SLING: Sino LINGuistic Evaluation of Large Language Models

Yixiao Song[♢] **Kalpesh Krishna**[♠] **Rajesh Bhatt**[♢] **Mohit Iyyer**[♠]

[♢]Department of Linguistics, UMass Amherst
[♠]Manning College of Information and Computer Sciences, UMass Amherst
{yixiaosong,bhatt}@umass.edu
{kalpesh,miyyer}@cs.umass.edu

Abstract

To understand what kinds of linguistic knowledge are encoded by pretrained Chinese language models (LMs), we introduce the benchmark of Sino LINGuistics (SLING), which consists of 38K minimal sentence pairs in Mandarin Chinese grouped into 9 high-level linguistic phenomena. Each pair demonstrates the acceptability contrast of a specific syntactic or semantic phenomenon (e.g., The keys *are* lost *vs.* The keys *is* lost), and an LM should assign lower perplexity to the acceptable sentence. In contrast to the CLiMP dataset (Xiang et al., 2021), which also contains Chinese minimal pairs and was created by translating the vocabulary of the English BLiMP dataset, the minimal pairs in SLING are derived primarily by applying syntactic and lexical transformations to naturally-occurring, linguist-annotated sentences from the Chinese Treebank 9.0, thus addressing severe issues in CLiMP's data generation process. We test 18 publicly available pretrained monolingual (e.g., `BERT-base-zh`, `CPM`) and multi-lingual (e.g., `mT5`, `XLM`) language models on SLING. Our experiments show that the average accuracy for LMs is far below human performance (69.7% vs. 97.1%), while `BERT-base-zh` achieves the highest accuracy (84.8%) of all tested LMs, even much larger ones. Additionally, we find that most LMs have a strong gender and number (singular/plural) bias, and they perform better on local phenomena than hierarchical ones.[1]

1 Introduction

While large-scale pretrained language models (LMs) have achieved considerable downstream success (Devlin et al., 2019; Xue et al., 2021; Brown et al., 2020, a.o.), it remains challenging to evaluate how much linguistic knowledge they have acquired. One approach is to design *minimal pairs* consisting of two sentences that differ only in a critical word or phrase, which renders only one of the sentences acceptable (e.g., *The keys are lost* vs. *The keys is lost*). If an LM is sensitive to the phenomenon exemplified by the minimal pair (in this case, plurality), it should assign a lower perplexity to the acceptable sentence. This methodology can be used to test an LM's understanding of a wide range of linguistic phenomena; for example, the BLiMP dataset (Warstadt et al., 2020) contains 67K minimal pairs automatically generated via manually-constructed grammars that span 12 high-level English phenomena.

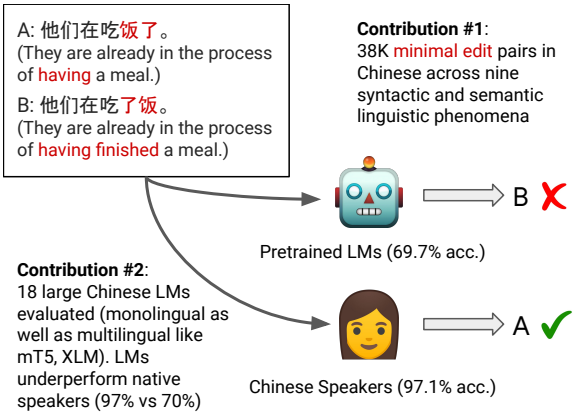

Figure 1: An illustration of the SLING dataset. The *A* sentence is acceptable but *B*, a minimal edit counterpart of *A*, is not. LMs see one sentence at a time and are expected to assign a lower (pseudo-)perplexity to the acceptable sentence. Overall, LMs underperform Chinese native speakers on SLING (97% vs 70%), making it an exciting benchmark for future Chinese LM research.

Can we create similar datasets to study linguistic phenomena in a different language, such as Chinese? As a first step in this direction, Xiang et al. (2021) introduce CLiMP, a Chinese dataset of minimal pairs. However, we identify two major issues with CLiMP's construction process: (1) its vocabulary is translated from BLiMP's vocabulary, which due to morphological differences between English and Chinese (e.g., the latter lacks numeral or verbal inflections) results in a large number of unin-

[1]The SLING data and code can be found https://github.com/Yixiao-Song/SLING_Data_Code.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4606 - 4634
December 7-11, 2022

telligible sentences; and (2) the grammatical templates for several phenomena (anaphor agreement, classifier-noun agreement, and filler-gap dependencies) are inadequately designed, which along with the vocabulary issue results in minimal pairs that do not have any clear contrast.[2]

To address these issues, we introduce SLING (Sino LINGuistics benchmark), a dataset of 38K minimal pairs to study nine Chinese linguistic phenomena, many of which are unique to the Chinese language. Instead of translating BLiMP, we construct SLING primarily using the Chinese Treebank 9.0 (Xue et al., 2016), which was annotated by trained linguists (see Table 1 for a comparison). We extract subtrees from human-validated constituency parses in this treebank and then carefully edit them using manually-designed linguistic templates to create minimal pairs. SLING does not suffer from the issues we found in CLiMP, and it additionally includes semantic as well as syntactic phenomena, seven of which are not found in CLiMP. A human validation of SLING with 16 native speakers confirms that its minimal pairs unambiguously show the acceptability contrast across all phenomena, yielding an *almost perfect* inter-annotator agreement (Fleiss' $\kappa = 0.88$).

We evaluate a total of 18 publicly-available pretrained LMs on SLING, including monolingual Chinese (e.g., `bert-base-chinese`, `PanGu-`α) and multilingual models (e.g., `mT5`, `XLM-R`). Our results reveal that: (1) no LM consistently outperforms others on SLING; (2) larger LMs do not necessarily outperform smaller ones; (3) monolingual Chinese LMs generally perform better than multilingual ones; and (4) humans significantly outperform all LMs (97.1% vs 69.7% average across LMs). We observe that the ranking of models on CLiMP differs from that on SLING: for example, `bert-chinese-base` is the best-performing model on SLING (average accuracy 84.8%), while `chinese-pert-base` performs best on CLiMP (81.2%). This result is due in part to the issues in CLiMP's construction process, as well as the different phenomena that we test in SLING. Additionally, SLING is more discriminative than CLiMP (i.e., LMs vary more across the phenomena in terms of accuracy), which makes it more useful as a diagnostic benchmark especially given the large gap between human and model performance.

[2]Note that although Xiang et al. (2021) report a high human accuracy of 97.1% on CLiMP, this number is calculated using majority vote of 16 annotators, and the inter-annotator agreement is not reported.

	CLiMP	SLING
vocab. source	BLiMP's vocab. translated	Chinese Treebank 9.0
vocab. size	actual 1272 types (w/ 230 proper names) (claimed 3456)	11988 types
grammar	9 syntax phenomena (16 paradigms)	3 semantics + 6 syntax (5 syntax differ from CLiMP) (38 paradigms)
evaluated LMs	monolingual only 1 bert-base-chinese 3 LSTM 2 5-gram	10 mono- & 8 multilingual 1 LSTM 3 Causal LMs 14 Masked LMs

Table 1: An comparison between CLiMP (Xiang et al., 2021) and SLING. SLING is created with a natural and diverse vocabulary, covers new semantic and syntactic Chinese linguistic phenomena, and is evaluated on large pretrained LMs, including multilingual models like mT5.[3]

2 Evaluating Chinese LMs with Minimal Pairs: CLiMP and Its Shortcomings

Using minimal pairs to detect a function of a single element (e.g., phoneme, affix, or word) is a common practice in linguistics. In Figure 1, by changing the position of 了, sentence *A* is transformed into the ungrammatical sentence *B*, and we know how the two aspect markers 在 and 了 interacts. In this paper, following BLiMP and CLiMP, we call each major grammatical category a *phenomenon*, and minimal pair types within each phenomenon *paradigms*. The *A* and *B* sentences in Figure 1 form a minimal pair of a paradigm in the *aspect* phenomenon of SLING.[4]

Xiang et al. (2021) created CLiMP to evaluate 9 Chinese syntactic phenomena with 16 paradigms. However, the dataset suffers from two major issues: (1) faulty minimal pair generation templates and (2) its translated vocabulary. In this section, we discuss the issues in detail and show why they hamper CLiMP's utility as a diagnostic dataset for LMs.

CLiMP's minimal pairs often do not show the desired acceptability contrast. This problem is especially prominent in the *ba* construction, binding/anaphor, and filler-gap dependency phenomena, on which Xiang et al. (2021) conclude that LMs perform poorly. The templates used to generate data for these phenomena are the primary cause of these errors, as we show below.

[4]More examples of minimal pairs can be found in Appendix D.

***ba* construction:** Many minimal pairs associated with this construction do not exhibit the acceptability contrast.[5] We examine the first 50 minimal pairs of this phenomenon in CLiMP and discover that 6 pairs actually have the wrong acceptability label:

Sentences	CLiMP	Actual
报告把大学转移了。 The report relocated the university.	✓	✗
报告被大学转移了。 The report was relocated by the university.	✗	✓

at least 9 minimal pairs contain two acceptable sentences:

Sentences	CLiMP	Actual
吴宇涛把图书馆调查了。 Wu investigated the library.	✓	✓
吴宇涛被图书馆调查了。 Wu was investigated by the library.	✗	✓

and 4 pairs are unintelligible or nonsensical:

Sentences	CLiMP	Actual
王萍把嘴举了 Wang lifted a mouth.	✓	✗
王萍被嘴举了 Wang was lifted by a mouth.	✗	✗

The primary reason for the low quality of these pairs is that CLiMP does not carefully control the source of unacceptability (Abrusán, 2019), which we discuss further in the Limitations section. Specific to the *ba* construction, CLiMP does not include essential information about thematic relations[6] in the vocabulary. Another contributing factor is the small size of the CLiMP vocabulary, which is translated from that of BLiMP despite many annotated features of BLiMP not applying to Chinese (e.g., number features, verb forms, or cases). For example, the English verb *buy* has six forms in BLiMP, listed in Table 2, which differ from each other in seven verb-related features. These inflections are useful in English for distinguishing sentence acceptability in several BLiMP phenomena (e.g., *Passive*, *Irregular Forms*, and *Subject-Verb Agreement*); however, they do not apply to Chinese because the language lacks inflection, and thus they cannot help construct Chinese paradigms. In Chinese, the same forms can be represented and built based on the three words shown in bold: *mai* (buy), *(zheng) zai* (progressive marker), and *le* (perfective marker). They do not need to be redundantly listed in the vocabulary. After removing the redundant word types, CLiMP's vocabulary size is 1,272 (including 230 proper names), not 3,456 as Xiang et al. (2021) report. This lack of diversity in the vocabulary contributes to the generation of nonsensical sentences using their minimal pair templates.

Chinese	English	Features
mai	buy	bare
zheng zai mai	buying	ing
mai **le**	bought	finite, past
mai le	bought	en
mai	buy	finite, pres
mai	buys	finite, pres, 3sg

Table 2: An example of the repetitive word types in CLiMP's vocabulary (*mai* here). ing = progressive, en = participle, pres = present, 3sg = third person singular.

Binding and anaphor paradigms: These two paradigms test whether the gender feature of the object anaphor agrees with that of the subject. Issues in the binding and anaphor paradigms stem from the fact that CLiMP uses proper names, which were added to CLiMP's vocabulary in addition to the one translated from BLiMP. However, Chinese proper names do not always unambiguously show gender. If the gender of the subject is ambiguous as in (1) where Ye Zi can be either gender (similarly for *Alex* in English), the performance of the LMs is not representative of whether they know the function of the reflexive anaphor, which is exactly what the binding and anaphor paradigms want to test.

(1) 叶梓逃离了他/ 她自己。
Ye Zi escaped from him- / herself.

Other issues with these two paradigms are discussed in detail in Appendix D.2.

Filler-gap paradigm: To create minimal pairs for the filler-gap paradigm in CLiMP, Xiang et al. (2021) use what they call the topicalization construction. However, (2a), taken from CLiMP, does not contain a filler-gap topicalization dependency. A real topicalization filler-gap structure should be the one in (2b), in which the direct object of the verb *buy* is topicalized and moved to the beginning of the sentence, leaving a *(gap)* at its base generated position (Huang et al., 2009, Section 6.1). Unfortunately, the minimal pairs associated

[5]The *ba* construction is a way to move the object from its base position (after a verb) to the position before the verb. The construction expresses the meaning of *settlement* and focuses on what is happening to the object.

[6]A thematic relation represents the semantic relation that a noun phrase bears with respect to an event denoted by a verb. For example, the thematic relation that *John* holds to the verb *eat* in *John eats an apple.*is that of agent, which means *John* is the agent of an apple eating event.

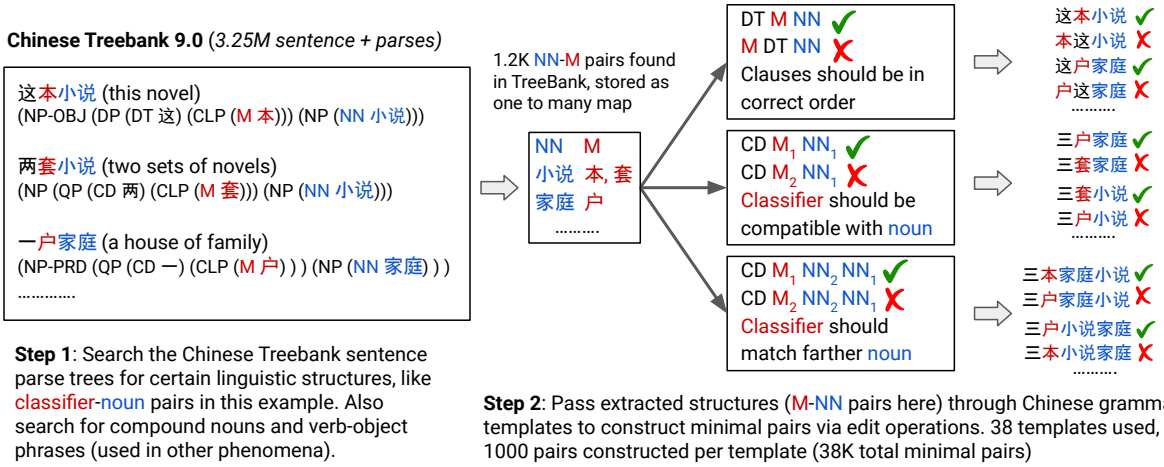

Figure 2: An illustration of the minimal pair generation process used to construct SLING.

with this paradigm are generated based on an erroneous template, which means no conclusions can be drawn from model performance on it.

(2) a. 门，我买了这东西。
Door, I bought this thing.
b. 门，我买了*(gap)*。
Door, I bought *(gap)*.

3 Creating the SLING Benchmark

This section describes our process of generating minimal pairs for SLING. We make use of the Chinese Treebank 9.0 (Xue et al., 2016), a Chinese corpus with linguist-annotated constituency parses that contains 2,084,387 words. This treebank allows us to use naturally-occurring sentences to construct our minimal pairs, unlike the synthetic and sometimes nonsensical sentences of CLiMP. Also, unlike CLiMP, whose linguistic templates rely solely on one grammar book (Po-Ching and Rimmington, 2015), our linguistic templates are constructed by a native Chinese linguist (the first author of this paper) based on multiple works in linguistics. Details of the construction of each phenomenon and the cited works can be found in Appendix D. The general minimal pair generation process is to identify a linguistic pattern, search for relevant linguistic structures in the Treebank, and form minimal pairs by applying hand-crafted transformation rules on the extracted structures. Figure 2 provides an overview of this process, with the same running example as this section.

3.1 Corpus: Chinese Treebank 9.0

Chinese Treebank 9.0 is a corpus of parsed text (3,247,331 Chinese and foreign characters) from various resources, both formal and colloquial. The Treebank contains 132,080 sentences; we extract a subset of these sentences that contains linguistic structures of interest and then manipulate those sentences to create minimal pairs for SLING.

3.2 Pattern Search

The most important patterns and corresponding strings extracted from the Treebank are classifier-noun phrases, compound noun phrases, and verb-object phrases. To demonstrate the extraction process, we will use classifier-noun phrases as an example. We extract classifier-noun phrases by searching for subtrees that have `NP` as their root node and contain a classifier `M`, for example, (3).

```
(3)  (NP-OBJ (DP (CD 两)
                 (CLP (M 套)))
             (NP (NN 小说)))
```

For each sub-tree, a classifier-noun pair is extracted as shown in Figure 2. Because each noun may have multiple compatible classifiers, a dictionary is created with the nouns as keys and the compatible classifiers as the values. Compound noun phrases and verb-object phrases are extracted in a similar way but stored as sub-trees only.

3.3 Sentence Generation

Minimal pairs are generated based on linguistic templates and the extracted strings. Using the classifier-noun agreement phenomenon as an example, the template is `CD M Noun`. For the acceptable phrases, the `M` is taken from the classifiers that are compatible with the noun in the dictionary. For the unacceptable phrases, `M` is randomly chosen from a classifier list (after making sure it is not in the list of compatible classifiers).

