From BERT to LLMs： Comparing and Understanding Chinese Classifier Prediction in Language Models

Ziqi Zhang, Jianfei Ma, Emmanuele Chersoni, Jieshun You, Zhaoxin Feng
Department of Language Science and Technology, The Hong Kong Polytechnic University
{blameredens.zhang, jianfei-mark.ma,
jieshun.you, zhaoxinbetty.feng}@connect.polyu.hk,
emmanuele.chersoni@polyu.hk

Abstract

Classifiers are an important and defining feature of the Chinese language, and their correct prediction is key to numerous educational applications. Yet, whether the most popular Large Language Models (LLMs) possess proper knowledge the Chinese classifiers is an issue that has largely remain unexplored in the Natural Language Processing (NLP) literature.

To address such a question, we employ various masking strategies to evaluate the LLMs' intrinsic ability, the contribution of different sentence elements, and the working of the attention mechanisms during prediction. Besides, we explore fine-tuning for LLMs to enhance the classifier performance.

Our findings reveal that LLMs perform worse than BERT, even with fine-tuning. The prediction, as expected, greatly benefits from the information about the following noun, which also explains the advantage of models with a bidirectional attention mechanism such as BERT.

1 Introduction

Chinese classifiers constitute a morphosyntactic category that semantically marks noun classes (Ahrens and Huang, 2016). They precede a head noun and combine with numerals or demonstrative pronouns to convey quantity or frequency within noun phrases (Li and Thompson, 1989). As illustrated in Figure 1, such linguistic devices construct a complex system describing different semantic features of head nouns that they precede (Huang and Shi, 2016). The large classifier inventory in Chinese often allows different classifiers to combine with the same head noun, conveying distinct semantic nuances (Shi, 2014; Huang and Chen, 2014). For example, both individual classifiers "个" and "位" can modify the noun of people, while the former is a more generic one, the latter is restricted to highly-regarded professions and conveys a polite tone. Improper collocations can result in semantic or pragmatic violations (Chan, 2019).

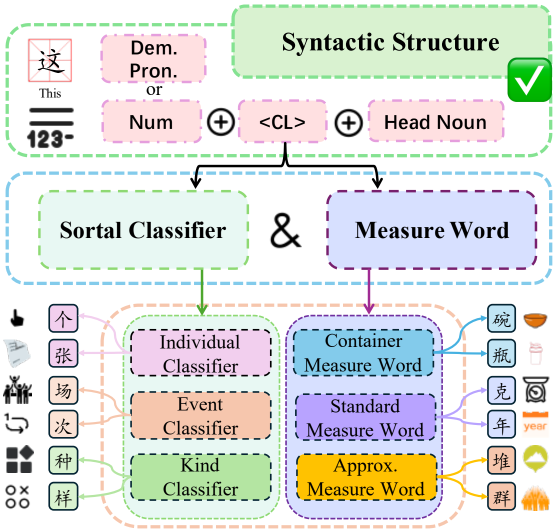

Figure 1: This figure elaborates on correct syntactic structures of the Chinese classifier system, the types of Chinese classifiers, and corresponding examples, where Num, CL, Dem. Pron., and Approx. stand for numeral, classifier, demonstrative pronoun, and approximation, respectively. Notably, several classifier examples on both sides are accompanied by icons that illustrate the approximate meanings they convey. A detailed explanation is provided in Appendix C.

In the natural language processing (NLP) literature, there has been extensive exploration of Natural Language Understanding (NLU) using Pretrained Language Models (PLMs) (Wang et al., 2022) and Large Language Models (LLMs) (Chang et al., 2024; Ma et al., 2025). However, all current studies on Chinese classifiers only include behaviours of traditional models for classifier prediction or selection with limited interpretability efforts (Peinelt et al., 2017; Järnfors et al., 2021), while to the best of our knowledge, there is no evaluation study with more recent autoregressive LLMs. In addition to the interest in evaluating LLMs on this essential component of Chinese grammar, it should

Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 317–329
November 9, 2025 ©2025 Association for Computational Linguistics

be kept in mind that learning classifier systems has been proven to be particularly challenging for learners of Chinese as L2 (Liang, 2008; Liu, 2018), and thus NLP technologies with a robust knowledge of classifiers would be a precious resource to develop educational tools.

Given the above-mentioned gap in the literature, we address the following research questions: How do LLMs perform in this classifier prediction? What are the semantic contributions of the different elements of a sentence to the process of selecting a classifier, and can this be observed from the attention mechanism of the model?

With this goal in mind, we establish a control task by iteratively inserting various classifiers into a blank classifier position within sentence and ranking them based on their Language Model (LM) log probabilities. With the same setups, we randomly extract sentence samples and mask the token of the classifier and then fine-tune the models to examine how well they can perform. Finally, we carry out additional analysis by modifying the model's attention mask, in order to make them ignore the surrounding words in the sentence and quantify their information contribution to task performance.

The control task shows that BERT (Devlin et al., 2019) and LLMs achieve good accuracy in the prediction, yet the former exhibits a distinct advantage and higher improvement potential with fine-tuning. Due to the strong semantic link between classifiers and head nouns, models exhibit a high dependency on the corresponding head noun during prediction. Intriguingly, the results also confirm an additional (albeit weak) contribution from the remaining contextual information. The same experiment also reveals that the bidirectional attention mechanism plays a critical role, despite the bigger parameters and training data size of autoregressive LLMs.

