Chinese Morpheme-informed Evaluation of Large Language Models

Yaqi Yin[1,2], Yue Wang[1,2], Yang Liu[1,2*]
[1]National Key Laboratory for Multimedia Information Processing, Peking University
[2]School of Computer Science, Peking University
yyqi@stu.pku.edu.cn, {wyy209, liuyang}@pku.edu.cn

Abstract

Previous evaluations of large language models (LLMs) focused on the perspective of various tasks or abilities. In this paper, we propose to evaluate from a linguistic viewpoint and argue that morpheme, a potential linguistic feature that captures both word-formation and lexical semantics, is another suitable component for evaluation that remains largely unexplored. In light of this, we construct MorphEval, a morpheme-informed benchmark, including three datasets following the bottom-up levels of characters, words, and sentences in Chinese, and then evaluate representative LLMs with both zero- and few-shot settings under two metrics. From this perspective, we reveal three aspects of issues LLMs nowadays encounter: dysfunctions in morphology and syntax, challenges with the long-tailed distribution of semantics, and difficulties from cultural implications. In these scenarios, even a smaller Chinese-targeted model may outperform ChatGPT, highlighting the actual challenges LLMs face and the necessity of language-specific improvements when applied to non-English languages. This new approach could also help guide model enhancements as well as get extended to other languages.

Keywords: Large Language Models, Evaluation, Chinese Morphemes

1. Introduction

Large language models (LLMs) such as GPT-4 (OpenAI, 2023) and LLaMA (Touvron et al., 2023) have demonstrated remarkable performance across a variety of NLP tasks, exceeding expectations in zero- and few-shot settings (Brown et al., 2020a). With the continuous development of LLMs, they have now been integrated into diverse real-world challenges (Sadik et al., 2023; Peeters and Bizer, 2023; Zhu et al., 2023). This growing reliance on LLMs makes it critical to conduct evaluations and explore the extent of their capabilities.

There are now two main approaches to evaluating LLMs. The first focuses on specific tasks and utilizes public data tailored for them, such as GLUE (Wang et al., 2020) and SQuAD (Rajpurkar et al., 2016). The second targets the challenging human-level abilities of LLMs. It often involves human exam questions from various domains and levels, such as MMLU (Hendrycks et al., 2020).

As one of the most widely used languages, Chinese has witnessed a significant increase in LLMs that exhibit impressive performance (Zeng et al., 2022; Sun et al., 2023). Chinese benchmarks primarily follow these two approaches as well. CLUE (Xu et al., 2020) exemplifies traditional task-oriented evaluations, while datasets such as AGIEval (Zhong et al., 2023) offer a human-centered alternative assessment.

However, we contend that prior evaluations may pose potential risks. Firstly, LLMs mistake in a task could be caused by both a lack of task ability and improper language comprehension, especially in non-English languages (Zhang et al., 2024; Robinson et al., 2023; Bang et al., 2023; Jiao et al., 2023). Second, previous assessments heavily rely on data from human interactions, which tend to be dominated by more common words (Zipf, 1949). This can lead to biased and incomplete judgments, as it neglects the less frequent aspects of language.

Task	**Translate the sentence into English.**
Ex.1	这个天气干打雷。
Pred.	The weather is **dry** and thundering. (✗)
GT	The weather is **just** thundering. (✓)
Ex.2	这个天气干打雷，不下雨。
Pred.	The weather is so **dry** and thundering, but it's not raining. (✗)
GT	The weather is **just** thundering, but it's not raining. (✓)
Ex.3	别干着急了。
Pred.	Don't be in a hurry. (✔)
GT	Don't be in a hurry **in vain**. (✓)

Table 1: An example where ChatGPT answers incorrectly in the machine translation task.

As an illustration, consider the Chinese character "干" *(futile)* in the machine translation task shown in Table 1. Despite specific context, such as "不下雨" *(it's not raining)*, ChatGPT[1] (Brown et al., 2020b) still

[*]corresponding author

Our code and resources are available at https://github.com/COOLPKU/MorphEval.

[1]In this study, the employed ChatGPT models are ChatGPT-3.5-turbo-0613, referred to as ChatGPT.

LREC-COLING 2024, pages 3165–3178
20-25 May, 2024.

Example	Linguistic Unit	Target Text	Model Prediction	
Ex.2	Sentence	这个天气干打雷	The weather is **dry** and thundering.	✗
	Word	这个	This.	✓
		天气	The weather.	✓
		干打雷	Dry and thundering.	✗
	Character	干	Very; Extremely	✗
		打雷	Thundering.	✓
Ex.3	Sentence	别干着急了	Don't be in a hurry	✓✗
	Word	别	Don't; stop.	✓
		干着急	Anxious; in a hurry.	✓✗
	Character	干	Do something; engage in something.	✗
		着急	Anxious; worried.	✓

Table 2: Exploration of the causes of ChatGPT's erroneous translations. The queried prompt is "请解释*{context}*中*{target text}*的意思，并把*{target text}*翻译成英文。" *(Please explain the meaning of {target text} in the context of {context}, and translate {target text} into English.)*

Ex.2	Para. 1	Translate the following sentence into English: 这个天气只打雷。
	Pred.	The weather is **only** thundering. (✓)
	Para. 2	Given that in the phrase "干打雷", the character "干" means "futile". Translate the following sentence into English: 这个天气干打雷。
	Pred.	This weather is thundering **in vain**. (✓)
Ex.3	Para. 1	Translate the following sentence into English: 别徒劳地着急了。
	Pred.	Don't worry **in vain**. (✓)
	Para. 2	Given that in the phrase "干着急", the character "干" means "futile". Translate the following sentence into English: 别干着急了。
	Pred.	Don't be **in vain** hurry. (✓)

Table 3: Verification of the causes of ChatGPT's erroneous translations.

erroneously translates "这个天气干打雷" to "*the weather is so dry and thundering*". This demonstrates that while contextual cues may assist the model in some instances, its susceptibility to errors due to insufficient language comprehension persists, thereby affecting task performance. Furthermore, in Ex. 3, the model translates "别干着急了" *(don't be in a hurry in vain)* as "*don't be in a hurry*," which could be considered correct in a machine translation task. However, a closer examination reveals that it captures only a portion of the intended meaning without conveying "in vain". This highlights the limitations of relying solely on overall indicators and upper-level language units to detect such errors. Given the increasing use of LLMs in real-world applications, it is crucial to identify these issues in order to further stabilize and optimize their performance.