Phenomenon	Acceptable Example	Unacceptable Example	Syn	Sem	Distractor	Distance	Hierarchy
Alternative Question	tamen shi laoshi haishi mujiang? they are teacher or carpenter "Are they teachers or carpenters?"	tamen shi laoshi haishi mujiang **ma**? they are teacher or carpenter SP	✓				
Anaphor (Gender)	nan dianyuan kanjianle **ta**(他)-ziji. male shop assistant saw himself "The male shop assistant saw himself."	nan dianyuan kanjianle **ta**(她)-ziji. male shop assistant saw herself	✓		✓		✓
Anaphor (Number)	nan dianyuan **men** kanjianle **tamen**-ziji. male shop assistant PL saw themselves "The male shop assistants saw themselves."	nan dianyuan men kanjianle **ta**-ziji. male shop assistant PL saw himself	✓		✓		✓
Aspect	ta **qu** nian zhiding zhengce le. he last year establish policy AS "He established policies last year."	ta **ming** nian zhiding zhengce le. he next year establish policy AS	✓				✓
Classifier-Noun	yi **ming** tielu jingcha one M railway policeman "a railway policeman"	yi **tiao** tielu jingcha one M railway policeman (**tiao** is a wrong classifier for *policeman*)		✓	✓	✓	✓
Definiteness Effect	zheli/nali you **yi** jia yingyuan. here/there exist one M cinema "Here/there exists a cinema."	zheli/nali you **zhe/na/mei** jia yingyuan. here/there exist DT/DT/every M cinema		✓			
Polarity Item	ta **bu** fazhan renhe youhao guanxi. she not develop any friendly relations "She does not develop any friendly relations."	ta fazhan renhe youhao guanxi. she develop any friendly relations		✓			
Relative Clause	ta jianle na ge zhizhile baoli de nü jingcha. she saw DT M stoped crime DEC female police "She saw the female police officer who stopped the crime."	ta jianle na ge ta zhizhile baoli de nü jingcha. she saw DT M she stoped crime DEC female police	✓				✓
Wh-fronting	tamen shang ge yue daodi goujie le **shenme**? they last M month on earth collude with AS what "What on earth did they collude with last month?"	**shenme** tamen shang ge yue daodi goujie le? what they last M month on earth collude with AS	✓				

Table 3: An overview of the phenomena present in SLING along with their properties. The table indicates whether the paradigms within each phenomena represent syntactic (syn) or semantic (sem) knowledge, whether they involve a distractor (e.g., the ***roses*** in the vase ***are***/*is ...), whether there are long distance dependencies (e.g., ***these*** beautiful red blooming ***roses***), and whether the LMs need hierarchical knowledge of the language (e.g., Figure 3) to distinguish acceptable sentences from unacceptable ones. Details of each phenomenon are given in Appendix D.

In addition to phrases extracted from the Treebank, we also extract the transitive verbs[7] used in CLiMP's anaphor and binding phenomena,[8] and for certain phenomena we also utilize word lists (e.g., locations, pronouns, and occupations) to build the minimal pairs. Finally, for each paradigm in SLING, we generate one thousand minimal pairs.

3.4 Phenomena

As summarized in Table 3, SLING includes 9 major Chinese linguistic phenomena in syntax and semantics. Several minimal pair paradigms are designed to test an LM's robustness to distance and distractors in a dependency relation as well as whether they have the essential linguistic knowledge of hierarchy in Chinese; more details are provided in Appendix D. Here we describe the gist of each phenomenon. The **alternative question** phenomenon tests the knowledge that the disjunctor *haishi* and the polar question marker *ma* may not co-occur. In the **anaphor agreement** phenomenon, we first use baselines to test the LMs' gender and number bias (see Appendix D.2). Then, the morpheme *ziji* (self) is added to test if the LMs knows the function of *ziji* and agree the gender/number of the reflexive with the sentence subject. To avoid the issue caused by Chinese proper names in CLiMP, we use *gender + occupation* as the subject of sentences to clearly indicate the gender. The **aspect** phenomenon tests the knowledge of the perfective aspect markers *le* and *guo* in the sense of their interaction with tense and the progressive marker *zai*. The **classifier-noun agreement** is observed when a noun is modified by a numeral or demonstrative. One noun can be compatible with more than one classifier and the matching can be idiosyncratic. The **definiteness effect** phenomenon is established on the observation that demonstrative *zhe* (this)/*na* (that) and the quantifier *mei* (every) may not occur in the post-verbal position of an existential *you* (there is) sentence. **Polarity item**s (PI) are words or phrases whose occurrence is restricted to certain contexts (e.g., negative or affirmative). We test two negative PIs, *renhe* (any) and *shenme* (what), as well as one positive PI *huoduo huoshao* (more or less). Chinese **relative clauses** exhibit a filler-gap dependency relationship. If the gap is a simple subject or direct object position, no resumptive noun or pronoun is allowed. Lastly, the ***wh*-fronting** phenomenon shows that in absence of a specific context (e.g., an echo question), a *wh* phrase must stay in situ.

[7]The transitive verbs from CLiMP are used in a small portion of the minimal pairs in SLING's *Anaphora* dataset, which requires transitive verbs that take animate subjects and objects. The acceptability contrast of sentences does not rely on those verbs. Extracting such verbs from the Treebank was impossible because animacy of nouns is not encoded in the parse.

[8]The vocabulary and data generation code of CLiMP can be found here https://github.com/beileixiang/CLiMP.

3.5 Human Validation

Two rounds of human validation were conducted on PCIbex (Zehr and Schwarz, 2018) to verify the quality of the generated minimal pairs.[9] Eleven students from the University of Massachusetts Amherst were recruited as annotators for the first round, and five for the second round. Each student has finished at least senior high school in China, and they all use Chinese on a daily basis. For the first round evaluation, every annotator rated 20 pairs from each of the 30 paradigms (not the baselines).[10] The annotators were shown one minimal pair at a time and asked to choose the more acceptable sentence. In total, the annotation task took 1.5 to 2 hours on average, and the annotators were paid $40 each. Details on the second annotation round can be found in Appendix E. The final raw human accuracy mean over all paradigms is 97.12% (median = 97.27%, SD = 2.29%). The inter-annotator agreement as measured by Fleiss' κ is 0.8823, indicating *almost perfect agreement* (Landis and Koch, 1977).

4 Experimental Setup

Evaluated Models: There are many publicly available pretrained monolingual Chinese LMs and multilingual LMs. While Xiang et al. (2021) only test `bert-base-chinese`, three LSTM LMs, and two 5-gram LMs in their work on CLiMP, we experiment with the 18 LMs listed in Table 4.[11] There are 6 pairs of LMs (color coded in Table 4) in which one model is either trained with more parameters than the other in the pair or with larger training data.[12] Although `lstm-zh-cluecorpussmall` and `gpt2-zh-cluecorpussmall` also differ in their model structure, we pair them to see whether a Transformer-based architecture leads to better model performance. We run the same suite of LMs on CLiMP, show the results in Table 7, and discuss them in Section 5.6.

LM	Param	Tr. Size	Source
(monolingual models)			
lstm-zh-cluecorpussmall	25.8M	14G	(Zhao et al., 2019)
gpt2-zh-cluecorpussmall	102M	14G	(same as above)
CPM-Generate	2.6B	100GB	(Zhang et al., 2021a)
PanGu-α	2.6B	1.1TB	(Zeng et al., 2021)
bert-base-zh	110M	25M sent.	(Devlin et al., 2019)
zh-pert-base	110M	5.4B	(Cui et al., 2022)
zh-pert-large	330M	5.4B	(same as above)
mengzi-bert-base	103M	300G	(Zhang et al., 2021b)
mengzi-bert-base-fin	103M	320G	(same as above)
ernie-1.0	110M	173M sent.	(Sun et al., 2019)
(multilingual models)			
GPT-3-Davinci	175B		(Brown et al., 2020)
XLM-R-base	270M	2.5TB	(Conneau et al., 2020)
XLM-R-large	550M	2.5TB	(same as above)
BERT-base-multiling-cased	110M		(Devlin et al., 2019)
MT5-small	300M	26.76TB	(Xue et al., 2021)
MT5-large	1.23B	26.76TB	(same as above)
Byt5-small	300M	26.76TB	(Xue et al., 2022)
Byt5-large	1.23B	26.76TB	(same as above)

Table 4: The set of Chinese language models evaluated in this work. We consider both large monolingual models and multilingual models (separated by double line). Tr. size = training data size; zh = Chinese; sent. = sentences. Color coded LM pairs were released in the same paper, and differ in size or training data.

Evaluation: To evaluate the performance of an LM on SLING, we use perplexity for the causal LMs and pseudo-perplexity (Salazar et al., 2020) for the masked LMs (see Appendix B for details). Given a minimal pair, the LMs should assign a lower (pseudo-)perplexity to the acceptable sentence. The accuracy of each LM on a paradigm is the proportion of the minimal pairs in which the model assigns the acceptable sentence a lower (pseudo-)perplexity.

Why perplexity? We choose to use perplexity instead of other metrics (e.g., raw probability) because some phenomena in SLING have systematic difference in sentence length within minimal pairs (e.g., *Polarity Item*, *Relative Clause*). Thus, we require a length-normalized metric like perplexity, since metrics such as probability can prefer shorter sentences by nature (Wu et al., 2016; Koehn and Knowles, 2017; Brown et al., 2020; Holtzman et al., 2021). Additionally, perplexity (or pseudo-perplexity) is applicable to all phenomena and all LMs that are tested in SLING (details in Appendix B). We considered other evaluation metrics such as prefix methods (Linzen et al., 2016; Gulordava et al., 2019; Wilcox et al., 2019), by-word surprisal (Futrell et al., 2018), and training an acceptability classifier (Warstadt et al., 2019) but eventually decided not to use them for reasons

[9] After the first round, the human accuracy on the two compound noun paradigms were 61.36% and 77.27%. To improve the quality of SLING, we revised the generation process of the two paradigms and re-evaluated their quality.

[10] Ten practice and 24 filler item pairs were created to test whether the annotators understood and paid attention to the task. Those pairs are irrelevant to the paradigms of interest. All annotators did these tests with 100% accuracy.

[11] Most LMs tokenize an input sentence into characters but `CPM-Generate` and `PanGu-`α occasionally cuts an input into words, and the `ByT5` models use bytes.

[12] The `mengzi-bert-base-fin` model is `mengzi-base` further trained with 20G extra financial news and research reports.

detailed in Appendix C.

5 Results & Analysis

Table 5 reports the human performance and the results of the LMs on each phenomenon.[13] Overall, LM performance (`bert-base-zh` 84.8% being the best) lags far behind human performance (97.1%). Looking into each phenomenon, although some LMs occasionally perform better than humans (e.g., in the definiteness effect), no single LM performs consistently well. Comparing the monolingual LMs to the multilingual ones, the former performs in general better than the latter.[14] In the following subsections, we provide analyses of the model performance from the aspects of model size, distance, and hierarchy. By-phenomenon results and analyses are in Appendix F.

5.1 Model Size

To investigate whether a larger model performs better on SLING, two-tailed pairwise Wilcoxon signed rank tests were conducted on each LM pair in Table 4. The tests indicated that the performance of the LMs in the `pert` and `mengzi` LM pairs statistically significantly differed from each other while there is no statistical difference in other LM pairs. Further one-tailed pairwise Wilcoxon signed rank tests on these two pairs revealed (unintuitively) that the smaller LMs (`pert-base`, `mengzi-base`) perform better than the larger ones (`pert-large`, `mengzi-fin`). The test results can be found in Table 9 in Appendix G.3. The finding here coincides with the conclusion drawn in BLiMP and CLiMP that increasing model size does not necessarily improve the model performance.

5.2 LMs are Affected by Distance

The classifier-noun phenomenon was designed to test if the LMs are affected by distance in a dependency. For example, in (4), the classifier is separated from the noun by a long adjective,[15] making the local dependency distant. The noun phrase can also be a compound noun (5), in which case the classifier should agree with the second noun.

(4) 三户非常优秀且高效的家庭
3 households of very excellent and efficient families

(5) 三本非常优秀且高效的家庭小说
3 copies of very excellent and efficient family fiction

Two two-tailed paired Wilcoxon signed rank tests were conducted to compare the simple noun paradigm with and without a long adjective as well as the ones with compound nouns. The results indicated that there was a statistically significant difference between the model performance when the long adjective was present and absent in the simple noun paradigms. There was no such difference in the compound noun paradigm. Further one-tailed Wilcoxon signed rank tests showed that, with a long adjective, the LM performance of the simple noun paradigms decreased. The p values are reported in Table 10.

5.3 LMs struggle with Hierarchy

All LMs struggle with hierarchical phenomena and are vulnerable to linear closeness. This is shown in the results for the anaphor and classifier-noun phenomena. The anaphor phenomenon was designed to test whether the LMs prefer linear or hierarchical closeness. For the LMs to correctly choose the acceptable sentences, they should prefer hierarchical closeness. In the example in Figure 3, $\texttt{DP}_5$ can only agree in its gender feature with $\texttt{DP}_1$, which is hierarchically closer. If the LMs are distracted by the linearly closer $\texttt{DP}_3$, they would pick the unacceptable sentence in which the $\texttt{DP}_5$ is *herself*.

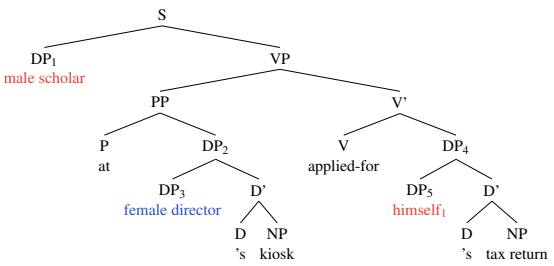

Figure 3: The syntax structure of the sentence 男学者在女导演的店里申请了他自己的退税。 (The male scholar applied for his own tax return at the female film director's shop.) The reflexive anaphor *himself* must be bound by $\texttt{DP}_1$, which is hierarchically closer, rather than $\texttt{DP}_3$, which is linearly closer. Details of the tree can be found in Appendix D.

Two two-tailed paired Wilcoxon signed rank tests were conducted on the male and female anaphor paradigms with and without a `PP` respectively. The results show that there is a statistically significant decrease in the performance when the

[13] The accuracy of each paradigm in all phenomena can be found in Appendix G.2, along with a visualization in Figure 7.

[14] The poor performance of `PanGu-`α is partially due to its strong bias toward singular number in the anaphor (number) phenomenon.

[15] In SLING, the long adjective is chosen to be eight characters of two conjoined adjectives modified by an adverb *very* as in (4-5).

Phenomenon	**human**	lstm	gpt2-zh	CPM	PanGu	bert-base-zh	pert-base	pert-large	mengzi-base	mengzi-base-fin	ernie	xlm-R-base	xlm-R-large	bert-base-multi	mt5-small	mt5-large	byt5-small	byt5-large	gpt3
Alternative question	**97.3**	13.5	47.4	85.8	10.0	93.1	89.8	79.2	75.6	73.0	94.3	53.1	56.9	6.5	45.3	10.3	25.9	55.1	14.9
Anaphor (gender)	**98.5**	74.9	67.5	71.1	99.0	88.3	60.8	50.3	92.2	89.3	81.6	59.5	61.0	82.5	50.6	37.7	53.9	37.7	63.2
Anaphor (number)	**96.5**	99.6	100	92.3	0.0	99.9	99.8	98.8	80.3	75.5	99.5	95.2	85.2	94.7	27.3	7.3	93.6	73.0	99.9
Classifier-noun	**96.4**	79.9	85.7	52.7	74.8	95.3	94.9	82.2	93.9	93.5	94.4	87.1	90.2	87.5	68.0	84.3	52.7	53.0	89.1
Aspect	**97.6**	52.4	71.9	61.2	55.8	84.1	81.6	68.4	76.3	78.3	74.3	54.1	68.9	45.0	49.8	65.1	55.3	50.9	71.5
Definiteness effect	**96.8**	97.0	99.4	70.4	68.5	96.4	95.4	73.9	96.6	96.1	88.7	63.5	72.8	94.1	72.2	49.0	14.2	9.0	81.5
Polarity item	**92.0**	90.3	86.0	78.9	79.6	72.0	90.4	94.7	97.9	98.2	81.3	96.5	96.5	44.2	78.2	81.6	59.5	62.9	85.9
Relative clause	**99.1**	72.1	44.9	50.4	14.3	34.2	38.0	89.3	18.9	13.1	33.1	43.7	48.7	13.2	42.2	50.2	2.8	18.3	65.2
wh fronting	**100**	100	99.7	93.7	94.3	99.8	99.8	99.6	99.8	99.4	99.8	97.4	99.4	67.8	81.1	98.6	13.1	44.7	100
Average over phenomena	**97.1**	75.5	78.0	72.9	55.1	84.8	83.4	81.8	81.3	79.6	83.0	72.2	75.4	59.5	57.2	53.8	41.2	45.0	74.6

Table 5: The average percentage accuracy of the LMs and human performance on each phenomenon (random guessing is 50%). Overall, humans significantly outperform all LMs. No LM performs well on all phenomena, but monolingual LMs perform better than multilingual ones. A larger model size does not imply better performance. The vertical line separates the mono/multilingual models. The anaphor phenomenon accuracies include the baselines.

distractor is present.[16] The descriptive and test statistics can be found in Table 11.