2 Related Work

2.1 Chinese Classifiers

Chinese classifiers serve as obligatory syntactic elements bridging numerals and head nouns, forming grammatically complete noun phrases (Li and Thompson, 1989) while encoding semantic features, including shape and function, and taxonomic categorization (Lakoff, 1986; Croft, 1994). Current research on this element centers on the usage patterns across diverse population groups (Zhan and Levy, 2018; Shi, 2021) and its nuanced idiosyncrasies (Liu et al., 2019).

However, recent computational studies of the prediction task remain scarce. Existing studies most focus on early approaches, covering SVMs (Guo and Zhong, 2005) and Word2Vec embeddings (Peinelt et al., 2017), later augmented with mutual information metrics (Liu et al., 2019).

The Transformers marked a turning point. But only Järnfors et al. (2021) demonstrated BERT's superior performance after fine-tuning, though revealing persistent deficiencies in implication covering politeness and plural markers. This limitation motivates investigating whether modern LLMs' enhanced contextual awareness and linguistic knowledge can achieve more robust classifier prediction.

2.2 Attention Mechanism in Lexical Semantics

BERT's bidirectional attention provides comprehensive contextual awareness by processing both left and right contexts of target words (Devlin et al., 2019). This architectural advantage has been applied to LLMs and empirically validated across NLP tasks, like syntactic parsing and named entity recognition (BehnamGhader et al., 2024; Springer et al., 2025). Building on this foundation, (Feng et al., 2025) demonstrates that bidirectional architectures particularly excel in semantic tasks requiring precise context resolution with the framework constructed by BehnamGhader et al. (2024).

While autoregressive LLMs are inherently constrained by unidirectional attention, their substantially expanded pretraining corpora and enhanced world knowledge (Wei et al., 2022; Brown et al., 2020) might offer compensatory advantages across various NLU tasks. However, specifically, for Chinese classifier prediction, the trade-off effect of this architectural dichotomy on classifiers remains unexamined. Furthermore, how bidirectional attention boosts BERT's accuracy and how head nouns impact the performance of bidirectional models both require investigation.

2.3 Masking Strategies for Probing

Masking strategies enable controlled experiments by selectively processing target regions to assess performance changes or predicted output (Petroni et al., 2019; Kassner and Schütze, 2020; Zhong et al., 2021). One of the typical approaches is to modify LMs' attention masks, zeroing selected token weights to study attention mechanisms' effects (Liong et al., 2024). In the computational linguistics domain, this approach helps assess specific linguistic components' contributions. For instance, Metheniti et al. (2020) showed that masking non-verbal linguistic elements improves BERT's alignment with human intuitions for role fillers, while Cho et al. (2021) demonstrated similar benefits for event location prediction by masking context and forcing attention on verb phrases.

Following the previous work, we mask the target classifier to trigger the model's prediction at this position and adjust the attention mask to examine how effectively context-based attention contributes to the target classifier.

2.4 Classifier Ranking by Log Probability

Log Probability (LogProb) has been proven effective for various token-level tasks like assessing grammatical correctness and semantic plausibility, where its outputs often align with human judgments and outperform direct prompting (Hu and Levy, 2023; Kauf et al., 2024).

We acknowledge that this metric is not without limitations. It is known to be sensitive to confounding variables such as word frequency[1] and output length (Salazar et al., 2020; Holtzman et al., 2021). But they are difficult to isolate due to the uneven distribution nature of pre-training data. Despite this sensitivity, LogProb also be recognized as the robust choice for check the certainty of LMs' output because its validity and superiority for semantic tasks are strongly supported by Kauf et al. (2024).

To specifically mitigate the influence of output length,we insert classifier candidates into a sentence to compute the average LogProb of each filled sentence to directly obtain the score without any redundant generation In this framework, the candidate that yields the highest sentence-level LogProb is considered the best fit.

3 Methodology

This study evaluates the performance of two types of LMs in Chinese classifier prediction: (1) the model with bidirectional attention mechanisms, BERT, with masked language modeling and fine-tuning; and (2) autoregressive LLMs, including local deployments (Qwen3-1.7B, 4B, 8B and corresponding fine-tuned versions) and full-parameter APIs (DeepSeek-R1 and GPT-4). Due to the complex mapping between head nouns and classifiers, we obtain accuracy based on log probability ranking for evaluation. The detailed workflow is demonstrated in Figure 2.

3.1 Dataset Constructions

We employ the Chinese Classifier Dataset (Peinelt et al., 2017), a comprehensive resource with annotated classifier-noun pairs in sentential contexts, convenient to adapt classifier prediction tasks. This dataset contains 681,104 sentences, encompassing 172 distinct classifiers that nearly cover the entire commonly used Mandarin classifiers. Additionally, the Stanford constituent parser (Levy and Manning, 2003) was applied to annotate the head noun in each sentence. Although classifiers exhibit diverse pairings in pragmatic contexts, their syntactic component combinations are highly fixed, as supported by (Li and Thompson, 1989), suggesting that a relatively small number of examples is sufficient to effectively evaluate their accurate usage and prediction. Due to this and computational resource limitations, we initially randomly sampled 11,986 sentences that span all classifiers and preserve their original distribution. After manual screening to remove 69 erroneous cases (e.g., annotation errors or syntactic anomalies), we obtained 11,917 valid sentences for further processing. These sampled instances were split into training and test sets at an 85:15 ratio to support fine-tune and evaluation.