To elucidate the origins of such errors, we prompt the model to generate definitions for target texts within contexts. This exploration follows the coarse-to-fine linguistic levels of "sentence-word-character". The results in Table 2 show that the model gives wrong explanations for the word "干打雷" and character "干" in Ex. 2, along with a similar issue in Ex. 3. Therefore, it could be concluded that the model's lack of understanding of these words and characters led to the incorrect translation of these sentences. The verification in Table 3 further substantiates this conclusion, as the model could translate correctly, provided the character is paraphrased or explained.

To address these risks, we advance LLM evaluation from a linguistic perspective and introduce morphemes as the basis. Similar to or slightly different from that in languages such as English, a Chinese morpheme is the smallest semantic and sound-bearing component (Zhu, 1982), representing different usages and meanings of characters (Lv, 1979). Table 4 shows different morphemes of character "干", as well as their usages in word-formation and sentence-making. The causes of wrong trans-

Char	Morpheme	PoS	Morpheme Sense	Word	Sentence
干	干$_1$	动 v.	做事情 doing something	干活 working	他一整天都在认真干活 He has been working hard all day.
	干$_2$	副 adv.	无效；徒劳 in vain	干着急 worry in vain	别干着急了。 Don't worry in vain.
	干$_3$	名 n.	古代指盾牌 shield	干戈 symbol of war	这个国家已经有几十年没有经历过干戈的洗礼了。 This country has not experienced war for decades.

Table 4: Different morphemes, word-formation, and sentence-making for the character "干".

lations we observed on "干" are actually due to the same morpheme "干$_2$", which proves the sufficiency of a morpheme-based evaluation.

In general, this new approach could benefit the evaluation in light of the following considerations: First, a morpheme-informed assessment can decompose language into its most basic components, allowing for the identification of the fundamental reasons underlining LLMs' linguistic misconceptions. Second, morphemes could effectively represent semantics (Zhang, 1997; Luo, 2013), as 14,291 morphemes, corresponding to 3,500 frequently-used characters, could cover 99.48% of a large-scale corpus from various regions (Fu, 1988). This ensures a high level of coverage over lexical semantics.

To meet this demand, we construct **MorphEval**, a Chinese **Morph**eme-informed **Eval**uation benchmark[2]. We collect data from a dictionary-based resource (Liu et al., 2018), emerging a well-designed and balanced distribution of semantics. Following the levels of characters, words, and sentences in the Chinese language, we build MorphEval into three datasets. We then evaluate typical English- and Chinese-targeted LLMs by testing the composition of language units through morphemes. This research could provide analyses and suggestions on morpheme usages with different parts-of-speech (PoS), long-tailed distribution of semantics, and cultural implications.

To sum up, our main contributions are as follows:

(1) We incorporate a linguistic perspective into LLMs evaluation and develop **MorphEval**, a large-scale Chinese morpheme-informed benchmark. This allows a bottom-up assessment that uncovers LLMs' potential defects.

(2) Our evaluation reveals three weaknesses neglected by previous work: dysfunctions in morphology and syntax, challenges with the long-tailed distribution of semantics, and difficulties from cultural implications. It also illustrates the need for language-specific improvements, as a smaller Chinese-targeted model, Alpaca-13B, for example, could even surpass ChatGPT in these scenarios.

[2]A carefully paraphrased version of MorphEval will be released subsequently.

Notably, this morpheme-informed evaluation approach can also apply to other languages, such as English, where morphemes contribute to word-formation and word meanings as well, similar to those in Chinese.

2. Related Work

2.1. Evaluations for LLMs

With the growing reliance on LLMs, a host of evaluations have emerged. Some benchmarks are designed towards advanced capabilities that only arise with increased model scales, such as reasoning (Cobbe et al., 2021), hard math problem-solving (Hendrycks et al., 2021), and coding (Chen et al., 2021), etc. Some other benchmarks aggregate a wide range of NLP tasks for exhaustive evaluation, including MMLU (Hendrycks et al., 2020), BIG-bench (Srivastava et al., 2022), and HELM (Liang et al., 2022).

For Chinese-targeted benchmarks, CLUE (Xu et al., 2020), the Chinese counterpart of GLUE, is widely used. AGIEval (Zhong et al., 2023), C-Eval (Huang et al., 2023) and CMMLU (Li et al., 2023a) evaluate models on human exams across diverse domains. Super-CLUE (Xu et al., 2023) incorporates evaluations of Chinese characters and cultural backgrounds, yet the datasets for these aspects remain relatively small and have not been made public yet. ZhuJiu (Zhang et al., 2023), following Super-CLUE, also includes evaluations for Chinese-specific abilities but are mainly on the level of words or higher. ACLUE (Zhang and Li, 2023) and TMMLU (Hsu et al., 2023), on the other hand, target evaluations for ancient or traditional Chinese.

2.2. Chinese Morpheme-Related Resources

Due to the powerful productivity of Chinese morphemes in word-formation (as well as sentence-making), research on morphemes-related resources has a long-standing history.

Yuan and Huang (1998) introduced a morpheme knowledge base by manually describing 17,470

morphemes. However, as the morphemes were listed independently, the resource could not be used for computational purposes.

Ji and Feng (2015) introduced a "morpheme and sense database with categories" by extracting 2,268 single-character morphemes from the Contemporary Chinese Dictionary (CCD) published by the Commercial Press, one of the most influential Chinese dictionaries, and annotated each morpheme with its sense category. However, the data are imbalanced and relatively small in size, making it challenging to meet computational requirements.

Liu et al. (2018); Lin and Liu (2019) proposed the Chinese Object-Oriented Lexicon (COOL) by extracting 20,855 single-character morphemes with PoS and inter-morpheme relations from CCD. COOL covers not only frequently used characters and their morphemes but also those rarely used, allowing for a reflection of historical and cultural implications. Based on it, Zheng et al. (2021) developed a word-formation dataset, FiCLS, for Chinese Word Sense Disambiguation (WSD) and provided a well-performing model.