The classifier-noun phenomenon is designed to test whether the LMs are aware of the right headedness of Chinese compound noun and match the classifier with the second noun in a compound noun rather than the first one (cf. (4) and (5)). If the LMs do not have this knowledge but prefer linear closeness, they would choose the wrong sentence in a minimal pair. The statistics and the results of two two-tailed Wilcoxon signed rank tests in Table 12 show that the LMs performed worse when the distractor was present.

5.4 Strong Gender and Number Bias

Because the LMs can have gender and number bias, in the anaphor phenomena, we use baselines (e.g., *The male baker likes him / her.*) to test the bias.[17] The higher the accuracy number is, the more biased a LM is towards *him*. Figure 9 in Appendix G.3 shows that, with a male subject, only four monolingual LMs (`gpt2-zh`, `CPM`, `pert-base`, and `ernie`) are gender neutral. When the subject is female, all LMs are biased towards a female object (see Figure 12).

One reviewer raised concern that the anaphora resolution in those baselines can only be reliably solved in context of the preceding text, which is true in real life situations. However, in our test setting, since there is no context, the models should ideally be gender neutral on average (Bordia and Bowman, 2019).

The LMs also have number bias. A baseline example is *The three male bakers like them / him*. The higher the accuracy number is, the more biased a LM is towards *them*. As seen in the results in Table 8 (Appendix G.3), while most LMs are biased to a plural object when the subject is plural, `PanGu-`α is strongly biased to a singular object.

The purpose of the baselines is to reliably test whether the LMs know that the gender/number of *ziji* (self) should agree with the subject's gender/number in the paradigms. As it turn out, the female and number features are not useful for our purpose because the LMs already achieve a 'high' accuracy in the baselines, making it ambiguous whether the high accuracy in non-baselines is because they know the function of *ziji* (self) or they are just biased. The male self paradigm, on the other hand, shows that most monolingual LMs were able to use *ziji* as a hint to agree the gender of the subject and object. Among the multilingual LMs, only `gpt3-davinci` achieved a meaningful accuracy increase.

5.5 Vulnerable to Uncertainty

In the current study, *haishi*, *le*, and *wh* phrases can have more than one usage depending on contexts. The observation is that the LMs performed worse on the paradigms with those phrases. This is most obvious in the aspect and polarity item phenomena.

In the aspect phenomenon, the possible position of *guo* is relatively fixed compared to *le*, and there is no interaction between *guo* and the progressive marker. The LMs performed better on the *guo* paradigms than on *le*.

In the polarity item phenomenon, the contexts

[16] This is even the case in the female paradigms where the LMs are strongly biased. The female baseline row in Table 8 shows that when the sentence subject is female, and there is no need for the object to agree with the subject of the sentence, the LMs strongly biased towards a female object. Detailed explanation of the baselines can be found in Appendix D.2.

[17] The Chinese baseline has the same structure as this English translation.

where the positive polarity item *more or less* can occur is more restricted than *any*, which is more restricted to *wh* phrases. And we see that the LM performance is the best on *more or less*, followed by *any*, and the worst on *wh* phrases.

5.6 Evaluating Our Set of 18 LMs on CLiMP

We ran the 18 LMs on CLiMP and compare model rankings and performance on CLiMP and SLING. We observe major differences: the best LM on SLING is `bert-base-chinese` (84.8%), and on CLiMP it is `chinese-pert-base` (81.22%). That said, monolingual LMs perform better than multilingual LMs on both datasets.[18] While the average performance of the LMs on both datasets is similar (SLING 69.7%, CLiMP 70.1%), on average LMs have significantly larger variation across phenomena on SLING (SD = 24.1%) than on CLiMP (SD = 13.2%). Thus, SLING is more discriminative of the strengths and weaknesses of LMs, as LMs tend to be more polarized to one direction across phenomena in SLING compared to those in CLiMP. Finally, because CLiMP does not test the LMs' bias in the gender and number features for their binding and anaphor paradigms, the LM performance on these two paradigms is uninformative since we do not know what role the bias plays in the tests. SLING corrects this issue by including 8 baseline paradigms and shows that the LMs can be strongly biased (see Section 5.4).

6 Conclusion

We present SLING, a new benchmark for evaluating Chinese linguistic knowledge in large scale pretrained LMs. Unlike the existing CLiMP dataset, in which we identify several critical issues, we construct SLING from naturally-occurring sentences in the Chinese Treebank. Our results show that monolingual Chinese LMs achieve better performance on SLING than multilingual LMs. We find that LMs are better at handling local dependencies than long-range dependencies or with distractors, and that they are better at syntactic rather than semantic phenomena. Overall, there remains a large gap between LM and human performance.

Limitations

As a benchmark of evaluating LMs' Chinese linguistic knowledge, SLING covers 9 major Chinese grammatical phenomena with 38k minimal pairs. However, there are still phenomena that are important but not included in the current work: for example, the *ba* and *bei* constructions. For those structures, unacceptability can have different sources (e.g., syntax or pragmatics).[19] Simple syntactic structure restrictions are not enough. When deciding which phenomena to include in SLING, we deliberately avoid such cases because the (un)acceptability of these phenomena can be mitigated by contextual or world knowledge. As a result, human judgement can vary significantly. As an example, take the *bei* construction (*Passive*): the sentence 王萍被嘴举了 (Wang was lifted by a mouth) is wildly bizarre to some people, while for others, it is acceptable because it is possible to imagine a world in which each body part is a mighty character that can lift things. Such "unacceptable" sentences are different from *The roses* ***is*** *red.*, which cannot be resolved by any context.

Another limitation is that even though Chinese Treebank 9.0 contains a rich and diverse vocabulary, it can still be inadequate at times. For example, for the classifier-noun agreement phenomenon in SLING, we were not able to extract enough high-quality compound nouns and thus had to manually create 196 minimal pairs, as described in Appendix E. One possible way to get around this limitation is to train a parser on the Treebank and use it to automatically parse even more raw Chinese data. We leave this for future work.

Ethical Considerations

Following best practices (McMillan-Major et al., 2021), we plan to open source our dataset along with a data card. We will follow the templates used in the GEM benchmark (Gehrmann et al., 2021)[20] and HuggingFace Datasets repository (Lhoest et al., 2021).[21] Overall, our project had a small computational cost since we did not need to do any model training. We performed inference on all 18 LMs on a single RTX8000 GPU with 48GB memory. All inference experiments in this paper can be completed within a day on the single GPU.

[18] Kendall Tau correlation of the two rankings for monolingual LMs is 0.42 and for multilingual LMs is 0.79.

[19] For possible sources of unacceptability of a sentence, please see (Abrusán, 2019).

[20] `https://gem-benchmark.com/data_cards`

[21] `https://huggingface.co/docs/datasets/v1.12.0/dataset_card.html`

Acknowledgements

First and foremost, we would like to thank all the anonymous reviewers for their valuable comments. We also thank the native Chinese speakers who helped us obtain human performance numbers on SLING. We are very grateful to Brian Dillon and Simeng Sun for helping formulate the project idea in the early stages of the project. We are also thankful to Yutao Zhou and all the participants in the Semantics Workshop at UMass Linguistics and the UMass NLP group for comments and suggestions during the project. Kalpesh Krishna was supported by the Google PhD Fellowship awarded in 2021.

References

Barbara Abbott. 1993. A pragmatic account of the definiteness effect in existencial sentences. *Journal of Pragmatics*, 19(1):39–55.

Márta Abrusán. 2019. Semantic anomaly, pragmatic infelicity, and ungrammaticality. *Annual Review of Linguistics*, 5:329–351.

Sigrid Beck. 2006. Intervention effects follow from focus interpretation. *Natural Language Semantics*, 14(1):1–56.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Li Chen. 2012. *Chinese polarity items*. Ph.D. thesis, City University of Hong Kong.

Lisa Lai-Shen Cheng. 1994. Wh-words as polarity items. *Chinese Languages and Linguistics*, 2:615–640.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yiming Cui, Ziqing Yang, and Ting Liu. 2022. Pert: Pre-training bert with permuted language model. *arXiv preprint arXiv:2203.06906*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. 2018. RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint arXiv:1809.01329*.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, et al. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Anastasia Giannakidou, Claudia Maienborn, Klaus von Heusinger, and Paul Portner. 2019. Negative and positive polarity items. *Semantics—Sentence and information structure*, pages 69–134.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2019. Colorless green recurrent networks dream hierarchically. *Proceedings of the Society for Computation in Linguistics*, 2(1):363–364.

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

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings*

of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jianhua Hu and Haihua Pan. 2008. Focus and the basic function of Chinese existential you-sentences. In *Existence: Semantics and syntax*, pages 133–145. Springer.

Cheng-Teh James Huang, Yen-hui Audrey Li, and Yafei Li. 2009. *The syntax of Chinese*, volume 10. Cambridge University Press Cambridge.

Edward L Keenan. 1987. A semantic definition of "indefinite NP". In Eric J. Reuland and Alice G. B. Ter Meulen, editors, *The Representation of (In)Definiteness*, pages 286–317. Mit Press.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Rajesh Kumar. 2013. *The syntax of negation and the licensing of negative polarity items in Hindi*. Routledge.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jing Lin and Anastasia Giannakidou. 2015. No exhaustivity for the mandarin NPI shenme. *Unpublished Manuscript*.

Jo-Wang Lin. 1998. On existential polarity-wh-0phrases in Chinese. *Journal of East Asian Linguistics*, 7(3):219–255.

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.

Angelina McMillan-Major, Salomey Osei, Juan Diego Rodriguez, Pawan Sasanka Ammanamanchi, Sebastian Gehrmann, and Yacine Jernite. 2021. Reusable templates and guides for documenting datasets and models for natural language processing and generation: A case study of the HuggingFace and GEM data and model cards. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 121–135, Online. Association for Computational Linguistics.

Haihua Pan and Peppina Lee. 2004. The role of pragmatics in interpreting the Chinese perfective markers-guo and-le. *Journal of Pragmatics*, 36(3):441–466.

Yip Po-Ching and Don Rimmington. 2015. *Chinese: A comprehensive grammar*. Routledge.

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

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

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019b. What do you learn from context? Probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019*.

Ildikó Tóth. 1999. Negative polarity item licensing in Hungarian. *Acta Linguistica Hungarica*, 46(1):119–142.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Leslie Fu-mei Wang. 2002. From a motion verb to an aspect marker: A study of guo in Mandarin Chinese. *Concentric: Studies in Linguistics*, 28(2):57–84.

Lianqing Wang. 1994. *Origin and development of classifiers in Chinese*. Ph.D. thesis, The Ohio State University.

Yu-Fang Flora Wang and Miao-Ling Hsieh. 1996. A syntactic study of the Chinese negative polarity item renhe. *Cahiers de linguistique-Asie orientale*, 25(1):35–62.

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.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Huiying Wen. 2020. Relative clauses in Mandarin Chinese. *Queen Mary's Occasional Papers Advancing Linguistics (OPAL, no. 46).*

Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. 2019. Structural supervision improves learning of non-local grammatical dependencies. In *Proceedings of NAACL-HLT*, pages 3302–3312.

Ying Wu. 2010. "haishi" de duoyixing yu xide nandu [the polysemy and the acquisition difficulty of *haishi*]. *TCSOL Studies*, pages 41–48.

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:1609.08144*.

Fei Xia. 2000. The segmentation guidelines for the penn chinese treebank 3.0. *IRCS Technical Reports Series. 37.*

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.

Liejiong Xu. 1995. Definiteness effects on Chinese word order. *Cahiers de linguistique-Asie orientale*, 24(1):29–48.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *NAACL-HLT*.

Nianwen Xue, Xiuhong Zhang, Zixin Jiang, Martha Palmer, Fei Xia, Fu-Dong Chiou, and Meiyu Chang. 2016. Chinese Treebank 9.0 LDC2016T13.

Keiko Yoshimura. 2007. *Focus and polarity: even and only in Japanese*. The University of Chicago.

Jeremy Zehr and Florian Schwarz. 2018. Penncontroller for internet based experiments.

Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, Chen Li, Ziyan Gong, Yifan Yao, Xinjing Huang, Jun Wang, Jianfeng Yu, Qi Guo, Yue Yu, Yan Zhang, Jin Wang, Hengtao Tao, Dasen Yan, Zexuan Yi, Fang Peng, Fangqing Jiang, Han Zhang, Lingfeng Deng, Yehong Zhang, Zhe Lin, Chao Zhang, Shaojie Zhang, Mingyue Guo, Shanzhi Gu, Gaojun Fan, Yaowei Wang, Xuefeng Jin, Qun Liu, and Yonghong Tian. 2021. Pangu-α: Large-scale autoregressive pretrained Chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369*.

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

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

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

Min Zhou and Jingquan Han. 2012. A phase-based approach to the derivation of relative constructions in Mandarin Chinese. *Journal of Foreign Languages*, 3(002).

Alessandro Zucchi. 1995. The ingredients of definiteness and the definiteness effect. *Natural Language Semantics*, 3(1):33–78.

A Ngram Count of CLiMP and SLING

CLiMP contains 16K minimal pairs (32K sentences) and SLING 38K (76K sentences). The average sentence length in CLiMP is 11.8 (median = 11) and in SLING is 12.5 (median = 12). Because of the difficulty of defining what counts as a word in Chinese, we report one to four ngram counts of types in Table 6, together with the word type counts returned by Jieba.[22] Because SLING has more sentences which can lead to larger type counts, we randomly shuffled the sentences and took 32K sentences to calcuate the ngram and Jieba counts of word types.

	CLiMP	SLING-32K	SLING-76K
1gram	1033	2756	2886
2gram	22289	33031	43122
3gram	62353	64257	92972
4gram	102772	87532	133900
Jieba	2335	9872	11987

Table 6: Counts of one to four ngram types in CLiMP and SLING and word type counts by Jieba.

One reason for having 1K sentence pairs in each paradigm is to cancel out the potential influence of word frequency on the perplexity of sentences. Having a diverse vocabulary surely helps in this sense.

B Metrics

Causal LMs Perplexity (PPL) is used for causal LMs to decide the preferred sentences. Each token w is assigned a probability p given the prefix being seen. The perplexity is calculated based on the log likelihood (L). For a sentence of length m, its perplexity is calculated as below:

$$L = \frac{1}{M}\sum_{i=1}^{m} \log p(w_i|w_1...w_{i-1})$$
$$\text{PPL} = \exp(-L)$$

Each sentence in a minimal pair is assigned a perplexity value. The one with the lower perplexity is taken as the good sentence that the models choose.

Masked LMs Pseudo-perplexity values (pseudo-PPL) are used to evaluate masked LMs (Salazar et al., 2020). Concretely, tokens in a sentence is masked one after another (w_j). The masked language models return a probability distribution over the vocabulary in the masked position given the context surrounding it. For a sentence of length m, its pseudo-perplexity is calculated as follows:

$$w_{\setminus i} = w_1...w_{i-1}, w_{i+1}...w_m$$
$$\text{pseudo-}L = \frac{1}{M}\sum_{i=1}^{m} \log p(w_i|w_{\setminus i})$$
$$\text{pseudo-PPL} = \exp(-L)$$

C Related Work: Methods of Evaluating Linguistic Knowledge and Their Limitations in SLING

To investigate what kind of and how much linguistic knowledge large-scale pretrained LMs have compared to human, previous works have focused on limited LMs and probed into the internal encoding of the linguistic knowledge (Tenney et al., 2019a,b; Clark et al., 2019). Other works investigate the LMs' linguistic knowledge of a small subset of English syntactic grammar by using prefix methods (Linzen et al., 2016; Gulordava et al., 2019; Wilcox et al., 2019), by-word surprisal (Futrell et al., 2018), or trained an acceptability classifier (Warstadt et al., 2019).

Prefix method Linzen et al. (2016) focus on English subject-verb dependencies and use a prefix method for evaluation, which requires LMs to assign probabilities to the next word given a prefix. The grammatical next word is expected to have a higher probability (e.g., *The keys **are*** vs. **The keys **is***). The task includes local subject-verb dependencies (e.g., *The keys **are*** vs. **The keys **is***) as well as dependencies in distance with distractors (e.g., *The roses in the vase by the door **are*** vs. **The roses in the vase by the door **is***). The prefix method is adopted in later works, for example, Gulordava et al. (2019) and Wilcox et al. (2019).

The limitation of the prefix methods is that it mostly applies to inflectional grammatical phenomena in a dependency relationship. For Chinese, a language that largely lacks inflection, the usage of the methods is very limited. Taking SLING as an example, the prefix methods are *not* applicable to all nine phenomena because the minimal pairs' acceptability depends on:

- the presence/absence of a crucial word (Alternative Question, Anaphor (number), Aspect, Polarity Item, Relative Clause);
- the word order (Aspect, *wh* fronting);

[22] https://github.com/fxsjy/jieba

- the choice of a crucial word in the middle of a sentence whose acceptability depends on the part of sentence that is after the word (Anaphor (gender), Classifier-Noun, Definiteness Effect, Polarity Item, Relative Clause).