3.2 BERT Classifier Prediction

Masked language modeling To evaluate BERT's performance in Chinese classifier prediction, we utilize the **BERT-base-chinese** model through masked language modeling (MLM). Our approach computes the conditional probability of candidate classifiers at the masked position, accommodating both single-token classifiers (e.g., "个") and two-token classifiers (e.g., "档子"). Given a tagged sentence $X = (x_1, \dots,$ <CL>, <h>, head noun, </h>, $\dots, x_n)$, where <CL> is the placeholder for the classifier and <h>, </h> demarcate the head noun, we replace <CL> with one or two "[MASK]" tokens based on the classifier's tokenization. We calculate the log probability for each candidate classifier $c \in C$, where C is a set of 172 classifiers, encompassing both single-character and two-character classifiers.

For single-token classifiers, the log probability

[1] The analysis of effect for word frequency on accuracy is attached in Appendix D

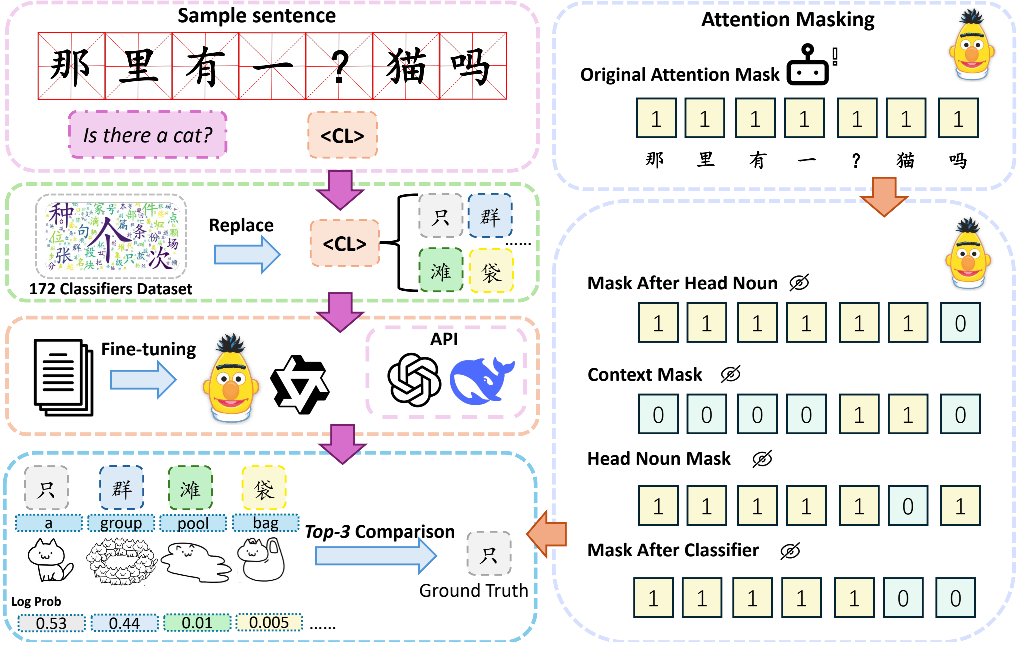

Figure 2: This figure shows the workflow of the project. As denoted by purple arrows, the given sample sentence is input into LMs for prediction. The sentences in the dataset vary in length (and are not necessarily limited to 7 tokens as in the example shown in the figure), while classifiers may consist of either one or two Chinese characters.

of a classifier c is computed as:

$$\log P(c|X) = \log\left(\text{softmax}\left(\text{BERT}(X_f(c))_{[I_1]}\right)_c\right) \tag{1}$$

where $X_f(c)$ is the sentence with <CL> replaced by classifier c, I_1 denotes the position of the single "[MASK]" token, and $\text{softmax}(\cdot)_c$ represents the probability of classifier c.

For two-token classifiers, where $c = (c_1, c_2)$, the joint log probability is calculated as:

$$\log P(c|X) = \sum_{m=1}^{2} \log\left(\text{softmax}\left(\text{BERT}(X_f(c))_{[I_m]}\right)_{c_m}\right) \tag{2}$$

where I_m is the position of the m-th `[MASK]` token ($m = 1, 2$), and $\text{softmax}(\cdot)_{c_m}$ is the probability of the m-th token of classifier c. The joint log probability sums the log probabilities of both mask positions, accurately capturing the combined likelihood of the two tokens.

Fine-tuning We use the full training set over 3 epochs with the AdamW optimizer (learning rate: 2×10^{-5}) with early stopping strategy.

3.3 LLM-Based Classifier Prediction

Sentence log probability Unlike BERT, we utilize sentence-level log probabilities for classifier ranking due to the autoregressive nature of LLMs. Since they can only access leftward context when predicting the classifier token, the isolated token probability fails to incorporate crucial information about the subsequent noun or other sentence elements. This lack of right-context access renders token-level probabilities unreliable for our task.