With the accumulation of these COOL-related resources, in this paper, we can then propose a morpheme-informed evaluation for LLMs.

3. The MorphEval Benchmark

In this section, we construct MorphEval by deliberately leveraging morpheme information from COOL-related resources. Following the levels of characters, words, and sentences, MorphEval is built into three datasets.

3.1. Construction of MorphEval

3.1.1. Dataset I - Character to Morpheme

Dataset I describes the connection between characters and morphemes.

Each entry in this dataset is a character-morpheme pair (c, m). For positive samples, we reuse the character-morpheme pairs described in COOL, namely (c, m_{pos}). For negative examples, we leverage Morphemic Concepts (MCs), the hierarchical synonymous morphemic sets (Liu et al., 2018) in COOL. These MCs represent distinct semantics and can thus help filter out negative ones. Given a positive pair (c, m_{pos}), a negative pair (c, m_{neg}) hereby ensures the morpheme m_{neg} shares the same high-level MC with m_{pos} but belongs to a different lower-level one than any morphemes of c.

The dataset contains 37,013 entries, with 20,855 positive samples and 16,158 negative ones, totaling 8,514 characters and 20,855 morphemes.

3.1.2. Dataset II - Word to Morpheme

Dataset II describes the connection between words and morphemes.

Each entry in this dataset is a word-morpheme pair (w, m), where the character c is used as morpheme m in word w. Note that w in the pair is monosemous, as different senses of polysemous words may result in different morpheme compositions. Thus, polysemous words cannot be disambiguated with their sole presence. To achieve a fair evaluation free of the impact of morpheme's productivity in word-formation, each morpheme will have only one paired word retained at random.

The dataset contains 13,276 entries, totaling 12,356 words and 13,276 morphemes, covering 63.66% of morphemes in Dataset I. Notably, this coverage is less than 100% due to that some of the Chinese free morphemes can only be used independently, such as "啊 *(ah)*".

3.1.3. Dataset III - Sentence to Morpheme

Dataset III describes the connection between sentences and morphemes.

Each entry in this dataset is a sentence-morpheme pair (s, m), where the character c is used as morpheme m in sentence s. The sentences are extracted from FiCLS (Zheng et al., 2021), which is also a reconstruction and expansion of COOL. To align with the above datasets, only sentences targeting COOL's morphemes are kept. Similar to Dataset II, each morpheme will have only one paired sentence retained.

The dataset contains 10,638 entries, totaling 10,638 sentences and 10,638 morphemes.

3.2. Analysis of Coverage

Dataset	Morpheme		Character		
	CCD	ZDic	CCD	ZDic	PD.
I	100.0	83.57	100.0	99.69	99.99
II	63.66	61.55	64.82	75.41	99.62
III	51.01	48.52	31.78	38.36	93.97

Table 5: Coverage rates(%) of MorphEval over different resources. *PD.* is short for People's Daily.

To assess the coverage of MorphEval over the Chinese language, besides CCD, we introduce the dictionary of ZDic[3] (an online Chinese dictionary), and the corpora of People's Daily[4] from 2018 to 2020 as external resources for comparison. The

[3] https://www.zdic.net/
[4] http://paper.people.com.cn/

corpora, with 105 million characters, covers various genres and themes in standard language, making it a critical and frequently used resource by researchers. Table 5 shows the results.

In general, MorphEval exhibits satisfying coverage over the additional resources. Due to the different scopes of usage and principles of construction, the relatively low coverage over ZDic is understandable. However, MorphEval's coverage on the corpora of People's Daily is sufficiently high, which demonstrates its usefulness for evaluation.

4. Evaluation Setup

4.1. LLMs for Evaluation

We select typical and widely used English- and Chinese-targeted LLMs for evaluation:

GPT series of models. They represent the top performance of LLMs and are the most widely used multilingual models. ChatGPT-3.5-turbo-0613 will be evaluated through the API;

LLaMA series of models. They have become one of the most extensively used LLMs in non-English languages with significant adaptability. The original LLaMA models (Touvron et al., 2023) will be evaluated through offline models, along with its two Chinese variants, Chinese-Alpaca(Cui et al., 2023) and Linly-ChatFlow(Li et al., 2023b), which are enhanced for Chinese in different ways.

4.2. Settings

4.2.1. Implementation Setup

To ensure a fair evaluation, all models are set to share the same set of parameters. We employ a temperature of 0 for greedy search generation, put the frequency penalty and top-k to 0, top-p to 1. We adjust their settings as closely as possible for models like Alpaca, which cannot precisely match the parameters above.

Zero-shot setting. In a zero-shot setting, models are evaluated on questions without prior examples of the specific task, which tests their innate ability to perceive and solve problems.

Few-shot setting. In a few-shot setting, models are provided with a few task-specific examples before being tested on new samples. For Dataset I, we extract two samples from ZDic. For Datasets II and III, we randomly select two samples with medium-sized candidate morphemes from the data not included in MorphEval.

4.2.2. Evaluation Prompts

On Dataset I, models are presented with a character and a morpheme sense, and then asked to determine whether they constitute a positive pair. On Datasets II and III, models are given a context, a target character, and multiple candidate morpheme senses. They are required to distinguish the most suitable morpheme. Figure 1 provides examples of prompts for the three datasets, respectively. All prompts in both zero- and few-shot settings are shown in Appendix A.

Figure 1: Examples of prompts for Datasets I, II and III in zero-shot setting.

Considering the sensitivity of LLMs to prompts, each sample will be subjected to three carefully designed prompts so as to demonstrate model performances properly. To further validate the effectiveness of these prompts, we instruct ChatGPT to generate ten additional prompts with varying expressions and lengths. They are then tested with 1,000 random samples each dataset in a zero-shot setting. Results suggest that our original prompts better stimulate the capabilities of all models except for Alpaca-13B, with all score differences less than 5 points. Table 6 provides a subset of results.

Model	ori_Avg. - new_Avg.
ChatGPT	2.04
LLaMA-13B	3.07
Alpaca-13B	-0.49
Linly-13B	1.10

Table 6: Results for validating prompts. *ori_Avg.* is the average score on original prompts, and *new_Avg.* is that on newly generated 10 prompts.

4.2.3. Evaluation Details

Post-processing. We adopt specific strategies to automatically extract one or more standard answers, which are optimized based on different features of generated text.