By-word surprisal Another evaluation method, inspired by the controlled psycholinguistic experimentation, is the by-word surprisal[23] and sentence completion methods proposed by Futrell et al. (2018) to explore LMs' knowledge of syntax. The surprisal reflects whether LMs are affected by the presence/absence of critical words in grammatical configurations. In the sentence completion task, LMs completes a sentence given a prefix. Human annotators then judge the grammaticality of the completed sentences.

The by-word surprisal method solves one limitation of the prefix methods (i.e., the acceptability depends on the presence/absence of a crucial word) but still does not account for the other two listed above. The sentence completion method faces similar restrictions and cannot be applied in a large scale because it requires human judgement of the completed sentences.

Acceptability classifier Warstadt et al. (2019) trained an acceptability classifier to perform a grammaticality judgement task, which consists of sentences collected from the linguistics literature marked for their acceptability.

There are several limitations of training a classifier. First, it involves many debatable design decisions (e.g., hyper-parameters). Second, LMs may learn the task from the training data (Hewitt and Liang, 2019; Voita and Titov, 2020). Our goal is to measure the linguistic capability of *pretrained* LMs without additional help from a training dataset that has the same distribution as the test set.

Overall, the previous methods are either only applicable to a subset of linguistic grammar or depend on the performance of a classifier. The minimal pair method used in BLiMP breaks through these limitations.

Minimal pair method To cover a wide range of linguistic phenomenon, Warstadt et al. (2020) introduced minimal pair evaluation for LMs and created the Benchmark of Linguistic Minimal Pairs for English (BLiMP). It evaluates the linguistic knowledge of twelve English grammatical phenomena including syntax and semantics. Each of them consists of minimal pair paradigms representing different aspects of the phenomena. All minimal pairs are code-generated using templates created by linguists and an annotated vocabulary that contains 3000 words. The dataset is human validated.

The results on BLiMP show that the LMs tested in BLiMP are good at local dependency relations (e.g., morphology agreement) but bad at phenomena involving hierarchy and semantic knowledge. Concerning the training size and model size, while increasing training size can improve model performance, increasing model size does less so.

Other possible metrics and their limitations Other possible metrics are probability and a masked-token method. However, probability is not a suitable metric to use in SLING for at least two reasons. First, probability is only useful for minimal pairs whose sentences have the same length. Otherwise, probability by nature prefers shorter sentences. Second, the sentences in a minimal pair need to have similar word orders. This is because tokenizers might tokenize a sentence in different ways depending on the word order, causing the sentence length of the sentences in a minimal pair to be different. In the masked-token method, we can mask out the crucial word in each sentence in a minimal pair and ask a LM to give probability of the two masked words. This method is not applicable to causal LMs. For masked LMs, it is only applicable to Anaphor (gender), Classifier-Noun, and Definiteness Effect in SLING where the word order does not change. In those cases, since SLING uses minimal pairs, the masked token in those phenomena will be exactly the part in which the sentences in a minimal pair differ. Hence, the masked-token method will return the same results as the pseudo-perplexity.

D Linguistic Phenomena

The current work focuses on six syntax and three semantics phenomena in Chinese. Table 3 offers an overview. There are 30 test paradigms. The anaphor phenomenon has 8 baseline paradigms to detect LMs' gender (male/female) and number (singular/plural) biases.

All phenomena have at least one paradigm that can be solved by checking the linear order of tokens. Some phenomena require a negative co-occurrence of words. For example, in the alternative question

[23] Surprisal is the log inverse probability of a word given its prefix (Hale, 2001).

phenomenon, the disjuntor *haishi* and the polar question particle *ma* may not co-occur. Other phenomena require a positive co-occurrence. For example, in the polarity item phenomenon, the grammaticality of *renhe* (any) dependes on the occurrence of negation.

Three phenomena contain paradigms that require the LMs to use the knowledge of hierarchy. If LMs use linear closeness rather than hierarchical closeness, they will wrongly assign a lower perplexity to the unacceptable sentence in a minimal pair. The anaphor phenomenon, for example, contains such paradigms.

The anaphor, classifier-noun agreement, and relative clause phenomena have paradigms that test LMs' robustness to distractors and long distance dependencies. A distractor is an element that intervenes between the head and its dependent in a dependency/agreement relation. For example, in *The **roses** in the vase **are** ...*, **roses** and **are** are in a dependency relation, and vase is the distractor. By distance, it is meant to be the case that the head and its dependent is separated from each other (e.g., ***these** beautiful red blooming **roses***).

This section introduces phenomena in turn. If a phenomenon is in CLiMP, a comparison between CLiMP and the current work will be provided.

D.1 Alternative Questions with *haishi*

Chinese alternative questions (AltQ) are most reliably marked by the disjunctor *haishi* (Huang et al., 2009). Although *haishi* has different usages (Wu, 2010), when it is used as the disjunctor, the polar question particle *ma* (SP) cannot occur. Minimal pairs like (6) test whether LMs are aware of this. The paradigm concerns only linear co-occurrence.[24]

(6) tamen shi laoshi haishi mujiang (*ma)?
they are teacher or carpenter (*SP)
"Are they teachers or carpenters?"

D.2 Anaphor

Mandarin Chinese has two reflexive pronouns: *ziji* and *ta(men)-ziji*. The former is morphologically simple with no person, number, or gender features. The latter, *ta(men)-ziji*, has the pronoun *ta* which encodes gender features in writing: 她 for singular female third person, 他 for singular male third person, and 它 for singular non-human third person. The character *men* indicates plurality. Because of this morphological richness, *ta(men)-ziji* is used to form minimal pairs. Since CLiMP contains the binding phenomenon, their implementation will be first introduced, followed by the binding phenomenon in the current work.

Binding Phenomenon in CLiMP Xiang et al. (2021) use singular female and male third person reflexives *ta-ziji* to test the LMs' knowledge of binding. There are two paradigms. The first one has a simple SVO structure in which the object is an anaphor and needs to match the gender feature of the subject. The second paradigm involves a distractor between the antecedent and the reflexive (e.g., DP_2 in Figure 4). The distractor is different from the true antecedent in its gender feature. The distractor is linearly closer to the reflexive but hierarchically farther. It turns out that the LMs struggle with this paradigm. The results show that the LMs did no better than chance. One of the acceptable binding sentences in Xiang et al. (2021) is cited below. We provide its syntax in Figure 4. The corresponding unacceptable sentence changes *herself* to *himself*.

(7) Huang Xiuying danxin Wang Hao
female.name worry-about male.name
zhihou guanchaguo ta-ziji.
after observe herself
"After Huang Xiuying worried about Wang Hao, she observed herself."

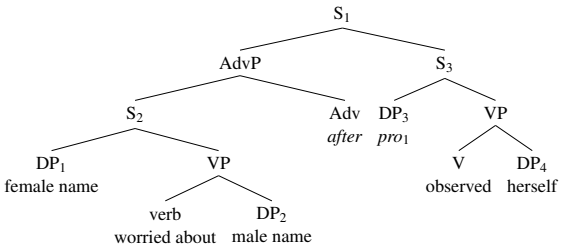

Figure 4: The syntax structure of (7).

Although, by comparing the two paradigms, Xiang et al. (2021) find the models are bad at dealing with hierarchy and distractors, there are four shortcomings in the minimal pair design that weaken the strength of the observation. First, it was not tested whether the LMs knew the gender of the proper names. Because Chinese names do not always clearly indicate the gender, this can cause the LMs guessing randomly. Second, the syntax of the second paradigm is complex because it involves ellipsis.[25] With the presence of ellipsis, it is

[24] The notation *(*ma)* in (6) means that the sentence is good without *ma* but bad with it.

[25] The ellipsis is presented as pro_1 in DP_3 in Figure 4. The

not for sure that the models did bad because they preferred a linearly closer agreement or because they couldn't recover the omitted subject correctly. Third, CLiMP does not have a baseline for the gender biases of the LMs. Hence, we cannot know if the models know the function of *ziji* or they simply prefer one gender. Fourth, CLiMP does not have separate corpora for the two genders. Thus, we do not know if the LMs are bad in both female and male reflexive agreements or only in one of them.

Paradigms in Current Work To amend the four shortcomings, the current work includes baseline paradigms to test LMs' gender bias. Sentences have a simple SVO structure. Instead of using proper names as the subject, the paradigms use gender plus occupations to indicate the gender of a noun. The female and male reflexive agreements are tested separately.

To form the baseline minimal pairs for the male reflexive agreement, an occupation and a transitive verb were chosen randomly. Following the verb is either a male or female pronoun. Example (8) is one resulting minimal pair.

(8) nan dianyuan baituole **ta** / **ta**.
male shop assistant got rid of him / her
"The male shop assistant got rid of him."

Both sentences are acceptable. The purpose is to see whether the models are gender biased when there is no clue for any gender agreement. Other baselines are formed in the same way.

With the baseline being established, the minimal pairs for the reflexive agreement are created by adding *ziji* to the end of the sentences in the baselines. This turns (8) into (9). Because the presence of *ziji*, the gender of *ta* should agree with the gender of *the male shop assistant*. Hence, *himself* is acceptable but *herself* is not. Such agreement can be solved by linear closeness.

(9) nan dianyuan baituole **ta**- / ***ta**-ziji.
male shop assistant got rid of him- / *herself
"The male shop assistant got rid of himself."

The next paradigm tests whether LMs prefer a linearly closer or a hierarchically closer noun as the antecedent of an anaphor. An example is (10). The syntax of the grammatical sentence is in Figure 5.

(10) nan xuezhe zai nü daoyan de dian
male scholar at female director DEG shop
shenqingle **ta**- / ***ta**-ziji de tuishui.
applied-for him- / *herself DEG tax return
"The male scholar applied for his own tax return at the female film director's shop."

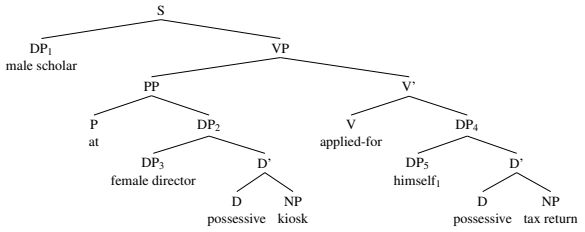

Figure 5: The syntax structure of the sentence in (10) with *himself* being bound by DP1.

Like Figure 4, Figure 5 involves a distractor DP3 but has no ellipsis. It is a SVO sentence with a preposition phrase (PP) modifying the verb phrase. The antecedent of DP5 can only be DP1 which c-commands *himself* while DP3 is embedded deeply in PP. DP1 is hierarchically closer to *himself* while DP3 is linearly closer. The LMs will fail if they have no knowledge of hierarchical structure.

The current work also uses the number feature to test LMs. Baselines are used to see if the tested LMs are biased to singularity or plurality. The gender feature is kept constant so that any distinct behaviour is only caused by the number feature.

D.3 Aspect Marker *le* and *guo*

The morphemes *le* and *guo* often function as perfective aspect markers.[26] Although they can occur in sentences of various tenses, without the help of a future oriented adverb together with morphemes as *cai* or *jiu*, they only occur in sentences of past tenses. A paradigm is built on this observation. An example is in (11).

(11) ta **qu** / ***ming** nian zhiding zhengce le.
he last / *next year establish policy AS
"He established policies last year."

The next paradigm is based on a restriction on *guo* that it cannot co-occur with the progressive marker *zai*, as in (12).

(12) tamen zai shi (***guo**) na ge fuwu.
they AD try (*AS) DT M service
"They are trying out that service."

The above paradigms can be solved linearly but the interaction between *le* and *zai* requires the knowledge of hierarchy. The morpheme *le* can co-occur with *zai* if *le* takes scope over *zai* but not

index 1 indicates its antecedent is DP1.

[26] For the other usages of *le* and *guo*, see Huang et al. (2009), Wang (2002), and Pan and Lee (2004), among others.

the other way. Based on this, two paradigms are formed. The first one (13) tests the knowledge that *le* cannot scope under *zai*. The other paradigm (14) shows that *le* can scope over *zai*.

(13) tamen zai guancha (***le***) xuanju.
they AD observe (*AS) election
"They are observing the election."

(14) a. tamen zai jiao fakuan **le**.
they AD pay fine AS
"They are (already in the process of) paying the fine."

b. * tamen zai jiao **le** fakuan.
they AD pay AS fine

D.4 Classifier-Noun Agreement

Classifiers are pervasive in Mandarin Chinese.[27] They match with nouns and indicate in what unit a noun is quantified (Huang et al., 2009). The difficulty in classifier-noun agreement is that the matching can be idiosyncratic, and one noun can be compatible with multiple classifiers.

CLiMP includes the classifier-noun agreement phenomenon which consists of three paradigms. However, because the variables in their minimal pairs are not well controlled, the experiment results are not conclusive.

Classifier-Noun Agreement in CLiMP Their first paradigm is the local classifier-noun matching. The second paradigm inserts an adjective with two to four characters between the classifier and the noun to increase the distance of the two. There is no distractor in the adjective. The third paradigm further increase the distance by having a relative clause instead of an adjective. Without showing the results of each paradigm, Xiang et al. (2021) report that the mean of the model performance is 71.66% (median 70.1%). Chinese BERT performs the best (92.9%). The overall human accuracy of the paradigms is 99.7%.

There are two issues with the paradigms. First, some minimal pairs do not show a clear contrast. Example (15) is taken from CLiMP, in which the classifier *jia* is intended to be unacceptable. However, both *liang* and *jia* are compatible with the noun *bike*.

(15) Sun Yingying zhengzai reng yi **liang** / *__jia__ zixingche.
female name PROG throw one M / *M bike
"Sun Yingying is throwing a bike."

The reason for the issue is that each noun in the CLiMP vocabulary is associated with only one classifier. However, as mentioned before, the classifier-noun matching can be a many to many relation. The second issue is the relative clauses in the third paradigm. Some relative clauses contain a distractor. In certain cases, the distractor even matches the classifier.

Paradigms in Current Work The current work has five paradigms for the classifier-noun agreement. To avoid the issues in CLiMP, we built a classifier-noun dictionary. Each noun is associated with a group of classifiers. When creating the minimal pairs, it is ensured that the classifier in the unacceptable sentences is not listed as a compatible classifier of the noun.

In the five paradigms, one paradigm tests models' knowledge of the linear order of demonstratives (DT) or numerals (CD) and classifiers (M) before a noun. The other four paradigms test LMs' knowledge of classifier-noun agreement.

The first of the four paradigms involves local classifier-noun agreement. The second paradigm inserts a long adjective between the classifier and the noun but, still, no knowledge of hierarchy is needed. The third paradigm is based on compound nouns. An example is given in (16).

(16) yi **ming** / *__tiao__ tielu jingcha
one M / *M railway policeman
"a railway policeman"

A Chinese compound noun can be formed by two nouns, noun1 (*railway*) and noun2 (*policeman*), with noun1 modifying noun2. The classifier agrees with noun2 (Huang et al., 2009). Hence, noun1 functions as a distractor. In (16), *ming* is the classifier for *policeman* while *tiao* is for *railway*. The last paradigm adds a long adjective after the classifier in the third paradigm. For the compound noun paradigms, the knowledge of hierarchy is needed. That is, the LMs should know the right-headedness of Chinese compound nouns.

D.5 Definiteness Effect

It has long been noticed that certain strong determiners cannot be in the postverbal position in an

[27] In the current paper, the word 'classifier' is used as a cover term for both classifiers and measure words. For the differences between classifiers and measure words, interested readers can refer to Wang (1994).

English existential *there*-sentence (Keenan, 1987; Abbott, 1993; Zucchi, 1995). Similar effects have been observed in Chinese (Xu, 1995; Hu and Pan, 2008). The phenomenon to be tested here involves Chinese *you* (have), a close counterpart to the *there*-construction. The demonstratives *zhe* (this) and *na* (that) as well as the quantifier *mei* (every) are used as an equivalence to the strong determiners in English. The phrase *yi* (one) + M is used as a counterpart of English weak determiners. This paradigm can be solved by checking the linear co-occurrence of two elements, *here/there* and the strong determiners. An example is in (17).

(17) a. zheli/nali you **yi** jia yingyuan.
here/there exist one M cinema
"Here/there exists a cinema."

b. * zheli/nali you **zhe/na/mei** jia yingyuan.
here/there exist DT/DT/every M cinema

D.6 Polarity Items

Polarity items (PI) are common in natural languages (Tóth, 1999; Yoshimura, 2007; Kumar, 2013; Giannakidou et al., 2019, a.o.). English, for example, has *any*, *ever*, and *yet*, etc. In Chinese, *renhe* (any) and *shenme* (what) are two actively investigated negative PIs. They occur in negation, polar questions, and conditionals (Cheng, 1994; Wang and Hsieh, 1996; Lin, 1998; Chen, 2012; Lin and Giannakidou, 2015). The phenomenon contains three paradigms. There is no complex hierarchical structure involved. All paradigms can be solved by just checking the linear co-occurrence or absence of certain tokens. The first one concerns *renhe* (any). The acceptability contrast is established by the presence of negation.[28]

(18) ta *(**bu**) fazhan renhe youhao guanxi.
she not develop any friendly relations
"She does not develop any friendly relations."