With locally deployed Qwen3, we replace the empty classifier position (indicated with an underscore) in each sentence with each of 172 candidate classifiers and use the **IncrementalLMScorer** from the **minicons**[2] to extract the log probability of each filled sentence by averaging the token-level log probabilities.It can be represented as:

$$\log P(S_c|X) = \frac{1}{T}\sum_{t=1}^{T} \log\left(P(w_t|w_{<t}, X_f(c))\right) \tag{3}$$

where S_c is the sentence with classifier c inserted, $X_f(c)$ is the sentence with <CL> replaced by c, T is the total number of tokens, w_t is the t-th token, and $w_{<t}$ is the preceding context. This approach evaluates the overall coherence of the sentence with the inserted classifier, averaging the log probabilities of all tokens to normalize for sentence length.

Prompting via API For the full-parameter models DeepSeek-R1 and GPT-4, we designed prompts

[2] **minicons** is a Python library for efficient probability scoring of transformer-based language models. Please refer to its' github link: https://github.com/kanishkamisra/minicons

to guide them to generate the most probable Chinese classifier for each given sentence with an empty classifier position. To diminish extraneous reasoning and maintain diversity of responses, we configured the temperature to 0, top-p to 0.9, and maximum token length to 32. To ensure uniqueness, outputs are further refined using a set-based deduplication method.

For GPT-4, we set **logprobs** parameter to be true in the API call, enabling the model to return the logarithmic probabilities of each output token. Thus, we can ensure that predicted classifiers can be sorted by their log-probability in descending order as Qwen. DeepSeek-R1 API, however, does not support LogProb extraction. Hence, we perform repeated generations with multiple candidate outputs and select the first result containing three distinct single-character classifiers as formal selections.

Metrics Predictions were evaluated using two metrics: *Accuracy* and *R-Rank*. *Accuracy* measures the proportion of samples where the model's top predicted classifier matches the correct classifier. *R-Rank*, based on previous work (Camacho-Collados et al., 2018; Peng et al., 2022), evaluates the model's nuanced understanding of classifier selection by considering the rank of the correct classifier within the top 3 predictions. Specifically, for each sample i, we define $rank_i$ as the rank of the correct classifier among the top 3 predictions, or 4 if it is not among them. These metrics are defined as follows:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left(y_i = y_i^l[0]\right) \tag{4}$$

where n is the total number of samples, y_i is the correct classifier for the i-th sample, $y_i^l[0]$ is the model's top predicted classifier for the i-th sample, and $\mathbb{1}\left(y_i = y_i^l[0]\right)$ is an indicator function that returns 1 if the top prediction matches the correct classifier, and 0 otherwise.

$$\text{R-rank} = \frac{1}{n}\sum_{i=1}^{n} rank_i \tag{5}$$

where n is the total number of samples, and $rank_i$ is the rank of the correct classifier for the i-th sample within the model's top 3 predictions (1, 2, or 3), or 4 if it is not in the top 3. The choice of a top-3 cutoff for R-rank was deliberate. Given that many Chinese nouns collocate with multiple classifiers, evaluating the top-3 candidates provides a sufficiently broad and practical assessment of a model's discriminative ability. Extending this range further would dilute the metric's practical significance, as lower-ranked candidates are increasingly unlikely to be contextually appropriate.

4 Experimental Results and Analyses

4.1 Can LLMs Be Good Classifier Guessers?

Model	Accuracy	R-rank
BERT MLM	62.31	1.8298
BERT-ft	**69.54**	**1.6676**
GPT-4	50.70	2.1408
DeepSeek-R1	59.64	1.9400
Qwen3-1.7B	31.80	2.7821
Qwen3-1.7B-ft	39.03	2.5107
Qwen3-4B	33.46	2.7270
Qwen3-4B-ft	47.69	2.2698
Qwen3-8B	39.03	2.5107
Qwen3-8B-ft	39.94	2.4861

Table 1: Results of *accuracy* and *R-rank* of different LMs for Chinese classifier prediction. "MLM" stands for "Masked Language Modeling", and "ft" denotes "Fine-tuning".

The results in Table 1 demonstrate BERT's superior performance, achieving both the highest *accuracy* and the optimal *R-rank* scores, suggesting its effectiveness in Chinese classifier prediction. In contrast, autoregressive LLMs, including GPT-4 and the Qwen3 variants, generally underperform, with most models failing to surpass 0.5 *accuracy* and exhibiting *R-rank* values between 2 and 3. Notably, Deepseek-R1 is an exception, achieving a competitive R-rank and higher *accuracy* than other LLMs, though it still falls short of BERT's performance. While scaling model parameters yields marginal improvements, even the largest models in this study, Deepseek-R1 and GPT-4, do not close the gap with BERT. This suggests that architectural differences (e.g., masked language modeling vs. autoregressive generation) may play a more critical role than parameter size in this task.

4.2 Can LLMs with Fine-tuning Close the Performance Gap to BERT?

Different sizes of Qwen3 models exhibit significant improvements in both *accuracy* and *R-rank* after fine-tuning in Table 1. Interestingly, the scaling effect of model parameters does not align with the default models' performance trends, and the Qwen3-4B-ft achieves optimal performance in both metrics among the Qwen3 variants.