Model	Zero-shot setting						Few-shot setting					
	Dataset I	Dataset II		Dataset III		Avg.	Dataset I	Dataset II		Dataset III		Avg.
	exact	exact	fuzzy	exact	fuzzy		exact	exact	fuzzy	exact	fuzzy	
Random	50.00	32.47	32.47	20.88	20.88	31.34	50.00	32.47	32.47	20.88	20.88	31.34
ChatGPT	64.31	61.06	61.41	43.31	46.10	**55.24**	64.54	51.50	56.72	43.03	48.48	**52.85**
LLaMA-7B	/	20.88	34.61	22.32	36.39	28.55	52.89	24.65	25.66	19.55	21.44	28.84
LLaMA-13B	/	24.13	26.01	23.22	36.71	27.52	56.33	37.67	39.56	21.95	22.25	35.55
LLaMA-30B	/	32.85	35.67	28.87	31.56	32.24	57.11	43.90	44.03	26.45	27.17	39.73
LLaMA-65B	/	34.10	39.17	32.75	38.91	36.23	59.06	47.55	47.73	36.38	36.84	45.51
Alpaca-7B	57.23	44.49	46.10	31.86	40.70	44.08	56.73	16.84	24.73	24.67	46.02	33.80
Alpaca-13B	64.27	50.15	50.36	37.38	38.41	**48.13**	65.34	16.17	16.61	21.61	22.30	28.41
Linly-7B	64.75	45.57	45.81	36.35	36.42	45.78	60.13	41.71	41.82	28.76	28.76	40.24
Linly-13B	69.12	47.28	47.50	36.64	36.67	47.44	48.85	45.99	46.11	30.23	30.31	**40.30**
Full Avg.Δ	/	7.59↑	10.49↑	11.64↑	17.11↑	9.24↑	7.89↑	3.75↑	5.64↑	7.19↑	10.63↑	7.02↑

Table 7: Performance of LLMs on MorphEval. *Random* is the random baseline. *Full Avg.*Δ represents the average score changes over *Random*.

Evaluation metrics. MorphEval is constructed from dictionary-based resources. Thus, it tends to have finer-grained sense granularity that is not always necessary for computation. Also, there is no consensus on how morpheme senses should be divided for characters. With these considerations, we will set up two metrics for evaluation: **(1) Exact Matching.** requires model output to be completely consistent with the label. When applied to Datasets II and III, it eliminates multiple generated options, leaving only the first option for metric calculation; **(2) Fuzzy Matching.** allows a generated result to be seen as correct if it includes the golden answer, regardless of other choices. This metric is tailor-made for Datasets II and III, as Dataset I only has definite true/false responses.

5. Evaluation Results and Analysis

5.1. Results

Table 7 reports the general evaluation results. As LLaMA models have not undergone instruction tuning and cannot provide acceptable responses in a zero-shot setting on Dataset I, there are no experimental results for them in this part.

5.1.1. Comparison among models

In general, only ChatGPT achieves an average accuracy of over 50%, highlighting the challenges presented by MorphEval. Alpaca-13B excels as the second-best model in zero-shot setting, beating the larger LLaMA-65B with a substantial lead. Nevertheless, LLaMA-65B exhibits better flexibility to few-shot samples.

Chinese-targeted models show significantly superior performance than their English-targeted predecessors, LLaMA, with the same or even fewer parameters. Notably, Linly-7B and -13B outshine ChatGPT on Dataset I in a zero-shot setting, with Alpaca-13B outperforming it on the same dataset in a few-shot setting. These findings suggest potential benefits of using Chinese-targeted models in Chinese scenarios, even with smaller parameters.

We further compare the Alpaca and Linly series, which share the same LLaMA predecessor but are adapted to Chinese differently. Results show that Alpaca-13B outperforms Linly-13B in a zero-shot setting. In general, both models are adversely affected in a few-shot setting, with Linly showing greater adaptability.

5.1.2. Comparison between zero- and few-shot settings

With few-shot examples, LLaMA models improve by 6.27 points, while ChatGPT, Alpaca, and Linly drop by 9.01 on average. This disparity could stem from the fact that, according to Zhang et al. (2022), the impact of in-context learning relies heavily on few-shot examples. Examples suitable for one model may not fit another.

Specifically, Dataset I deviates from Datasets II and III as its task is unfamiliar to LLMs. In zero-shot settings, models, especially LLaMA for example, either fail to provide plausible responses or repeatedly produce identical answers, irrespective of positive or negative samples. However, with a few-shot setting, models can now deliver sensible answers despite the possible decline in indicators, thereby genuinely reflecting their inherent capabilities.

5.1.3. Comparison among language units

Model performances across datasets exhibit the following tendencies. First, although Dataset I is

evaluated with a binary classification task, with fewer candidates than Datasets II and III, model performances on it are only around 60%. Second, models' average score changes (*Full Avg.*Δ) improve gradually from Dataset I to Dataset III in both zero- and few-shot settings.

This implies that the richness of contextual information has impact on model performances.

5.2. Analysis

To obtain a more in-depth analysis, we randomly selected 100 prediction errors of ChatGPT from each dataset. It is observed that approximately 30% of cases are associated with functional morphemes such as prepositions and conjunctions, 17% with long-tailed distribution of semantics, and 16% with cultural implications. The rest are related to special cases, such as abbreviations, scientific terms, etc. Sequentially, we then focus on these three aspects in analysis: morpheme usages with different PoS, long-tailed distribution of semantics, as well as cultural implications.

As discussed in section 5.1.2, to ensure informative and undisturbed analysis, we consider results with exact matching in few-shot settings for Dataset I and zero-shot settings for Datasets II and III.

5.2.1. Morpheme usages with different PoS

In MorphEval, each morpheme has obtained its PoS from COOL. Table 8 shows its distribution across PoS. In addition to a plethora of content

Dataset	N.	V.	Adj.	Adv.	Func.
I	46.90	30.34	11.23	2.75	8.78
II	38.17	37.90	14.12	3.45	6.37
III	30.17	42.95	14.68	4.65	7.54

Table 8: Distributions of MorphEval across PoS (%). "Func." is short for functional morphemes.

morphemes (nouns, verbs, and adjectives), MorphEval comprises 2.75% adverbial and 8.78% functional morphemes (prepositions, conjunctions, pronouns, etc., connecting semantics within words or sentences). Though less discussed, functional morphemes are pivotal for accurate language perception, as evidenced in Table 9. In this case, functional morphemes with identical characters but differing senses can significantly alter sentence interpretations. To enhance clarity, we provide extra contextual information. However, even in isolation, "与" can convey multiple senses and affect the overall sentence meaning.