The second paradigm involves *shenme*, a multifunctional phrase. It is often seen in *wh*-questions (e.g., ni_{you} chi_{eat} $shenme_{\text{what}}$ "what do you eat?"). However, *shenme* also occurs in the contexts where typical negative PIs occur. The acceptability contrast is manipulated by the presence of negation. Yet, to avoid a *wh*-question reading, the adverb *shenzhi* (even) is used, which can occur in affirmative or negative contexts but not in *wh*-questions as it can be a focus intervener (Beck, 2006).

(19) tamen shenzhi *(**mei**) sheji shenme liyi.
they even not involve what interests
"They weren't even involved in any interests."

The last paradigm in the current phenomenon focuses on the adverb *huoduo huoshao* (more or less). It is less studied than *renhe* (any) or *shenme* (what). Nonetheless, by searching in the corpus CCL[29], it is confirmed that there is no sentence in which *bu* or *mei* (not) negates the verb within 10 characters before or after *huoduo huoshao*. Hence, the acceptability of the minimal pairs is built on the absence of negation.[30]

(20) tamen huoduohuoshao (***mei**) fadong le jingong.
they more-or-less (*not) start AS attack
"They more or less started the attack."

D.7 Relative Clauses

Relative clauses in Mandarin Chinese are head-final, meaning a modifying clause occurs before a modified noun. This characteristic is tested in CLiMP. Another characteristic of Chinese relative clauses is that it is a filler-gap construction and, in the gap position, a resumptive noun is out of the question, and a resumptive pronoun cannot occur freely. As cited in Wen (2020), Zhou and Han (2012) point out that resumptive pronouns may not occur in simple subject or direct object positions. The current study uses this property and constructs minimal pairs as in (21). If the LMs are not aware of the relative clause structure in those sentences, they can perform poorly because of the local coherence created by the filled-in gaps.

(21) ta jiandao le na ge (***nü** **jingcha / ta**)
she see AS DT M (*female police / she)
zhizhi le baoli de nü jingcha.
stop AS violence DEC female police
"She saw the female police officer who stopped the violence."

[28] The notation ***(bu)*** means that the sentence is unacceptable without *bu*.

[29] CCL is a Chinese corpus curated by Center for Chinese Linguistics at Peking University. It contains 581,794,456 characters in its Contemporary Chinese corpus. Text sources include transcribed spoken language, newspaper, practical writing, literature, etc. Details can be found at http://ccl.pku.edu.cn:8080/ccl_corpus/corpus_statistics.html.

[30] The minimal pairs of this paradigm differ in two aspects. First, the acceptable sentences contain *le* but the unacceptable ones do not. Second, the acceptable sentences do not contain *mei* but the unacceptable ones do. This seems render the pairs not minimally distinct. However, the morpheme *mei* is a negation that encodes the perfective aspect. This is what *le* does in the acceptable sentences. Keeping *le* in the unacceptable sentences will make them unacceptable for a reason that is not at issue here. Hence, even though on the surface the two sentences are not minimally distinct, they semantically are.

D.8 *Wh*-fronting

As mentioned in Section D.6, *shenme* is frequently used to form *wh*-questions. In canonical *wh*-questions, the *wh*-phrases stay in situ (Huang et al., 2009). Without a very specific appropriate context, *wh*-fronting is unacceptable. Hence, no matter whether *shenme* alone functions as an object or modifies a noun as in (22), the noun phrase containing it cannot be fronted. To force a question reading of *shenme*, the phrase *jiujing* or *daodi* (on earth) are added. There is no complex hierarchy in the sentences and the *wh* phrases are all objects.

(22) a. tamen shang ge yue daodi goujie
they last M month on earth collude with
le shenme (heidao)?
AS what mobster
"What (mobster) on earth did they collude with last month?"

b. * shenme heidao tamen shang ge yue
what mobster they last M month
daodi goujie le?
on earth collude with AS

E Second Round of Human Validation

The minimal pairs of the two compound noun paradigms were refined. Among the 2000 new minimal pairs, 1804 were code generated and 196 were manually created. To verify the minimal pair quality, a second round of human validation was conducted. Five annotators (3 female, 2 male) with an average age of 22.2 were recruited the same way as described in Section 3.5.

Twenty pairs of sentences were randomly sampled from both the code generated and manually created minimal pairs from each paradigm. The practice and filler items were used. Each annotator rated 114 pairs. They did the practice and filler items with 100% accuracy. The task took less than 10 minutes. The annotators were paid $5. The raw accuracy on the new validated pairs was 95.25% (κ = 0.8823). The manually created minimal pairs had a higher accuracy than the code generated ones (97.5% *vs*. 93%). After the second round, the raw human accuracy mean over all paradigms is 97.12%.

F By-phenomenon Results and Analyses

AltQ The multi-lingual LMs either prefer the sentences with *ma* or perform near chance. Although the mono-lingual LMs perform better, only `bert-base-zh` and `ernie` have an accuracy higher than 90%. There can be multiple reasons for the unsatisfactory performance. First, *haishi* is multi-functional, which might cause the LMs being unsure of its disjunctor usage. Second, *ma* only occurs in interrogative contexts, which can make the LMs prefer having it. Third, the LMs do not have a global view of the sentences but only attend to parts of them, which can be the reason of their random guessing.[31]

Anaphor (Gender) The LMs are gender biased. Figure 9 shows that, with a male subject, only four mono-lingual LMs (`gpt2-zh`, `CPM`, `pert-base`, and `ernie`) are gender neutral. When the subject is female, all LMs are biased (see Figure 12). The mono-lingual LMs strongly prefer a female object.

On one hand because the LMs are strongly biased, using the female gender to test the anaphor phenomenon is inconclusive. Compare Figure 13 to Figure 12, it is unclear whether the LMs achieved a high accuracy because they knew *ziji* or just because they liked the female feature. The male self paradigm, on the other hand, shows that most mono-lingual LMs were able to use *ziji* as a hint to agree the gender of the subject and object. Among the multi-lingual LMs, only `gpt3-davinci` achieved a meaningful accuracy increase.

Turning to the female self with `PP` paradigm in Figure 14, even thought the mono-lingual LMs prefer the female feature in the baseline, when there is a male distractor in the `PP` which is linearly closer to the reflexive, the LMs are affected, reflected as a decrease in the accuracy. Fewer multi-lingual LMs are affected by the distractor. As a matter of fact, `XLM-large` and `ByT5-small` even have an increase in accuracy. On the male self with `PP` paradigm, only the `mengzi` models and `gpt3-davinci` are relatively unaffected by the distractor.

Anaphor (Number) The plural number feature is used to elicit the anaphor agreement. The feature is imposed on the subject by using numeral + classifier or the plural marker *men*, or both. The plural feature on the object reflexive is reflected by adding *men* to it. As it turns out, the number feature is not a good choice because most LMs are strongly biased (see Table 8).

Aspect Compared to *le*, *guo* has a fixed position in a `VP` and cannot take a wide scope over the progressive marker *zai*. The results show that the

[31] The *A haishi B* disjunction and *ma* being at the end of a question are both locally grammatical.

LMs performed better on the *guo* paradigms than on *le*. There is no obvious reason why CPM in Figure 18 performs extremely bad.

Classifier-noun agreement The first paradigm tested the LMs' knowledge of the relative order of a demonstrative and classifier. Figure 20 shows that, except for the `CPM`, `PanGu-`α, `mt5`, and `ByT5` models, all LMs' accuracy are comparable to the human annotators.

Comparing the paradigms with simple nouns (Figure 21 and 22) to the ones with compound nouns (Figure 23 and 24), the multi-lingual models are more severely affected by the existence of a distractor (i.e., `noun1` in a compound noun) than the mono-lingual ones. The LMs are less affected by the distance created by the long adjective (Figure 21 *vs.* Figure 22, and Figure 23 *vs.* Figure 24).

Definiteness Effect Except for `CPM`, `PanGu-`α and `pert-large`, all mono-lingual models have a decent accuracy. On the multi-lingual side, the `ByT5` models are especially bad.

Polarity item Among the three PIs, *huoduo huoshao* (more or less) reliably occurs only in affirmative contexts. The negative PIs, *renhe* (any) and *shenme* (what), can occur in negative, interrogative, and affirmative contexts. Fifteen out of eighteen LMs reached an accuracy on *huoduo huoshao* comparable or even better than human. On the other two PIs, although there are quite a few LMs perform even better than human, overall, the accuracy values are worse and uneven.

Relative clause In the resumptive noun paradigm, only `CPM` and `pert-large` have a satisfying performance. The other models are either near chance (`lstm` and `mt5-small`) or strongly deviated by the repeated filler in the gap position. The reason could be that the LMs are vulnerable to repetition, or to local grammaticality. When the gap in the relative clause is filled by a pronoun that matches the gender of the head noun, fewer than half of the LMs are able to notice the minimal pair contrast.

***Wh*-fronting** All mono-lingual models performed well. Probably because *wh* in situ is a prominent feature of Mandarin Chinese. Except for the `mt5` and `ByT5` models, most multi-lingual models did well. The `gpt3-davinci` model even reaches a 100% accuracy.

G Results

G.1 CLiMP

The results are reported in Table 7 and Figure 6.

G.2 SLING

The results are reported in Table 8 and Figure 7 to Figure 33.

G.3 Statistic Tests

The results are reported in Table 9 to Table 12.

	lstm	gpt2-zh	CPM	PanGu	bert-base-zh	pert-base	pert-large	mengzi-base	mengzi-base-fin	ernie	xlm-R-base	xlm-R-large	bert-base-multi	mt5-small	mt5-large	byt5-small	byt5-large	gpt3
anaphor_agreement_gender_1000	82.6	79.5	79.9	92.6	86.2	90.5	71.1	96.1	96.2	93.7	82.1	78.0	73.0	46.2	69.3	55.4	49.4	83.3
binding_gender_1000.csv	49.1	45.1	51.3	61.2	50.8	51.5	39.6	64.8	64.0	54.7	48.4	50.6	44.4	51.7	44.7	51.7	51.6	47.1
ba_construction_1000	51.2	72.0	59.3	19.2	69.0	69.1	73.3	59.0	68.0	70.4	73.3	71.1	55.4	34.6	49.3	80.0	64.5	70.9
classifier_1000.csv	90.8	95.1	57.1	76.0	95.6	95.4	78.8	89.3	90.2	96.5	85.6	90.8	87.8	58.6	77.4	49.9	51.7	93.1
classifier_adj_1000.csv	80.3	91.9	55.5	69.1	93.2	94.3	76.9	90.4	90.7	95.8	81.1	88.0	84.7	58.4	74.1	50.6	50.7	88.3
classifier_clause_1000.csv	71.9	84.6	52.2	66.5	90.0	93.2	77.4	86.3	85.4	92.6	77.7	83.2	81.7	61.4	70.9	49.9	51.2	97.6
coverb_instrument_1000.csv	62.7	82.7	36.0	54.1	91.1	97.3	63.9	92.6	93.8	96.3	89.3	90.4	60.0	52.0	80.7	54.9	55.7	87.6
coverb_with_1000.csv	78.0	78.3	61.7	73.5	84.7	88.6	73.3	88.6	86.0	88.5	85.0	88.3	76.7	81.8	82.8	56.7	48.3	84.7
filler_gap_dependency_1000.csv	79.1	86.7	62.3	91.9	62.4	80.2	90.9	86.3	82.7	70.1	67.9	60.3	78.2	80.3	46.0	62.3	63.3	68.2
head_final_clause_1000.csv	68.3	77.0	86.5	65.6	53.1	83.9	73.3	82.5	78.9	78.0	76.2	87.1	72.0	85.2	85.8	43.6	60.6	73.0
passive_formal_1000.csv	69.2	61.6	47.0	61.6	67.7	67.3	44.0	46.4	47.1	68.7	55.0	48.1	73.2	57.3	51.4	54.2	52.4	54.5
verb_complement_direction_1000.csv	67.0	75.2	81.4	80.1	93.0	91.4	85.9	83.3	89.2	71.6	90.5	88.4	38.5	50.7	55.2	42.7	56.1	73.2
verb_complement_duration_1000.csv	96.1	99.1	83.6	82.6	90.2	96.4	89.1	98.4	96.8	94.1	86.4	90.4	76.3	64.6	51.0	12.7	18.9	55.4
verb_complement_frequency_1000.csv	98.5	99.2	48.8	75.6	97.8	91.5	78.7	75.9	75.0	87.5	23.6	21.5	90.9	69.8	71.4	44.2	32.5	96.0
verb_complement_res_adj_1000.csv	82.9	87.5	25.9	59.3	87.6	87.0	49.3	85.5	84.2	92.5	90.2	91.6	64.4	71.9	88.0	74.9	74.2	79.3
verb_complement_res_verb_1000.csv	99.4	98.5	96.7	90.1	96.2	88.8	68.9	85.9	87.2	92.3	53.6	66.1	92.4	65.0	78.6	27.5	33.2	97.0
Average over 8 phenomena	71.7	77.8	61.5	65.9	74.3	81.2	69.7	77.5	77.7	79.6	71.9	72.4	70.4	62.1	64.3	55.0	54.7	73.9
Std-dev over 8 phenomena	11.4	11.9	12.5	21.2	15.0	11.1	14.3	16.0	14.2	10.6	10.0	14.8	9.6	16.3	15.4	12.2	7.4	12.2

Table 7: Eighteen LMs' performance on CLiMP.

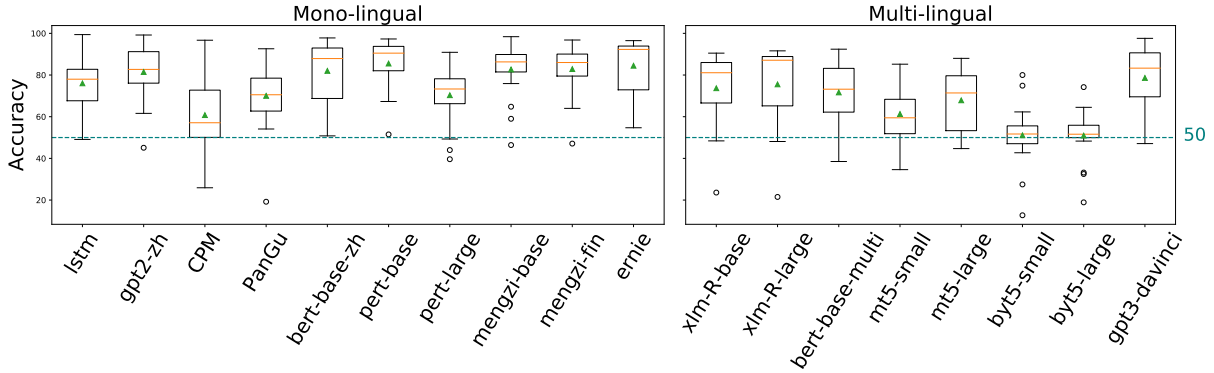

Figure 6: The box represents the inter-quartile range of the human and LM accuracy, with an orange line at the median accuracy and a green triangle at the mean. The whiskers extend from the box by 1.5 times. Dots are the accuracy values that past the end of the whiskers.