However, while fine-tuning leads to substantial performance gains, the enhanced Qwen3-4B model only reaches *accuracy* levels comparable to GPT-4, still falling significantly short of BERT's performance. Furthermore, when applying the same fine-tuning procedure to BERT, we observe an inverse relationship between the two metrics. Despite this, the fine-tuned LLMs fail to match BERT's performance in either metric, suggesting that fine-tuning alone may be insufficient to overcome LLMs' inherent limitations in classifier prediction tasks.

4.3 Can LLMs Balance Prediction Performance among Classifier Types?

While LLMs currently trail BERT in overall performance, their potential to leverage vast pre-training data to address BERT's key limitations, particularly inconsistent performance across task types and weaker fine-grained semantic discrimination, warrants further investigation. This motivates our detailed analysis of classifier performance across different task types and models.

As illustrated in Figure 3, we evaluate models' *accuracy* per classifier type (*R-rank* with similar trends). Contrary to expectations, LLMs fail to perform more balanced or superior semantic precision and R-rank than BERT despite their broader pretraining; in many cases, they lag behind.

For sortal classifiers, the individual classifiers yield the highest *accuracy* across models, likely due to their reliance on explicit head-noun features, straightforward classification logic, and high frequency in training data. However, event classifiers reveal only a marginal gap between BERT and LLMs, suggesting comparable challenges in modeling event semantics for both of them. Notably, BERT's strong performance in kind classifiers, paired with LLMs' decline, highlights the latter's typological understanding deficits.

The performance of LMs' measure classifiers reveals an important distinction, while standard measure classifiers achieve relatively strong results across all models due to their rigid syntactic patterns, both BERT and LLMs struggle with container and approximate classifiers. This performance dichotomy suggests that while models can effectively learn predictable, formulaic relationships, they face fundamental challenges in modeling more complex items like the container-contents relationship and quantifying abstract concepts.

The similar performance patterns between LLMs and BERT, coupled with LLMs' overall weaker results, suggest that neither their expanded pretraining data scale nor their enhanced capabilities from larger parameters lead to improved prediction performance. This persistent performance gap may warrant further investigation into architectural differences, particularly in attention mechanisms.

4.4 How LMs' Attention Mechanisms Contribute to Prediction?

Attention Mask Type	Token Visibility Pattern
Standard	[CLS] 那 里 有 一 [MASK] 猫 吗 ? [SEP]
Mask After Head Noun	[CLS] 那 里 有 一 [MASK] 猫 0 0 [SEP]
Context Mask	[CLS] 0 0 0 0 [MASK] 猫 吗 ? [SEP]
Head Noun Mask	[CLS] 那 里 有 一 [MASK] 0 吗 ? [SEP]
Mask After Classifier	[CLS] 那 里 有 一 [MASK] 0 0 0 [SEP]

Table 2: Token visibility patterns under different masking strategy types. The positions with 0s correspond to tokens in the input for which the model's attention is blocked(instead of 0 in the text sequence).

Given the classifiers' strong dependence on their head nouns, the differences in attention mechanisms between BERT and LLMs, and the above analysis, we further investigate how the architectural distinctions account for the gaps. We select BERT as our baseline reference and employ 4 different attention masking types distinct from the standard attention masking for BERT MLM. Examples and comparisons are shown in Figure 3.

Inspired by Metheniti et al. (2020), we design four masking strategies by zeroing out tokens in BERT's attention mask, shown in Table 2.

The results of adjusting attention masking in Table 3 show an obvious decline trend in both *accuracy* and *R-rank*. The minor decrease in performance for Mask After Head Noun and Context Mask indicates that directional contextual information (excluding the head noun) contributes marginally to prediction. Based on the changes

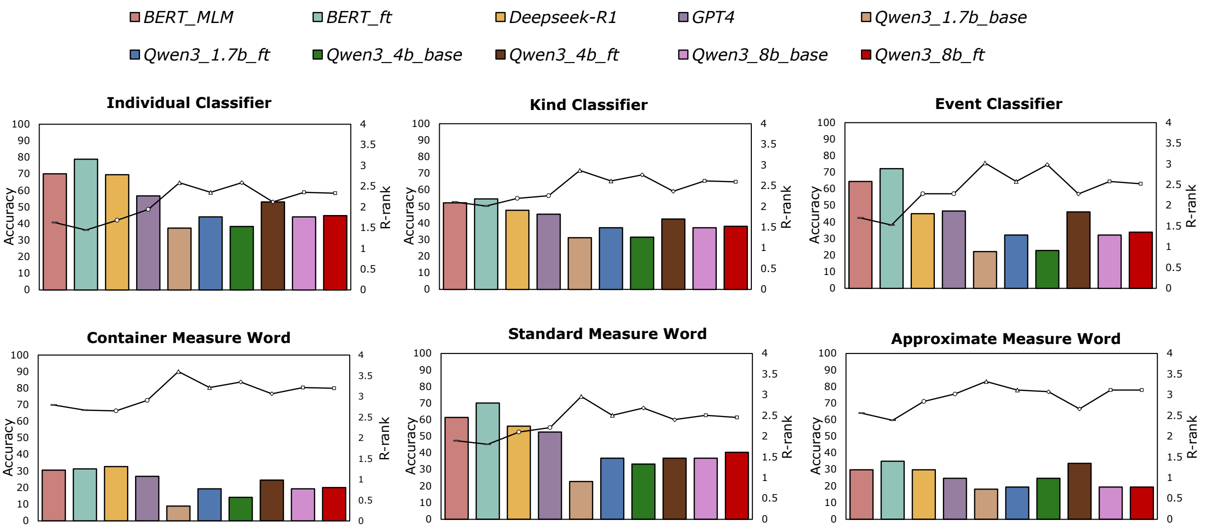

Figure 3: The fine-grained analysis of the six types of classifiers' predictions among the proposed LMs. The black polyline represents the *R-rank* value (the lower, the better); the bar charts in different colors represent the *accuracy* of specific models in this type of classifier (the higher, the better). The "ft" in model names represents that the LMs have applied a fine-tuning strategy.

in both metrics, the text after the head noun has a greater influence on R-rank, while the preceding content affects *accuracy* more significantly.