In subsequent analysis, we concentrate on English- and Chinese-targeted top-performing models, namely ChatGPT and Alpaca-13B. We also include LLaMA-13B, the predecessor model for Alpaca-13B. Table 10 provides models' overall performances across different PoS.

Character	Morph Sense	PoS	Sentence
与	跟；向	介	我与他讲了个故事。（他很喜欢听。）
to	targeting	prep.	I told a story to him. (He loved the story.)
与	和	连	我与他讲了个故事。（大家都被我们吸引了。）
with	together with	conj.	I told a story with him. (People were attracted by us.)

Table 9: Examples of functional morphemes with the same character changing sentence meaning.

Model	Data	Avg.	N.	V.	Adj.	Adv.	Func.
Rand	Avg.	34.44	36.50	34.51	35.05	31.54	34.60
ChatGPT	I	75.61	73.07	80.36	78.86	73.17	65.84
	II	61.06	64.50	61.95	58.48	52.91	45.27
	III	43.31	45.30	46.01	45.87	29.23	23.82
	Avg.	59.99	60.94	62.77	61.07	51.77	44.98
	$\Delta\uparrow$	25.55	24.44	28.26	26.02	20.23	10.38
LLaMA-13B	I	56.33	60.53	51.25	53.19	51.13	61.24
	II	24.13	25.24	24.08	22.91	21.54	21.91
	III	23.22	24.24	23.88	25.03	14.68	17.22
	Avg.	34.56	36.67	33.07	33.71	29.12	33.46
	$\Delta\uparrow$	0.12	0.16	1.44$\downarrow$	1.34$\downarrow$	2.42$\downarrow$	1.14$\downarrow$
Alpaca-13B	I	65.34	60.68	72.30	70.01	66.08	54.61
	II	50.15	53.65	48.25	47.51	46.72	48.19
	III	37.38	37.65	39.15	36.90	33.94	29.34
	Avg.	50.96	50.66	53.23	51.47	48.91	44.05
	$\Delta\uparrow$	16.52	14.16	18.72	16.42	17.37	9.45

Table 10: Overall results of ChatGPT, LLaMA-13B and Alpaca-13B with PoS breakdown. *Rand* represents the random baseline. $\Delta\uparrow$ is model's average score improvement over *Rand*.

To compare models, Alpaca-13B outperforms LLaMA-13B on all PoS, proving the potency of language-specific enhancements. However, Alpaca's least progress over LLaMA is made with functional morphemes, suggesting them to be a demanding aspect of model enhancements.

When comparing morphemes with different PoS, ChatGPT and Alpaca generally obtain lowest absolute scores on functional morphemes, while lowest relative improvements ($\Delta\uparrow$) are on nominal, adverbial and functional morphemes. To explore the underlying causes, we further analyze the accuracy of PoS between the predicted morpheme and the labeled morpheme, namely, the average accuracy

of PoS in prediction (PoS_Acc). Results are provided in Table 11. Although morpheme senses are not explicitly identified with PoS in the benchmark, they may implicitly possess definition patterns under certain PoS, which LLMs could capture.

Model	N.	V.	Adj.	Adv.	Func.
#Cand.	3.64	3.94	2.51	1.90	2.38
#Full_Cand.	5.78	6.36	5.90	6.93	6.82
Random	72.55	67.23	51.39	33.63	46.65
ChatGPT	**82.24**	**82.15**	**70.72**	49.34	50.15
LLaMA-13B	65.23	61.69	47.79	25.98	36.69
Alpaca-13B	76.22	75.69	61.90	**51.92**	**55.29**
Avg.	74.56	73.18	60.14	42.41	47.38
Avg.$\Delta\uparrow$	2.01	5.95	8.75	8.78	0.73

Table 11: Average accuracy of PoS in prediction (PoS_Acc). Within the same character, *#Cand.* is the average number of options on the same PoS, while *#Full_Cand.* is that of all options.

For nominal morphemes, models obtain low relative overall improvements, while high PoS_Acc in Table 11. The high PoS_Acc is due to the greater ratio of noun options in candidates, as shown by *#Cand.* and *#Full_Cand.*. The low relative improvements of overall results as well as PoS_Acc, on the other hand, could be due to their commonly extended usage in Chinese. For example, character "草" has its basic nominal morpheme with the sense of *"grass"*, which is then extended to another nominal morpheme with the sense of *"countryside or folk"*, as well as an adjective morpheme with the sense of *"sloppy"*.

For adverbial morphemes, models' low relative improvements of overall results are due to characters for adverbial morphemes usually having more other morpheme senses, as shown in Table 11.

For functional morphemes, models exhibit low absolute scores and relative improvements in overall results and PoS_Acc, with the ratio of functional options in candidates being relatively low. This indicates confusion not only with morphemes in other PoS but also among the functional type, revealing models' dysfunction with them. For instance, the character "当" in "当众 *(in public)*" is misinterpreted by ChatGPT from a prepositional sense of *"in front of"* to a verbal sense of *"serve as"*. On the other hand, the character "缘" in "路缘溪而建。 *(The road was built along stream.)*" is misconstrued from a prepositional sense of *"along"* to another prepositional sense of *"due to"*. Such misinterpretations of functional morphemes can lead to misunderstandings of entire words or sentences, emphasizing the need for greater attention to morphology and syntax in future model enhancements.

5.2.2. Long-tailed distribution of semantics

MorphEval, in addition to its extensive coverage of the People's Daily, contains 1,847 characters missing from the corpora. Furthermore, over half of the MorphEval characters appear less than 100 times in the corpora. Thus, the benchmark has the capacity for characters from the long tail, along with their semantics. We extract this long tail, namely the characters not found in the corpora, and show the evaluation results in Table 12.