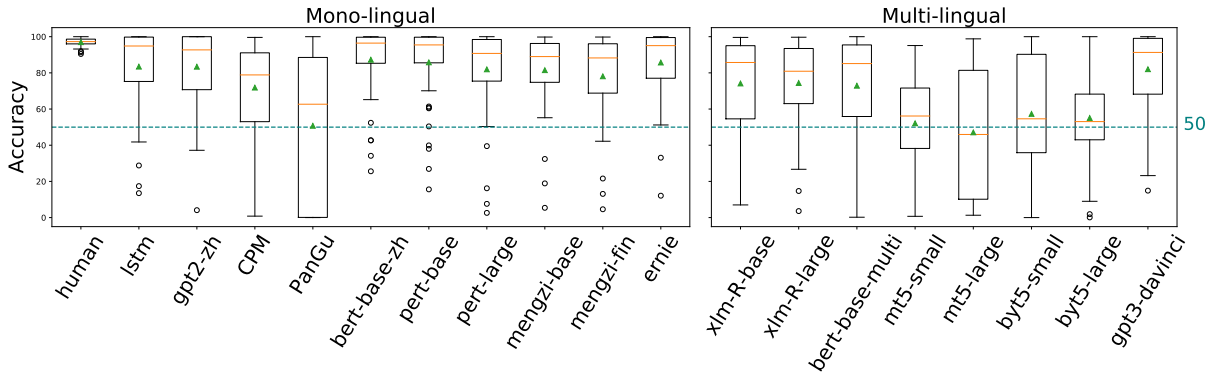

Figure 7: The box represents the inter-quartile range of the human and LM accuracy, with an orange line at the median accuracy and a green triangle at the mean. The whiskers extend from the box by 1.5 times. Dots are the accuracy values that past the end of the whiskers.

		human	lstm	gpt2-zh	CPM	PanGu	bert-base-zh	pert-base	pert-large	mengzi-base	mengzi-base-fin	ernie	xlm-R-base	xlm-R-large	bert-base-multi	mt5-small	mt5-large	byt5-small	byt5-large	gpt3
Alternative question	haishi	**97.3**	13.5	47.4	85.8	10.0	93.1	89.8	79.2	75.6	73.0	94.3	53.1	56.9	6.5	45.3	10.3	25.9	55.1	14.9
Anaphor (gender)	male_baseline		98.7	50.0	54.1	100	42.7	50.4	2.6	74.5	87.6	51.2	16.3	26.7	93.7	38.3	28.9	81.7	74.7	67.5
	male_self	**98.2**	99.6	87.6	57.7	100	92.6	80.9	7.6	99.6	98.9	88.9	28.3	61.9	99.6	10.5	46.9	48.6	59.2	91.6
	pp_male	**99.6**	64.9	37.2	41.3	99.9	75.7	39.9	16.2	91.4	91.6	71.8	13.9	14.7	71.5	37.4	25.3	34.3	28.7	89.6
	female_baseline		92.8	80.2	91.2	95.9	85.6	85.3	94.9	89.3	91.2	96.4	90.5	62.8	82.8	64.0	39.2	40.6	48.0	40.8
	female_self	**97.7**	93.4	80.8	89.9	97.3	98.2	95.5	93.9	99.7	98.1	99.7	98.2	71.0	82.7	84.6	47.4	61.0	44.0	48.5
	pp_female	**98.6**	41.8	64.5	95.4	98.6	86.7	26.9	83.5	78.1	68.6	65.8	97.5	96.2	76.1	70.0	31.1	71.6	18.9	23.2
Anaphor (number)	cl_male_baseline		100	100	76.4	0.0	99.7	100	97.7	82.3	67.8	100	94.3	81.7	99.7	80.6	7.0	99.9	99.5	99.9
	cl_self_male	**95.5**	99.5	100	98.7	0.1	99.7	100	99.3	78.5	69.4	99.8	98.5	88.1	97.3	68.1	5.0	99.4	61.7	99.8
	cl_female_baseline		99.5	100	80.6	0.0	99.6	99.9	91.2	80.0	65.8	99.2	67.5	42.3	98.5	19.6	5.9	91.8	65.2	99.0
	cl_self_female	**97.3**	98.7	100	98.6	0.0	100	99.6	98.9	72.6	83.4	99.6	94.3	69.6	89.6	2.1	8.1	63.1	42.9	99.1
	men_male_baseline		100	100	48.4	0.1	99.8	100	98.3	66.3	43.5	99.6	84.4	77.5	100	45.1	5.8	98.7	100	99.8
	menself_male	**97.3**	100	100	79.9	0.0	99.9	100	100	94.1	85.8	99.7	98.4	85.2	99.5	52.3	5.2	100	77.4	100
	men_female_baseline		100	100	50.5	0.0	98.0	99.9	96.6	65.4	48.1	97.6	54.4	63.6	99.6	7.1	10.1	99.4	99.9	96.3
	menself_female	**95.9**	100	100	80.9	0.0	99.8	100	98.9	95.0	93.0	98.6	90.5	85.8	89.4	2.2	14.2	99.9	99.4	99.9
	cl_men_male_baseline		100	100	86.0	0.0	99.9	100	99.1	73.1	49.5	100	88.1	66.7	100	43.3	1.3	99.9	99.9	98.8
	cl_menself_male	**95.9**	99.7	100	99.6	0.0	100	100	99.6	85.4	74.2	99.9	98.4	92.8	100	38.2	2.5	99.8	64.2	100
	cl_men_female_baseline		100	100	88.8	0.0	99.4	99.5	96.3	59.7	42.2	99.9	50.8	39.5	100	4.5	2.9	98.7	99.2	94.3
	cl_menself_female	**97.3**	99.8	100	96.2	0.0	100	99.4	95.9	56.3	47.0	98.1	91.3	89.6	92.4	0.7	8.5	99.2	92.5	99.9
Classifier-noun agreement	dem_cl_swap	**99.6**	99.6	99.8	52.5	85.7	99.8	99.5	92.3	94.2	92.2	99.0	92.8	94.1	98.9	78.5	81.4	63.0	57.5	98.3
	cl_simple_noun	**98.6**	95.6	96.7	61.2	85.0	98.5	98.4	88	96.4	96.6	95.9	94.5	95.9	92.4	77.9	90.4	50.1	53.1	96.3
	cl_adj_simple_noun	**93.2**	92.1	95.5	58.9	77.1	96.5	96.5	83.8	95.8	96.1	95.2	93.6	95.2	91.0	58.4	85.5	51.6	52.6	94.5
	cl_comp_noun	**94.0**	56.2	70.7	45.6	66.3	91.3	89.9	74.2	90.6	90.3	91.1	71.2	78.5	74.7	63.3	83.6	51.8	48.6	78.3
	cl_adj_comp_noun	**96.5**	56.2	65.6	45.1	59.9	90.3	90.1	72.9	92.7	92.1	90.9	83.3	87.4	80.4	62.1	80.8	46.8	53.3	77.9
Aspect	past_tense_guo	**99.1**	83.9	85.6	79.7	72.4	95.5	91.4	60.5	88.7	87.6	92.7	79.6	81.7	54.5	48.6	71.6	58.8	45.3	85.8
	zai_guo	**98.2**	55.6	88.2	78.6	65.4	97.9	98.2	90.3	87.3	88.9	97.7	61.3	84.5	65.7	67.3	89.5	49.6	54.2	91.0
	past_tense_le	**99.1**	76.2	70.7	78.8	73.9	65.2	61.4	39.5	81.0	77.3	51.6	49.5	64.6	37.4	31.2	37.9	53.9	43.0	57.7
	zai_no_le	**95.0**	17.4	50.8	0.8	16.1	85.2	86.7	72.8	55.2	62.5	75.6	31.9	51.4	44.6	56.2	81.4	53.9	43.0	63.1
	zai_le_scope	**96.4**	28.8	64.1	68.0	51.4	76.9	70.1	79.1	69.5	75.4	53.7	48.2	62.1	22.8	45.6	45.0	60.3	68.8	60.0
Definiteness effect	demonstrative	**96.8**	94.1	99.3	48.3	49.3	98.2	98.2	82.4	97.4	96.5	94.9	55.1	65.4	92.5	59.8	25.5	27.8	16.1	70.4
	every	**96.8**	99.8	99.5	92.5	87.7	94.6	92.6	65.3	95.8	95.6	82.4	71.9	80.1	95.7	84.5	72.5	0.6	1.9	92.6
Polarity item	any	**90.5**	87.6	89.9	95.9	93.6	65.8	86.3	94.9	97.4	97.6	75.6	93.2	95.0	33.8	61.8	83.7	60.0	45.2	63.8
	even_wh	**91.4**	85.3	70.3	42.3	47.7	52.4	87.4	99.1	99.5	99.5	70.8	99.1	99.6	7.1	77.6	97.4	33.0	66.6	96.2
	more_or_less	**94.1**	98.0	97.7	98.6	97.6	97.9	97.5	90.0	96.9	97.6	97.6	97.3	94.9	91.8	95.1	63.8	85.6	77.0	97.7
Relative clause	resumptive noun	**100**	50.9	4.1	82.1	16.7	25.6	15.6	98.5	5.4	4.6	12.1	7.0	3.6	0.2	56.1	26.1	0.0	0.1	39.4
	resumptive pronoun	**98.2**	93.2	85.7	18.6	11.8	42.7	60.4	80.0	32.4	21.6	54.1	80.3	93.7	26.2	28.3	74.2	5.5	36.5	90.9
wh fronting	bare_wh	**100**	100	99.9	96.6	99.7	100	100	99.7	99.7	98.9	99.6	99.6	99.7	75.6	86.1	98.4	7.0	36.6	100
	wh_as_modifier	**100**	100	99.4	90.7	88.8	99.5	99.5	99.4	99.8	99.8	99.9	95.2	99.0	60.0	76.1	98.8	19.2	52.8	100
Average over 9 phenomena		**97.1**	75.5	78.0	72.9	55.1	84.8	83.4	81.8	81.3	79.6	83.0	72.2	75.5	59.5	57.2	53.8	41.2	45.0	75.0

Table 8: Eighteen LMs' performance on SLING. The blue marked lines are baselines. The baselines are supposed to have an accuracy of 50%, meaning the LMs are gender/number neutral.

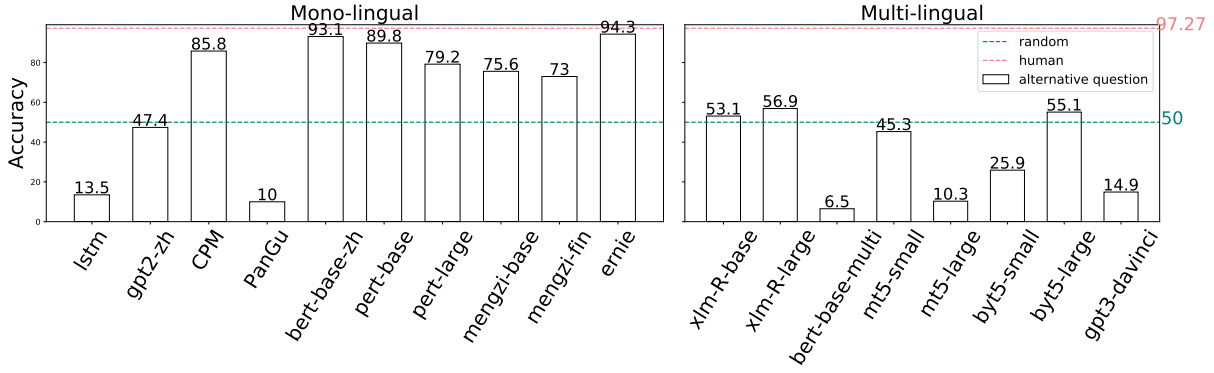

Figure 8: The LM accuracy on the alternative question phenomenon.

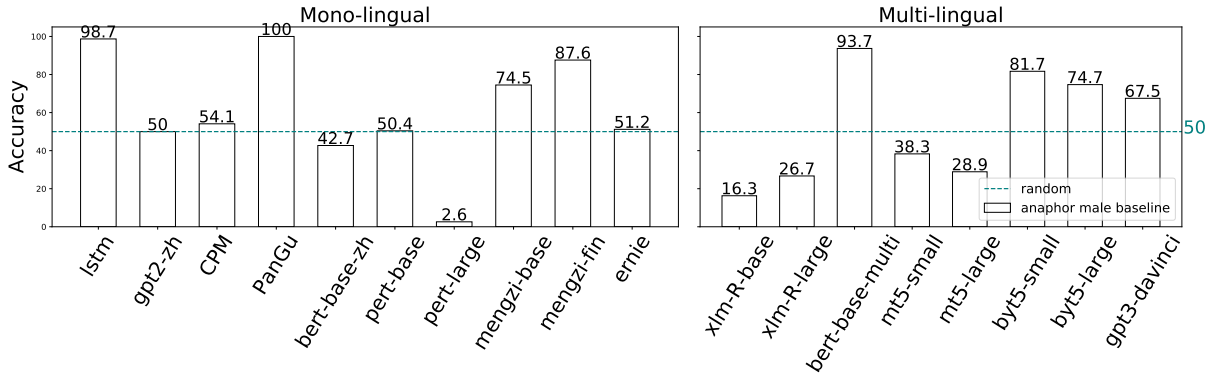

Figure 9: The LM bias towards a male object when the subject is male.

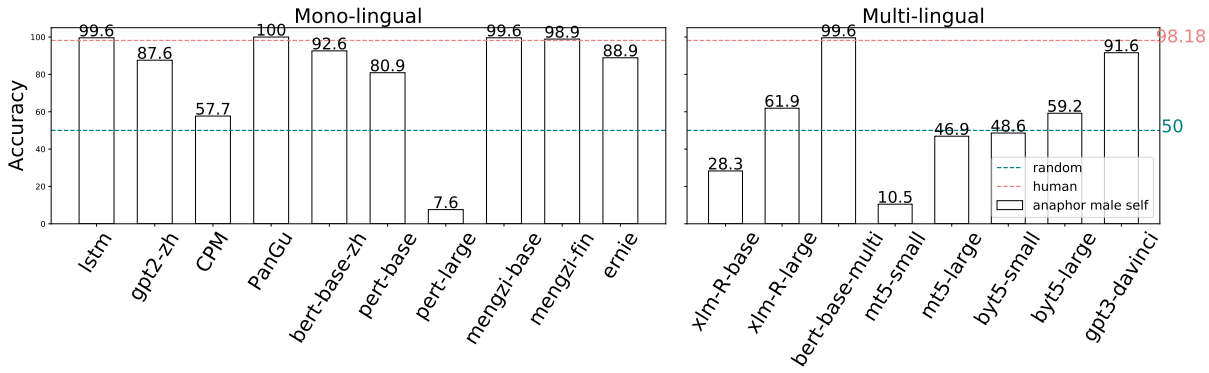

Figure 10: The LM accuracy on the anaphor male self paradigm.

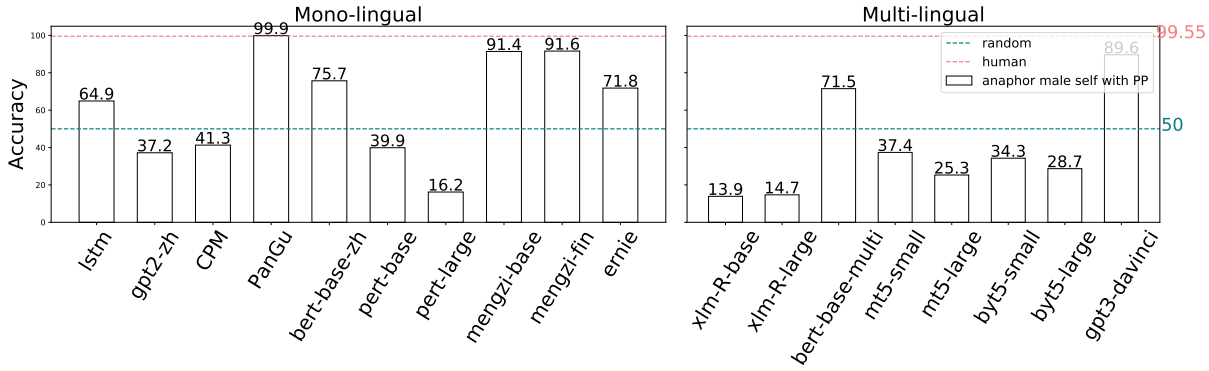

Figure 11: The LM accuracy on the anaphor male self with PP paradigm.

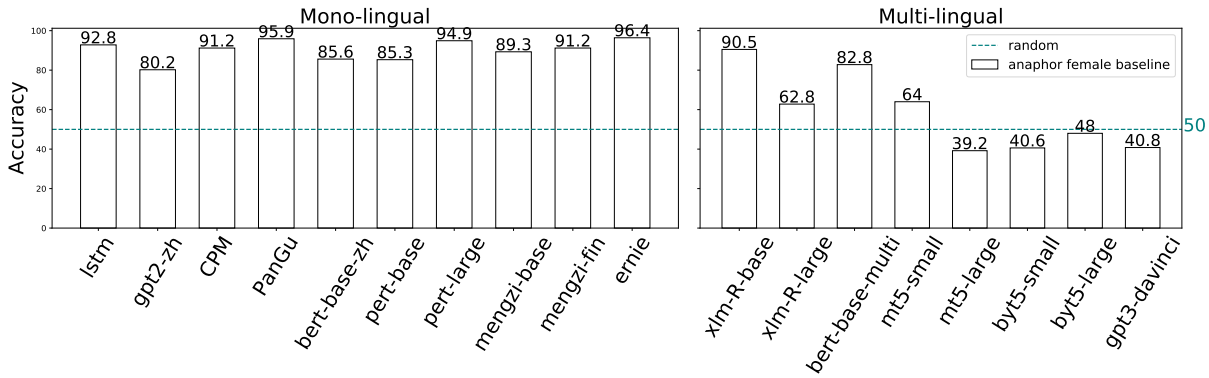

Figure 12: The LM bias towards a female object when the subject is female.