Attention Mask Type	Accuracy	R-rank
Standard	**62.31**	**1.8298**
Mask After Head Noun	60.92	1.8929
Context Mask	58.35	1.9272
Head Noun Mask	33.19	2.6670
Mask After Classifier	25.59	2.9443

Table 3: Performance for BERT with various attention masking strategies.

When the head noun is masked, the performance plummets compared to the standard condition, highlighting the high dependency of classifier prediction on language models. However, further masking the preceding context reveals an interesting pattern where *accuracy* experiences a significant drop, *R-rank* performance shows a slight rebound. This scenario mirrors LLM's unidirectional attention mechanism. With the scale of 110M parameters, BERT achieves only around 25% *accuracy* and a ranking score near 3. This result underscores the critical role of the bidirectional attention mechanism, which doubles the *accuracy* while reducing the ranking score by one.

The strong dependency on the head noun and the partial dependence on preceding contexts in classifier prediction seem to strictly require bi-directional attention for efficient modeling. This explains why increasing the parameter and training data size fail to compensate for the inherent limitations of the attentional mechanism.

5 Error Case Analysis

Although current LMs, particularly BERT, exhibit strong capabilities in Chinese classifier prediction, their varying performance across different models and various classifier categories underscores persistent challenges. To better understand these error patterns, we systematically analyze two primary types of failures, aiming to empirically investigate the underlying causes of these specific errors.

5.1 Unable to Capture Fine-grained Pragmatic Preferences

Current language models demonstrate systematic shortcomings in aligning with pragmatic preferences when selecting classifiers, consistently favoring statistically frequent but stylistically inappropriate options. As illustrated in Table 4, the models' universal top prediction of "件" (piece) in a colloquial negative-affect context, followed by other generic or semantically mismatched classifiers like "种" (kind) and "回" (occasion), reveals their inability to integrate register, affective tone, and habitual semantics into classifier choice. While all models recognize grammatical validity, they diverge in subsequent errors. Qwen3 persists with generic classi-

Model	Predictions	Carrier Sentence
BERT	件, 种, 回	早起这件事是要多痛苦有多痛苦。
Qwen3	件, 个, 桩	早起这件事是要多痛苦有多痛苦。
GPT-4	件, 桩, 回	早起这件事是要多痛苦有多痛苦。
Ds-r1	件, 桩, 种	早起这件事是要多痛苦有多痛苦。

Table 4: The models' responses demonstrate the failure to capture fine-grained pragmatic preferences. The most appropriate candidate is "档子". BERT and Qwen3 results are selected from the top three results of the base and fine-tuning models with the best performance. Ds-r1 denotes Deepseek-R1. The English translation of the carrier sentence is **"Getting up early, this thing is as painful as it get"**.

fiers, full-parameter LLMs incorrectly shift toward event classifiers, and BERT shows partial awareness of categorical distinctions, yet all share the critical failure to prioritize the pragmatically optimal "档子", which uniquely satisfies colloquialism, negative affect, and abstract habitual semantics.

This consistent neglect of stylistic and affective dimensions underscores that LMs treat classifier selection as a frequency-driven grammatical task rather than a pragmatic negotiation between linguistic constraints and communicative intent. The hierarchy of error from grammatical correctness to semantic coherence to pragmatic appropriateness exposes their inability to progress beyond coarse statistical patterns toward fine-grained sociolinguistic competence.

5.2 Hardly Further Check Whole Context

Current language models demonstrate a concerning tendency to make classifier predictions based on local noun-classifier associations rather than holistic context evaluation. This limitation becomes evident when examining BERT's performance in the raffle scenario, where its top prediction "笔" (pen) reveals a fundamental misunderstanding. While "一笔笔" could theoretically form a plural classifier for money, this interpretation completely disregards the actual context of awarding pens as a prize. Its subsequent predictions, though grammatically correct for describing individual pens, still fail to account for the pragmatic implausibility of awarding just one pen in a raffle setting, a scenario that typically involves more prizes unless explicitly stated otherwise.

Model	Predictions	Carrier Sentence
BERT	笔, 支, 把	后来抽奖,又抽到一笔笔,虽不算好,总比什么都没有的人强。
Qwen3	支, 把, 枝	后来抽奖,又抽到一支笔,虽不算好,总比什么都没有的人强。
GPT-4	支, 件, 份	后来抽奖,又抽到一支笔,虽不算好,总比什么都没有的人强。
Ds-r1	支, 管, 杆	后来抽奖,又抽到一支笔,虽不算好,总比什么都没有的人强。

Table 5: With similar descriptions and settings as Table 4. This table demonstrates the LMs may not check all the context for classifier selection. The English translation of the carrier sentence is **"Later in the raffle, I drew one pen, not great, but better than nothing"**. The proper classifier is "盒"(box). The classifiers' meanings are"支"(stick),"把"(grasp),"枝"(branch),"件"(piece),"份"(portion),"管"(pipe), and "杆"(rod).