Model	I	II	III	Avg.
Random	50.00	70.42	40.00	53.47
ChatGPT	56.93	78.35	**66.67**	**67.32**
LLaMA-7B	42.76	19.34	42.70	34.93
LLaMA-13B	60.64	26.55	41.20	42.80
LLaMA-30B	60.93	33.48	48.31	47.57
LLaMA-65B	60.46	44.23	58.80	54.50
Alpaca-7B	**64.82**	70.37	62.17	65.79
Alpaca-13B	57.42	75.69	62.92	65.34
Linly-7B	42.07	77.99	56.93	59.00
Linly-13B	39.39	**80.23**	65.17	61.60
LT Avg.Δ	3.94$\uparrow$	14.17$\downarrow$	16.10$\uparrow$	1.96$\uparrow$
Full Avg.Δ	7.89$\uparrow$	7.59$\uparrow$	11.64$\uparrow$	9.04$\uparrow$

Table 12: Performance of LLMs on long-tailed morphemes. *LT Avg.*Δ is the average score changes on long-tailed morphemes. *Full Avg.*Δ is their changes on the entire dataset.

The average score improvement on morphemes from the long tail (LT Avg.Δ) is only 1.96 points, with even a 14.17-point drop on Dataset II, which is significantly lower than the whole dataset's 9.04-point increase (Full Avg.Δ). This reveals the difficulties LLMs face when dealing with this long tail.

To compare models, Alpaca and Linly-13B surpass ChatGPT on Datasets I and II, respectively, proving the lead of smaller Chinese-targeted models.

Char	搦	
Word	搦战 Lure the enemy into war.	
Sent	华雄引铁骑下关，来寨前大骂搦战。 Xiong Hua, leading cavalry to the pass, luring them to war outside their camp.	
Cand	A.持；握；拿着。 B.挑；惹。	A. Hold; grip. B. Provoke; lure.
Label	B	

Table 13: A long-tailed morpheme ChatGPT fails to predict within a word but succeeds within a sentence.

When comparing language units, a gradual improvement from Dataset II to III is observed. For example, Table 13 shows how ChatGPT fails to choose the correct morpheme sense "*Provoke; lure*" within a word but succeeds within the same word in a sentence. This suggests that LLMs can deduce a correct answer from a more lengthy context, offering a possible solution for their poor performance on words from the long tail.

5.2.3. Cultural implications

MorphEval, constructed from dictionary-based resources, also contains many morphemes with cultural implications, as exemplified in Table 14. These morpheme senses are common in Chinese yet need some cultural background to understand.

Character	Morpheme Sense
上	旧时指皇帝。
up	The emperor in ancient China.
草	旧指山野、民间。
grass	The countryside or the folk.

Table 14: Morpheme senses with cultural implications.

A total of 500 morphemes with cultural implications are manually selected and evaluated, as shown in Table 15.

Model	I	II	III	Avg.
Random	50.00	16.03	20.88	28.97
ChatGPT	**65.05**	**54.83**	29.71	**49.86**
LLaMA-7B	55.24	16.59	18.73	30.18
LLaMA-13B	54.84	17.39	15.00	29.08
LLaMA-30B	55.05	20.85	22.55	32.82
LLaMA-65B	57.22	24.88	24.41	35.50
Alpaca-7B	56.89	30.19	19.61	35.56
Alpaca-13B	59.33	43.00	**31.86**	44.73
Linly-7B	51.47	30.03	30.49	37.33
Linly-13B	45.90	33.82	23.82	34.51
CI Avg.Δ	5.67↑	14.15↑	9.75↑	9.76↑
Full Avg.Δ	7.89↑	7.59↑	11.64↑	9.04↑

Table 15: Performance of LLMs with cultural implications. *CI Avg.*Δ is the average score changes on cultural implications, while *Full Avg.*Δ is that on the entire dataset.

When comparing language units, in Datasets I and III, score improvements on culture-implicated morphemes (CI Avg.Δ) are less than that on full datasets (Full Avg.Δ), indicating challenges to LLMs. Models score higher on Dataset II is due to that these culture-implicated words are common in Chinese, which LLMs may have met with a lot.

However, as cultural implications are often more flexibly used in sentences, performance gaps between Dataset II and III (in ChatGPT, Alpaca and Linly-13B) reveal that LLMs could be misled and show a noticeable drop in performance. As an illustration, for character "点", ChatGPT makes the correct choice of morpheme sense "*dim sum*" in the common word "糕点 *(pastry)*". When tested within the sentence of "这本书包含了一些细微洞察，是文学的柠檬糕点 *(This book contains some subtle insights and is a lemon pastry of literature.)*", the model would be misled to the wrong morpheme sense of "*slight spot*".

On the other hand, the performance gaps in Alpaca and Linly are smaller than that in ChatGPT. This suggests that language-specific enhancements could help improve models' adaption to more flexibly used cultural implications.

6. Conclusion

In this paper, we advance LLM evaluation from a linguistic perspective and introduce morphemes as the basis. We construct MorphEval, a Chinese morpheme-informed benchmark, including three datasets following the levels of characters, words, and sentences, respectively. We evaluate representative LLMs and reveal three under-emphasized yet critical LLMs' weaknesses, highlighting the need for language-specific improvements:

Dysfunctions in morphology and syntax. Analysis of morpheme usages with different PoS reveals that LLMs have dysfunctions in understanding functional morphemes, which could result in misunderstandings of the entire words or sentences. Despite Chinese-targeted models showing better performance, it is suggested to be a demanding aspect of model enhancements.

Challenges with the long-tailed distribution of semantics. Analysis of long-tailed distribution uncovers a noticeable drop in LLM performance when handling less common semantics, where smaller Chinese-targeted models even outperform ChatGPT. It's suggested that a language-specific model enhancement or a more lengthy context could help improve the semantic generalization of LLMs.

Difficulties from cultural implications. Analysis of cultural implications shows that LLMs could understand them within the common words but would be misled by more flexibly used of them in sentences. It's suggested that language-specific model enhancements, as well as human-centered knowledge bases, could help shorten this gap.

In future work, we will enlarge the benchmark and broaden it to multi-character morphemes. The benchmark-guided enhancements, as well as extensions to other languages, are also under consideration.