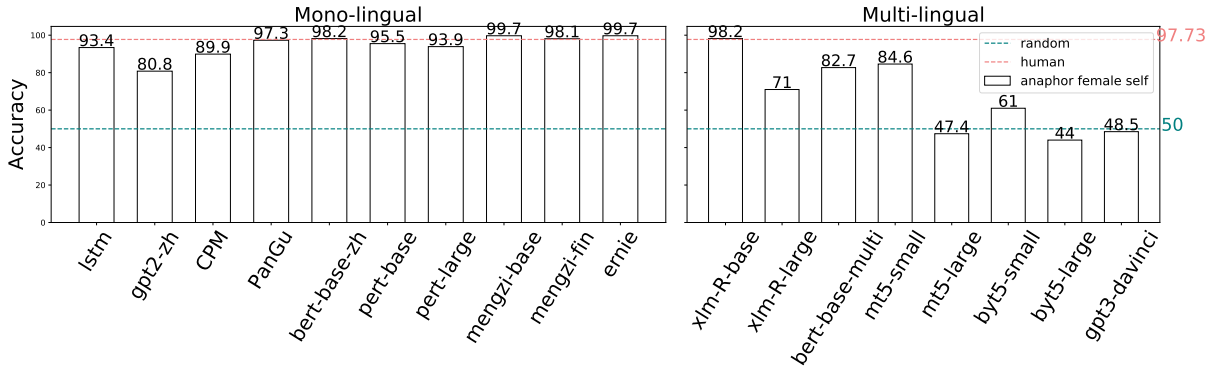

Figure 13: The LM accuracy on the anaphor female self paradigm.

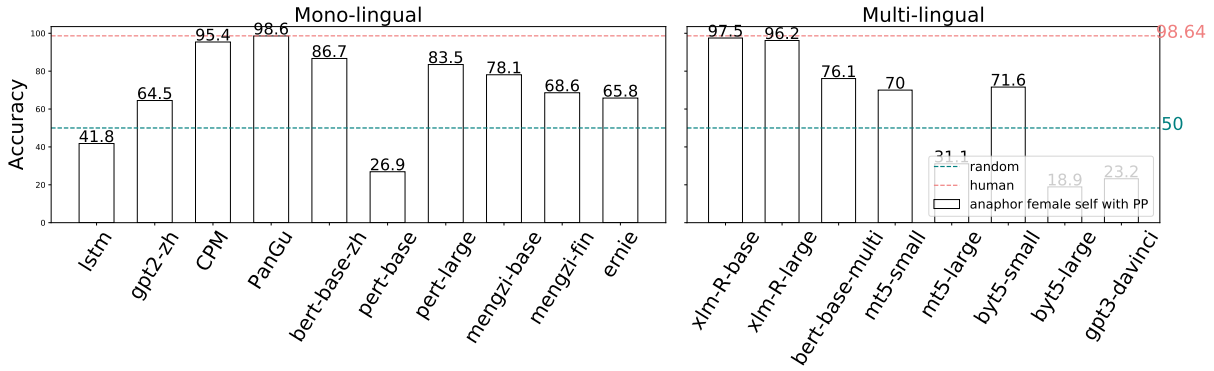

Figure 14: The LM accuracy on the anaphor female self with PP paradigm.

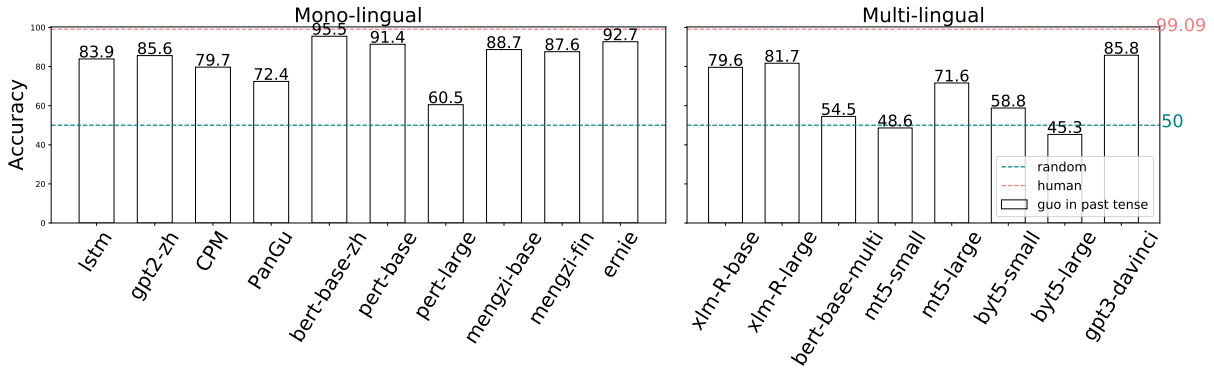

Figure 15: The LM accuracy on the guo in past tense paradigm.

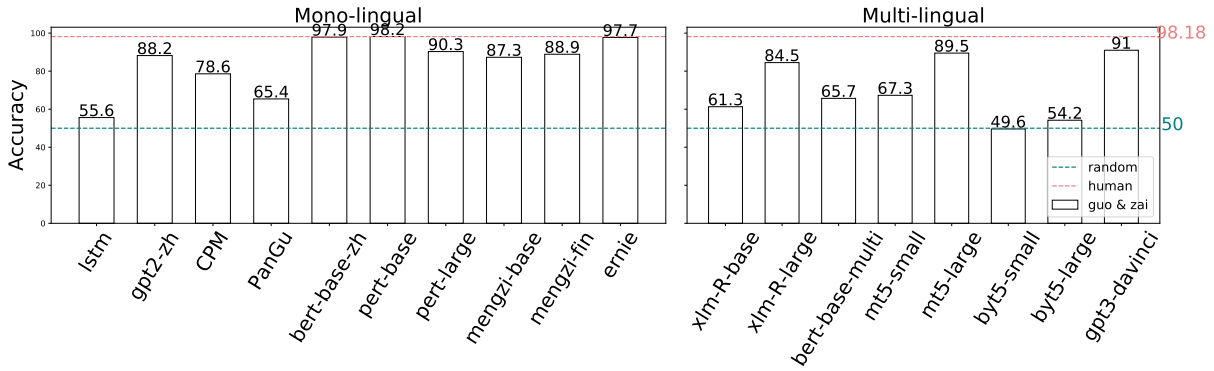

Figure 16: The LM accuracy on the guo & zai paradigm.

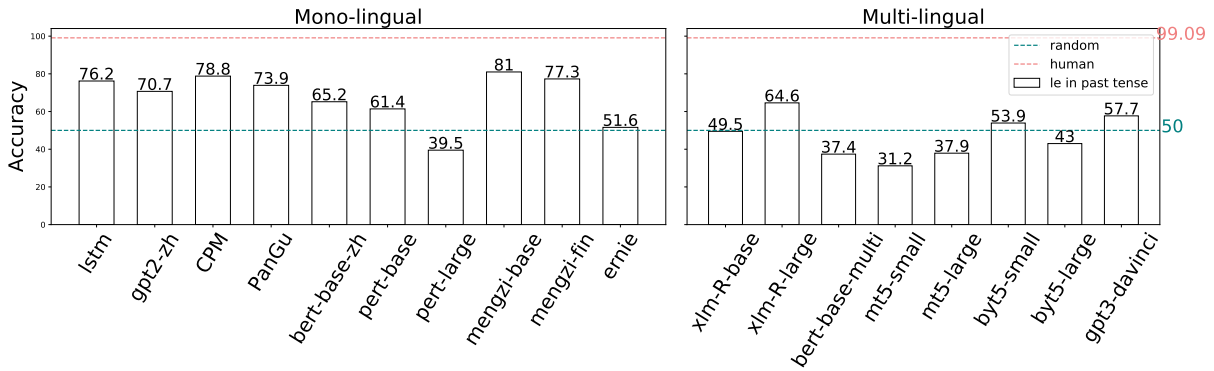

Figure 17: The LM accuracy on the le in past tense paradigm.

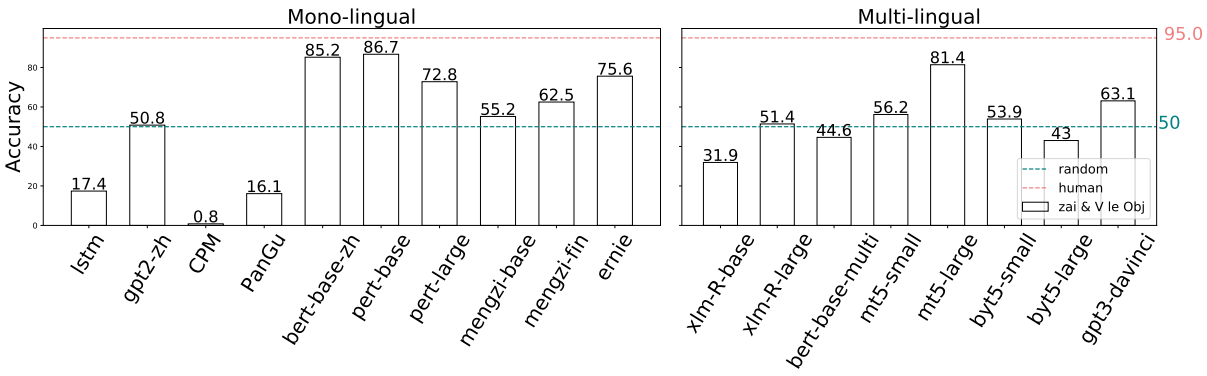

Figure 18: The LM accuracy on the zai & V le Obj paradigm.

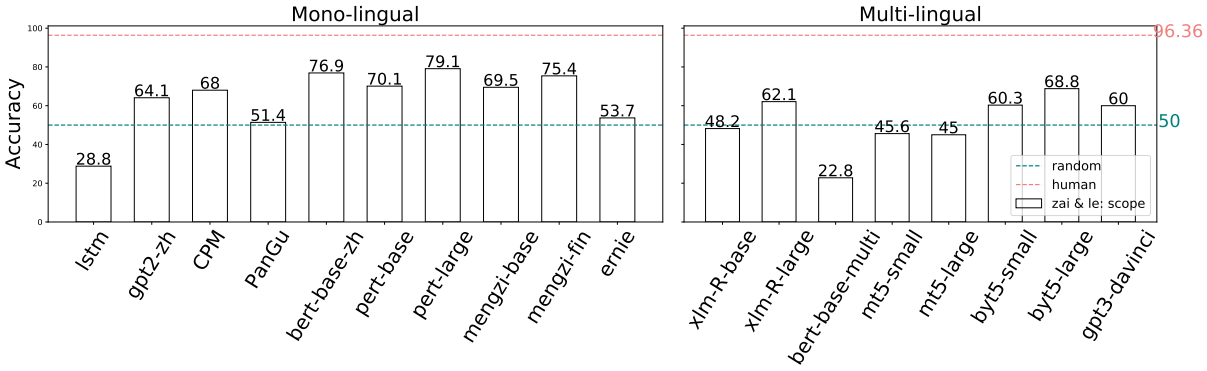

Figure 19: The LM accuracy on the zai & le scope paradigm.

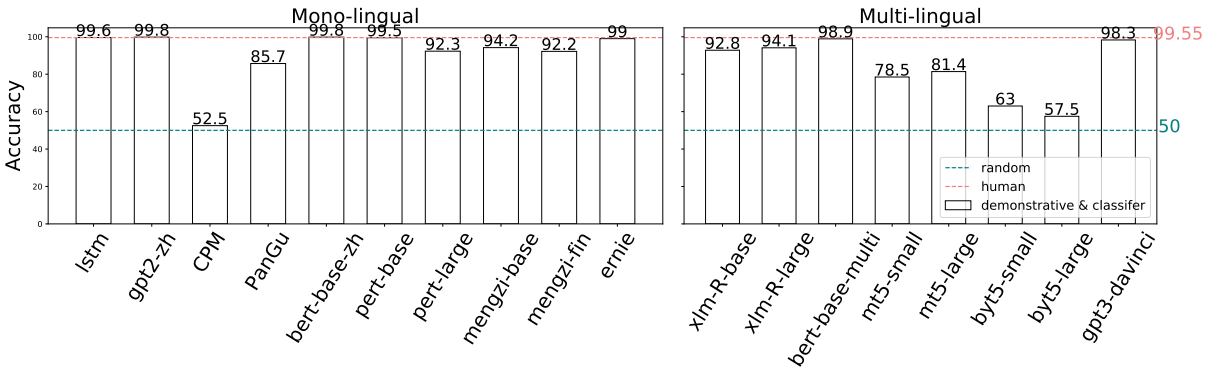

Figure 20: The LM accuracy on the demonstrative & classifier paradigm.

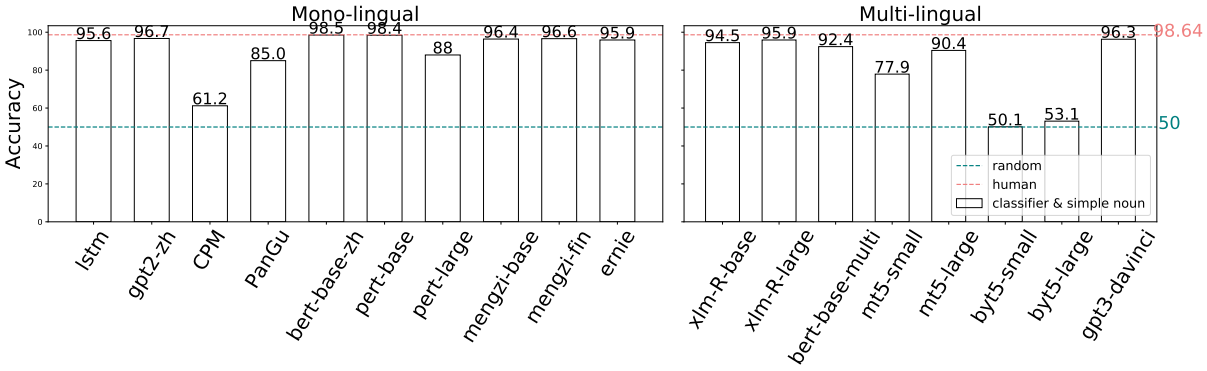

Figure 21: The LM accuracy on the classifier & simple noun paradigm.

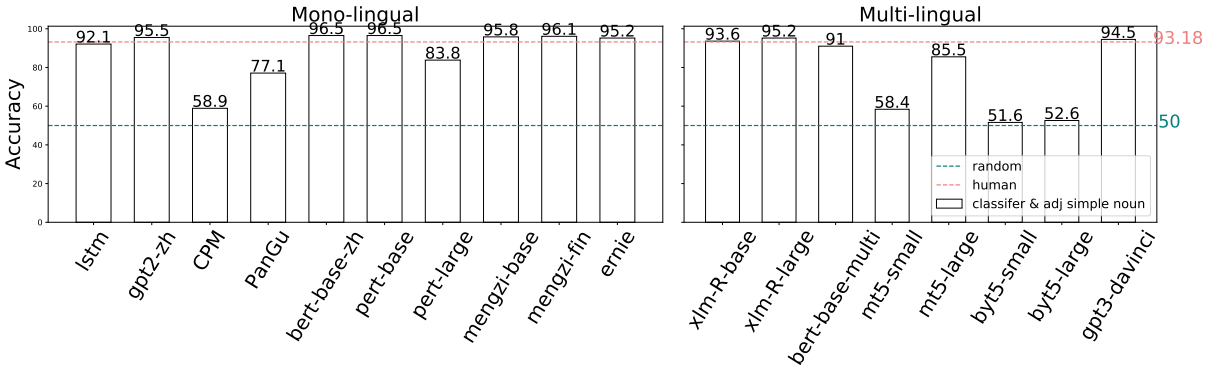

Figure 22: The LM accuracy on the classifier & adj. simple noun paradigm.

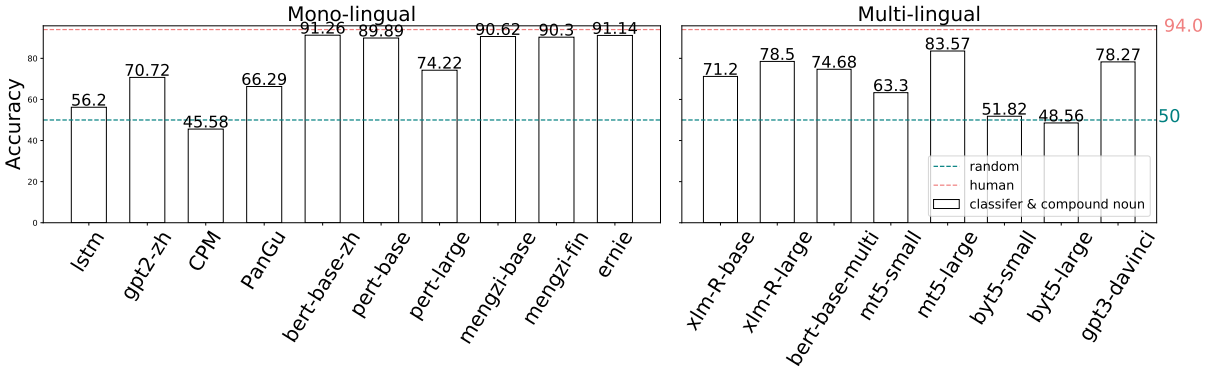

Figure 23: The LM accuracy on the classifier & compound noun paradigm.