The comparative performance of Deepseek-R1, which generated BERTs' subsequent similar candidates. While these properly match the semantic requirements for slender objects, they also overlook the unlikelihood of the single-pen raffle scenario. More alarmingly, Qwen and GPT exhibit even more severe limitations, producing completely unacceptable classifier-noun combinations in their secondary predictions. This degradation in performance highlights how advanced LLMs frequently fail to progress beyond basic noun-classifier matching to consider broader context.

While all models demonstrate basic grammatical competence in noun-classifier pairing, their ability to incorporate pragmatic considerations varies significantly. The most sophisticated models (like Deepseek-R1) at least maintain grammatical accuracy, whereas others (particularly Qwen and GPT) degrade to producing outright errors when forced beyond their primary predictions. This indicates current LMs lack robust mechanisms for contextual integration, instead relying on progressively weaker fallback strategies when their initial predictions prove contextually inadequate. The models' consistent failure to question the plausibility of the single-pen raffle scenario particularly illustrates their limited capacity for real-world reasoning.

6 Conclusions

Our study compares the performance of BERT and LLMs in Chinese classifier prediction, revealing that LLMs still underperform compared to BERT, and highlighting the critical role of attention mechanisms. While advanced LLMs possess strengths such as rich world knowledge and fine-grained semantic sensitivity, our results prove that BERT, with or without fine-tuning, achieves better performance in the task. Strikingly, when preceding attention is masked, BERT's performance declines sharply, even falling below that of Qwen3-1.4B.

This explains why LLMs with extensive knowledge bases still demonstrate a significant performance gap even when using enhanced prompts or fine-tuning. The inherent limitation lies in their unidirectional attention mechanism, which fundamentally constrains their effectiveness for this task. These findings highlight the critical role of bidirectional attention and suggest that future research should focus on new strategies to enhable bidirectional attention in LLMs, in order to combine the strengths of both architectures and advance Chinese classifier prediction performance.

Ethics Statement

We do not foresee any ethical risks related to our research.

Limitations

Though this research provides insights on the comparative performance of BERT and LLMs for Chinese classifier prediction, there are also limitations that should be acknowledged.

First, the evaluation methodology of BERT and LLMs has to be different due to architectural formula differences. Specifically, BERT, as an Encoder, allows single-token log probability retrieval for tokens that have been masked, but Decoder model like DeepSeek cannot produce log probability for a single token and only provide sentence-level average log probability instead. Furthermore, there are LLM APIs that do not even provide token-level log probabilities, thereby inevitably adding dissimilarities in the model being assessed's performance, across architectures.

Second, the log probability measurements utilized in our model forms have been shown to be a function of sentence length and word frequencies. Sentence lengths in this piece were not held fixed, and thus confounding factors may have been introduced as a consequence. Accordingly, differences seen across sentences and models cannot be separated fully from variations in sentence format, with the possible consequence of reducing the objectivity and interpretability of the findings.

Third, the evaluation dataset includes annotation ambiguities, especially in identifying the head noun. Imperfect or inconsistent annotation may corrupt the training as well as the evaluation performance, introducing noise and potential bias into the reported results.

Finally, this study does not explore the full spectrum of fine-grained semantic distinctions present within the Chinese classifier system. Subtle nuances between classifier usage. For example, those dependent on pragmatic or context-specific cues remain under-investigated and could represent important directions for further analysis.

Future work may address these limitations by standardizing evaluation metrics, curating high-quality annotated data, and performing a more in-depth analysis of classifier subtype distinctions.

References

Kathleen Ahrens and Chu-Ren Huang. 2016. Classifiers. In Chu-Ren Huang and Dingxu Shi, editors, *A Reference Grammar of Chinese*, Reference Grammars, pages 169–198. Cambridge University Press, Cambridge, England.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large language Models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.

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, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa-Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. 2018. SemEval-2018 Task 9: Hypernym Discovery. In *Proceedings of SemEval*.

Shiao-hui Chan. 2019. An elephant needs a head but a horse does not: An erp study of classifier-noun agreement in mandarin. *Journal of Neurolinguistics*, 52:100852.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3).

Won Ik Cho, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2021. Modeling the Influence of Verb Aspect on the Activation of Typical Event Locations with BERT. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2922–2929, Online. Association for Computational Linguistics.

William Croft. 1994. Semantic universals in classifier systems. *Word*, 45(2):145–171.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhaoxin Feng, Jianfei Ma, Emmanuele Chersoni, Xiaojing Zhao, and Xiaoyi Bao. 2025. Learning to look at the other side: A semantic probing study of word embeddings in llms with enabled bidirectional attention. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics.

Hui Guo and Huayan Zhong. 2005. Chinese Classifier Assignment Using SVMs. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

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.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. *arXiv preprint arXiv:2305.13264*.

Chu-Ren Huang and Dingxu Shi, editors. 2016. *A Reference Grammar of Chinese*. Cambridge University Press, Cambridge, England.