Acknowledgement

This paper is supported by the National Natural Science Foundation of China (No. 62036001) and the National Social Science Foundation of China (No. 18ZDA295). We thank all the anonymous reviewers for their constructive comments, and Shuhuai Ren for the helpful suggestions in preparing the manuscript.

7. Bibliographical References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

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, et al. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, et al. 2021. Evaluating large language models trained on code.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for Chinese LLaMA and Alpaca.

Yonghe Fu. 1988. Modern Chinese general character list: Appendix of the most and secondly frequently used characters. *Yuwen Jianshe*, (02):22–31.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Chan-Jan Hsu, Chang-Le Liu, Feng-Ting Liao, Po-Chun Hsu, Yi-Chang Chen, and Da shan Shiu. 2023. Advancing the evaluation of traditional chinese language models: Towards a comprehensive benchmark suite.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-Eval: A multi-level multi-discipline Chinese evaluation suite for foundation models.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023a. Cmmlu: Measuring massive multitask language understanding in chinese.

Yudong Li, Yuhao Feng, Zhe Zhao, Cheng Hou, Wen Zhou, Xiaoqin Wang, Shuang Li, Hao Li, Xianxu Hou, Yiren Chen, Jing Zhao, Ningyuan Sun, and Wenjun Tang. 2023b. Chinese Falcon & LLaMA & OpenLLaMA large language model.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models.

Can Luo. 2013. The classification of morphemes,words and its grammartical and semantic catesories in the 3000 characters.

Shuxiang Lv. 1979. *Chinese grammar analysis issues*. The Commercial Press, Beijing, China.

OpenAI. 2023. GPT-4 technical report.

Ralph Peeters and Christian Bizer. 2023. Using ChatGPT for entity matching.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Ahmed R. Sadik, Antonello Ceravola, Frank Joublin, and Jibesh Patra. 2023. Analysis of ChatGPT on source code.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Hang Yan, Xiangyang Liu, Yunfan Shao, Qiong Tang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, and Xipeng Qiu. 2023. Moss: Training conversational language models from synthetic data.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Roziere, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. SuperGLUE: A stickier benchmark for general-purpose language understanding systems.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023. Superclue: A comprehensive chinese large language model benchmark.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. GLM-130B: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Baoli Zhang, Haining Xie, Pengfan Du, Junhao Chen, Pengfei Cao, Yubo Chen, Shengping Liu, Kang Liu, and Jun Zhao. 2023. Zhujiu: A multi-dimensional, multi-faceted chinese benchmark for large language models.

Kai Zhang. 1997. Statistical analysis of Chinese morpheme-based characters. *Language Teaching and Linguistic Studies*, (01):43–52.

Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages with in-context linguistic descriptions. *arXiv preprint arXiv:2402.18025*.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yixuan Zhang and Haonan Li. 2023. Can large language model comprehend ancient chinese? a preliminary test on aclue.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. AGIEval: A human-centric benchmark for evaluating foundation models.

Dexi Zhu. 1982. *Grammar Lectures*. The Commercial Press, Beijing, China.

Junchen Zhu, Huan Yang, Huiguo He, Wenjing Wang, Zixi Tuo, Wen-Huang Cheng, Lianli Gao, Jingkuan Song, and Jianlong Fu. 2023. Moviefactory: Automatic movie creation from text using large generative models for language and images.

George Kingsley Zipf. 1949. *Human behavior and the principle of least effort.* Addison-Wesley Press.

8. Language Resource References

Zhiwei Ji and Minxuan Feng. 2015. A study on semantic word-formation of bi-character words for common unknown word understanding. *Journal of Chinese Information Processing*, 29(05):63–68+83.

Zi Lin and Yang Liu. 2019. Implanting rational knowledge into distributed representation at morpheme level. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33(01), pages 2954–2961.

Yang Liu, Zi Lin, and Sichen Kang. 2018. Towards a description of Chinese morphemic concepts and semantic word-formation. *Journal of Chinese Information Processing*, 32(2):12–21.

Chunfa Yuan and Changning Huang. 1998. Research on Chinese morphemes and word formation based on morpheme database. *Chinese Teaching in the World*, 2:7–12.

Hua Zheng, Damai Dai, Lei Li, Tianyu Liu, Zhifang Sui, Baobao Chang, and Yang Liu. 2021. Decompose, fuse and generate: A formation-informed method for Chinese definition generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5524–5531, Online. Association for Computational Linguistics.

A. Prompts for Evaluation

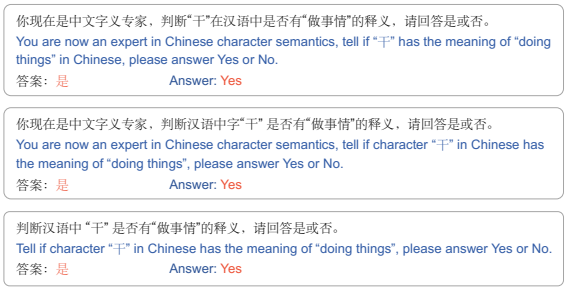

Figure 2: Prompts for Datasets I in the zero-shot setting.

Figure 3: Prompts for Datasets II in the zero-shot setting.

Figure 4: Prompts for Datasets III in the zero-shot setting.

你现在是中文字义专家，判断目标字在汉语中是否有目标释义的含义，请回答是或否。
You are now an expert in Chinese character semantics, tell if target character in Chinese has the meaning of target sense, please answer Yes or No.

目标字：笑。 目标释义：露出愉快的表情，发出欢喜的声音。
Target character: 笑. Target sense: Display a joyful expression and emit sounds of delight.
答案：是 Answer: Yes

目标字：雨。 目标释义：空气流动的现象。气象学特指空气在水平方向的流动。
Target character: 雨. Target sense: Air movement in meteorology.
答案：否 Answer: No

目标字：干。 目标释义：做事情。
Target character: 干. Target sense: Doing things.
答案：是 Answer: Yes

你现在是中文字义专家，判断汉语中目标字是否可以被解释为目标释义，请回答是或否。
You are now an expert in Chinese character semantics, tell if target character in Chinese can be explained as target sense, please answer Yes or No.

