MLP: mean LP; MCL: MORCELA; SLLN: SLLN-LP ($\alpha = 0.5$). Human ratings are from Chen et al. (2020) and Hu et al. (2025).

SLLN-LP consistently ranks second in performance across all tokenizers and settings, behind only MORCELA. While MORCELA achieves higher correlations, SLLN-LP provides a simpler conceptualization that doesn't require frequency normalization.

B Performance on Excluded Paradigms

We exclude 11 paradigms for their low human agreement, which may be caused by poor data quality or ambiguous acceptability. We find that Zh-Pythia-160M has a lower accuracy when evaluated on these excluded paradigms. We will release these paradigms and their human annotations too, to facilitate research on human agreement and LM scores.

	included	excluded
# paradigms	118	11
Human agreement	93.9	59.6
LP	85.9	77.0
mean LP	83.7	70.7
SLOR	86.9	79.3
MORCELA	88.4	80.6
SLLN-LP	87.0	78.2

Table 8: Comparison between model performance (Zh-Pythia-160M) on included paradigms and excluded paradigms.