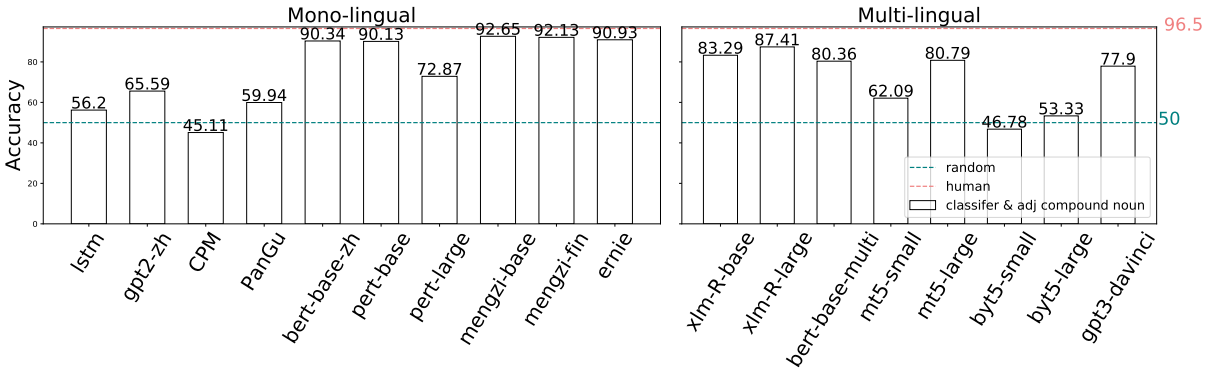

Figure 24: The LM accuracy on the classifier & adj compound noun paradigm.

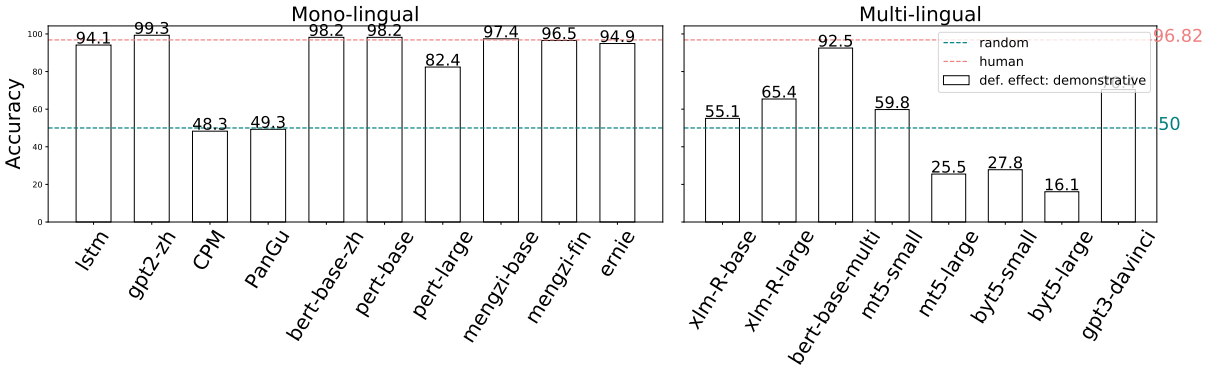

Figure 25: The LM accuracy on the definiteness effect with demonstrative paradigm.

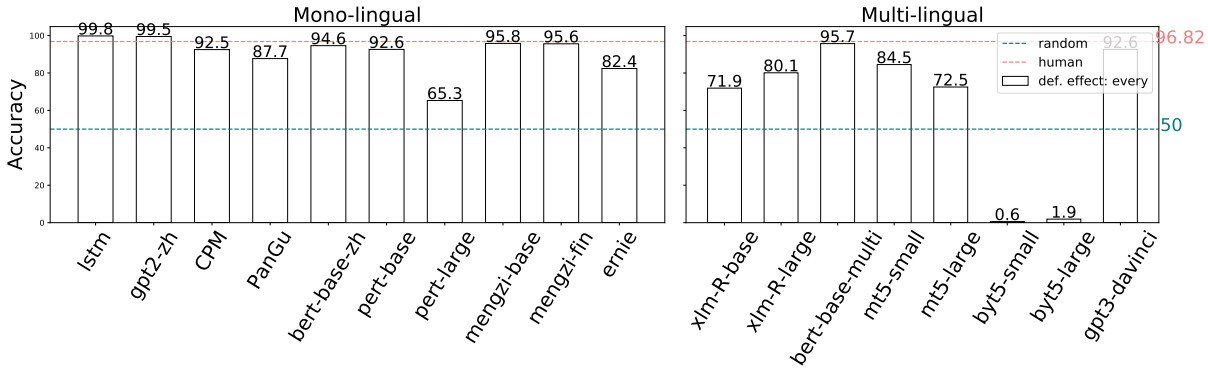

Figure 26: The LM accuracy on the definiteness effect with every paradigm.

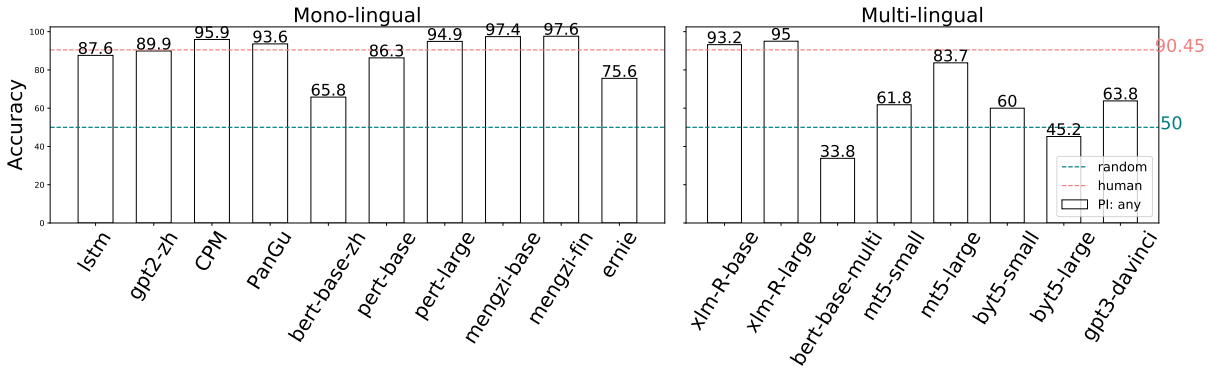

Figure 27: The LM accuracy on the polarity item any paradigm.

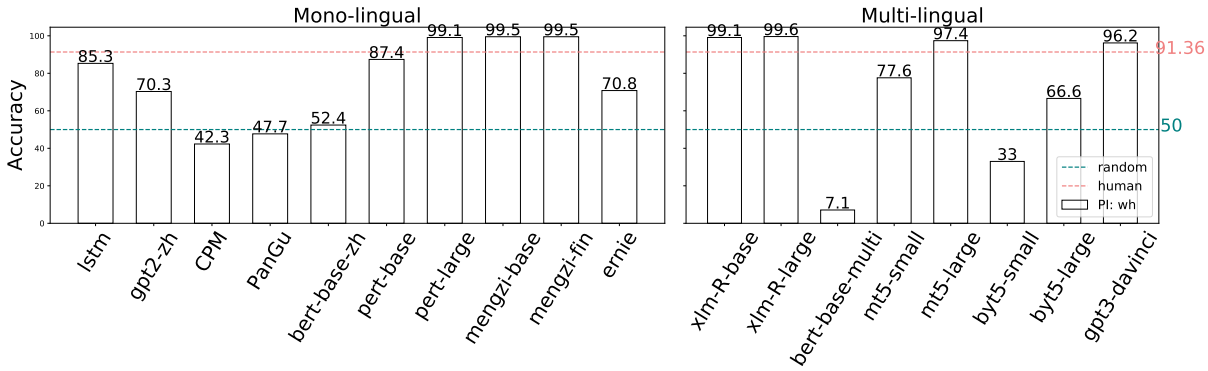

Figure 28: The LM accuracy on the polarity item wh paradigm.

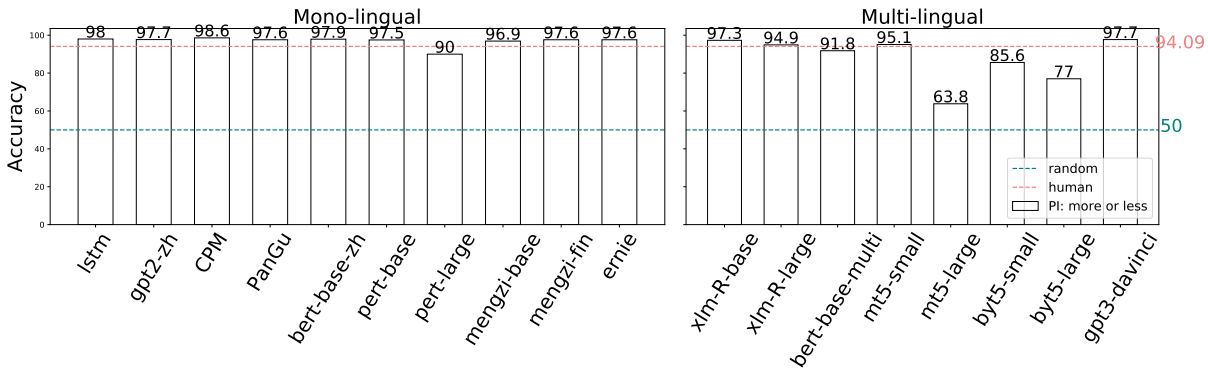

Figure 29: The LM accuracy on the polarity item more or less paradigm.

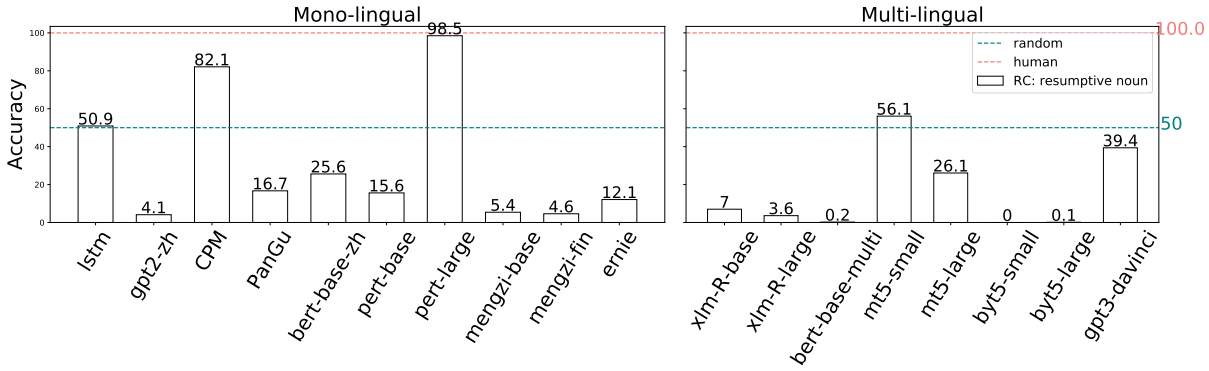

Figure 30: The LM accuracy on the relative clause with resumptive noun paradigm.

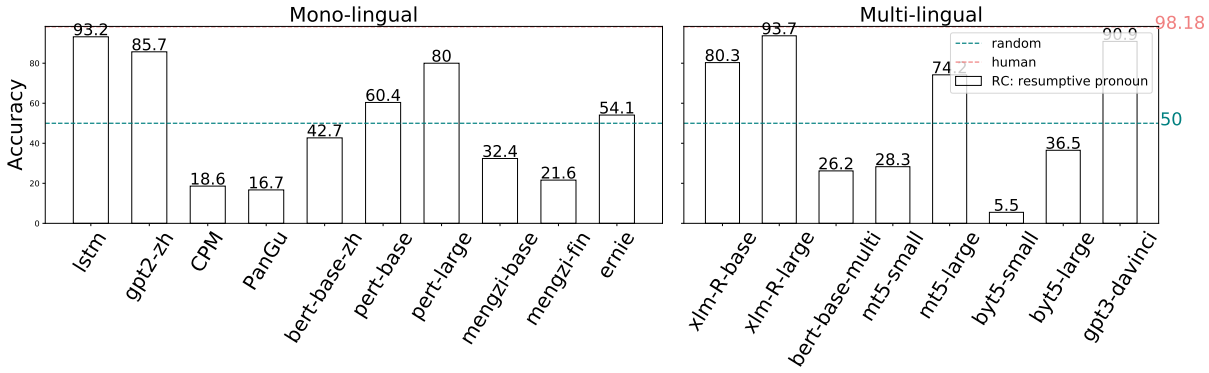

Figure 31: The LM accuracy on the relative clause with resumptive pronoun paradigm.

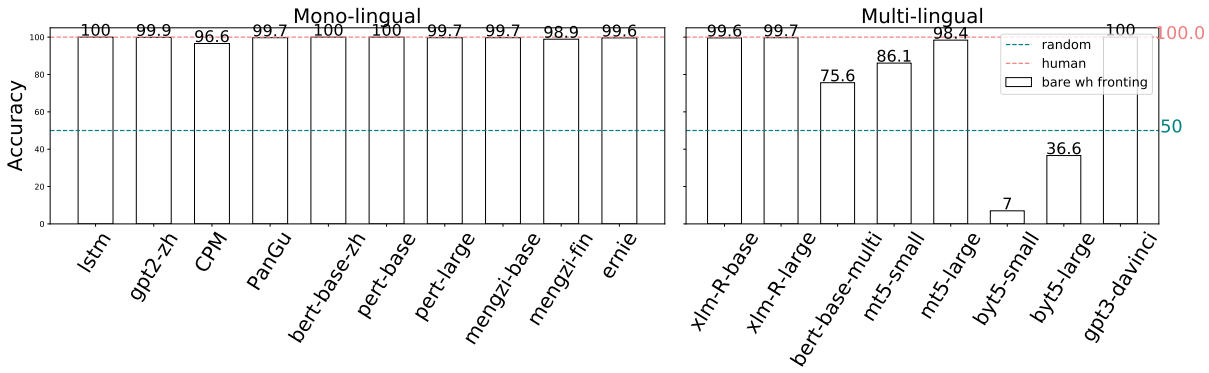

Figure 32: The LM accuracy on the bare wh fronting paradigm.

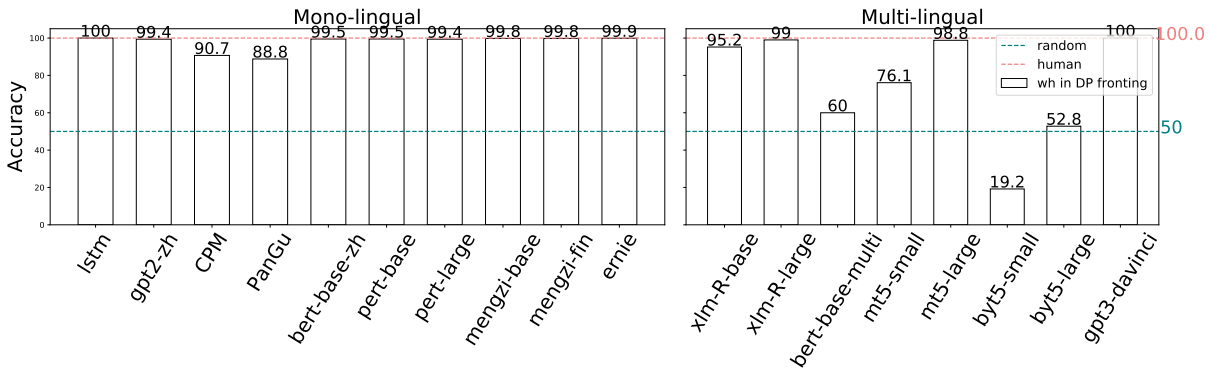

Figure 33: The LM accuracy on the wh in DP fronting paradigm.

LM1	LM2	two-tailed	greater	lesser
lstm	gpt2-zh	0.617	———	———
pert-base	pert-large	0.009**	0.005**	0.996
mengzi-base	mengzi-fin	0.004**	0.002**	0.998
xlm-R-base	xlm-R-large	0.913	———	———
mt5-small	mt5-large	0.293	———	———
byt5-small	byt5-large	0.277	———	———

Table 9: The *p* values of the Wilcoxon signed rank tests of LM pairs.

data	two-tailed	greater	lesser
simple & simple w/ adj.	0.000***	0.000***	1.000
compound & comp. w/ adj.	1	———	———

Table 10: The results of the Wilcoxon signed rank tests of the simple noun with/withouth a long adjective and the ones with compound nouns.

data	min.	median	mean	max.	SD	*p* value
male self	7.6	84.25	70	100	31.27	0.002**
male pp	13.9	40.6	52.52	99.9	29.42	
female self	44	91.65	82.44	99.7	19.54	0.008**
female pp	18.9	70.8	66.36	98.6	26.85	

Table 11: Descriptive statistics of the anaphor (fe)male self and (fe)male self with PP paradigms. The *p* values are from the Wilcoxon signed rank tests.

data	min.	median	mean	max.	SD	*p* value
simple	50.1	95.05	86.83	98.5	15.75	0.000***
compound	45.58	74.45	73.12	91.26	15.26	
simple w/ adj.	51.6	92.85	83.88	96.5	16.57	0.000***
comp. w/ adj.	45.11	79.13	73.77	92.65	16.43	

Table 12: Descriptive statistics of the classifier & (adj.) simple noun and compound noun paradigms. The *p* values are from the Wilcoxon signed rank tests.