目标字：笑。 目标释义：露出愉快的表情，发出欢喜的声音。
Target character: 笑. Target sense: Display a joyful expression and emit sounds of delight.
答案：是 Answer: Yes

目标字：雨。 目标释义：空气流动的现象。气象学特指空气在水平方向的流动。
Target character: 雨. Target sense: Air movement in meteorology.
答案：否 Answer: No

目标字：干。 目标释义：做事情。
Target character: 干. Target sense: Doing things.
答案：是 Answer: Yes

判断汉语中目标字是否有目标释义的含义，请回答是或否。
Tell if target character in Chinese has the meaning of target sense, please answer Yes or No.

目标字：笑。 目标释义：露出愉快的表情，发出欢喜的声音。
Target character: 笑. Target sense: Display a joyful expression and emit sounds of delight.
答案：是 Answer: Yes

目标字：雨。 目标释义：空气流动的现象。气象学特指空气在水平方向的流动。
Target character: 雨. Target sense: Air movement in meteorology.
答案：否 Answer: No

目标字：干。 目标释义：做事情。
Target character: 干. Target sense: Doing things.
答案：是 Answer: Yes

Figure 5: Prompts for Datasets I in the few-shot setting.

你现在是中文字义消歧专家，请你从候选释义中选择目标字在目标词中的释义。
You are now an expert in Chinese character sense disambiguation, choose from candidate senses the meaning of target character in target word.

目标词：候审 Target word: 候审
目标字：侯 Target character: 侯
候选释义： Candidate senses:
A. 详细；周密 A. detailed; thorough
B. 审查 B. review
...
答案：C Answer: C

目标词：全称 Target word: 全称
目标字：称 Target character: 称
候选释义： Candidate senses:
A. 适合；相当 A. suitable; comparable
B. 名称 B. name
...
答案：B Answer: B

目标词：干着急 Target word: 干着急
目标字：干 Target character: 干
候选释义： Candidate senses:
A. 做事情 A. doing things
B. 无效；徒劳 B. in vain
...
答案：B Answer: B

请从候选释义中选择目标字在目标词中的释义。
Please choose from candidate senses the meaning of target character in target word.

目标词：候审 Target word: 候审
目标字：侯 Target character: 侯
候选释义： Candidate senses:
A. 详细；周密 A. detailed; thorough
B. 审查 B. review
...
答案：C Answer: C

目标词：全称 Target word: 全称
目标字：称 Target character: 称
候选释义： Candidate senses:
A. 适合；相当 A. suitable; comparable
B. 名称 B. name
...
答案：B Answer: B

目标词：干着急 Target word: 干着急
目标字：干 Target character: 干
候选释义： Candidate senses:
A. 做事情 A. doing things
B. 无效；徒劳 B. in vain
...
答案：B Answer: B

请从候选释义中选择#内的字在目标词中的释义。
Please choose from candidate senses the meaning of target character in target word.

目标词：#候#审 Target word: #候#审
候选释义： Candidate senses:
A. 详细；周密 A. detailed; thorough
B. 审查 B. review
...
答案：C Answer: C

目标词：全#称# Target word: 全#称#
候选释义： Candidate senses:
A. 适合；相当 A. suitable; comparable
B. 名称 B. name
...
答案：B Answer: B

目标词：#干#着急 Target word: #干#着急
候选释义： Candidate senses:
A. 做事情 A. doing things
B. 无效；徒劳 B. in vain
...
答案：B Answer: B

Figure 6: Prompts for Datasets II in the few-shot setting.

请你从候选释义中选择目标字在上下文中的释义。
Please choose from candidate senses the meaning of target character in context.
目标字：视 Target character: 视
上下文：中央巡视组反馈，多地存在超职数配副秘书长等问题。
Context: Central inspection reveals excessive deputy secretaries in several regions.
候选释义： Candidate senses:
A. 看 A. looking at
B. 考察 B. inspect
... ...
答案：B Answer: B

目标字：菜 Target character: 菜
上下文：中国菜已经历数千年的发展历史。
Context: Chinese cuisine has a history of thousands of years.
候选释义： Candidate senses:
A．能做副食品的植物；蔬菜。 A. vegetable
B．专指油菜。 B. rapeseed plant
... ...
答案：C Answer: C

目标字：干 Target character: 干
上下文：他一整天都在认真干活 Context: He has been working hard all day.
候选释义： Candidate senses:
A. 做事情 A. doing things
B. 无效；徒劳 B. in vain
... ...
答案：A Answer: A

请你从候选释义中选择#内的字在上下文中的释义。
Please choose from candidate senses the meaning of character in # within context.
上下文：中央巡#视#组反馈，多地存在超职数配副秘书长等问题。
Context: Central #inspect#ion reveals excessive deputy secretaries in several regions.
候选释义： Candidate senses:
A. 看 A. looking at
B. 考察 B. inspect
... ...
答案：B Answer: B
上下文：中国#菜#已经历数千年的发展历史。
Context: Chinese #cuisine# has a history of thousands of years.
候选释义： Candidate senses:
A．能做副食品的植物；蔬菜。 A. vegetable
B．专指油菜。 B. rapeseed plant
... ...
答案：C Answer: C
上下文：他一整天都在认真#干#活 Context: He has been #working# hard all day.
候选释义： Candidate senses:
A. 做事情 A. doing things
B. 无效；徒劳 B. in vain
... ...
答案：A Answer: A

你现在是中文字义消歧专家，请你从候选释义中选择#内的字在上下文中的释义。
You are now an expert in Chinese character sense disambiguation, choose from candidate senses the meaning of character in # within context.
上下文：中央巡#视#组反馈，多地存在超职数配副秘书长等问题。
Context: Central #inspect#ion reveals excessive deputy secretaries in several regions.
候选释义： Candidate senses:
A. 看 A. looking at
B. 考察 B. inspect
... ...
答案：B Answer: B
上下文：中国#菜#已经历数千年的发展历史。
Context: Chinese #cuisine# has a history of thousands of years.
候选释义： Candidate senses:
A．能做副食品的植物；蔬菜。 A. vegetable
B．专指油菜。 B. rapeseed plant
... ...
答案：C Answer: C
上下文：他一整天都在认真#干#活 Context: He has been #working# hard all day.
候选释义： Candidate senses:
A. 做事情 A. doing things
B. 无效；徒劳 B. in vain
... ...
答案：A Answer: A

Figure 7: Prompts for Datasets III in the few-shot setting.