Shuping Huang and Jenn-Yeu Chen. 2014. The effects of numeral classifiers and taxonomic categories on chinese and english speakers' recall of nouns. *Journal of East Asian Linguistics*, 23(1):27–42.

Jani Järnfors, Guanyi Chen, Kees van Deemter, and Rint Sybesma. 2021. Using BERT for Choosing Classifiers in Mandarin. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 172–176, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. 2024. Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 263–277, Miami, Florida, US. Association for Computational Linguistics.

George Lakoff. 1986. Classifiers as a reflection of mind. In *Noun Classes and Categorization*, pages 13–51. John Benjamins Publishing Company, Amsterdam.

Roger Levy and Christopher D. Manning. 2003. Is it harder to parse Chinese, or the Chinese treebank? In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 439–446, Sapporo, Japan. Association for Computational Linguistics.

Charles N. Li and Sandra A. Thompson. 1989. *Mandarin Chinese: A Functional Reference Grammar*. University of California Press.

Neal Szu-Yen Liang. 2008. The acquisition of chinese shape classifiers by l2 adult learners. In *Proceedings of the 20th North American Conference on Chinese Linguistics (NACCL-20)*, volume 1, pages 309–326.

Khai Jiet Liong, Hongqiu Wu, and Hai Zhao. 2024. Unveiling vulnerability of self-attention. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17225–17236, Torino, Italia. ELRA and ICCL.

Jie Liu. 2018. *The L2 Acquisition of Chinese Classifiers: Comprehension and Production*. Ph.D. thesis, Michigan State University.

Shijia Liu, Hongyuan Mei, Adina Williams, and Ryan Cotterell. 2019. On the idiosyncrasies of the mandarin chinese classifier system. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4100–4106, Minneapolis, Minnesota. Association for Computational Linguistics.

Jianfei Ma, Zhaoxin Feng, Huacheng Song, Emmanuele Chersoni, and Zheng Chen. 2025. Reasoning or memorization? investigating LLMs' capability in restoring Chinese Internet homophones. In *Proceedings of the 3rd Workshop on Towards Knowledgeable Foundation Models (KnowFM)*, pages 120–139, Vienna, Austria. Association for Computational Linguistics.

Eleni Metheniti, Tim Van de Cruys, and Nabil Hathout. 2020. How relevant are selectional preferences for transformer-based language models? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1266–1278, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nicole Peinelt, Maria Liakata, and Shu-Kai Hsieh. 2017. ClassifierGuesser: A Context-based Classifier Prediction System for Chinese Language Learners. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 41–44, Tapei, Taiwan. Association for Computational Linguistics.

Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2022. Discovering Financial Hypernyms by Prompting Masked Language Models. In *Proceedings of the LREC Financial Narrative Processing Workshop*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Edward Ruoyang Shi. 2021. Schizophrenia through the lens of Chinese classifier: A preliminary study. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 392–401, Shanghai, China. Association for Computational Linguistics.

Yeli Shi. 2014. Comparison of individual classifiers and collective classifiers between chinese and english. *Theory and Practice in Language Studies*, 4(9).

Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2025. Repetition improves language model embeddings. In *The Thirteenth International Conference on Learning Representations*.

Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. 2022. Pre-trained language models and their applications. *Engineering (Beijing)*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Meilin Zhan and Roger Levy. 2018. Comparing theories of speaker choice using a model of classifier production in Mandarin Chinese. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1997–2005, New Orleans, Louisiana. Association for Computational Linguistics.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

A Experiment Settings

A.1 Hardware

All experiments were conducted using a single NVIDIA A100 GPU with 40GB of dedicated memory and a single NVIDIA H20 GPU with 96GB of memory, hosted on a system equipped with an AMD EPYC 9K84 96core processor (16 vCPUs) and 150 GB of system RAM. Each experimental run was configured to have a duration exceeding three hours.

A.2 Experiement Setup

This study involves Transformers packages with version 4.55.2 (Hugging Face) and platform of PyTorch with version 2.8.0.

A.3 Prompts Usage

We utilized the the prompt shown in table 6 to make our model inference and fine tune. Due to the task is Chinese classifier prediction, we directly apply Chinese as the target language for prompt, the English version is also attached to the table for reference.

B Instruction and Statistics of Classifier Annotation

For the sake of systematic investigation of Chinese measure words in the experiment, the quantifiers are sorted into six classes: individual classifier, event classifier, kind classifier, container classifier,

Prompt Type	Prompt Text
English	As a professional native speaker of Chinese, please complete the task of filling in the missing measure words. A Chinese sentence lacking a measure word will be input. Please identify and select only one single-character measure word that best fits the position indicated by the underscore “_” in the sentence, according to the rules of Chinese measure word usage. Each sentence contains only one underscore. Please note that the final output should only include the measure word you selected, without any additional information. Sentence lacking a measure word: Output measure word:
Chinese	你作为一个专业的中文母语者，现请你完成补全量词的任务。现在会输入一个缺乏量词的中文句子，请根据中文量词搭配规范，在输入的缺乏量词的句子中，在短下划线“_”的位置，找出且仅找出一个最合适这个位置的单字量词，每一个句子只存在一个下划线，请注意最后的结果只输出你选择的量词，而没有其他的任何信息。 缺乏量词的句子： 输出的量词：

Table 6: The demonstration of prompts used in both the inference and fine-tuning stages is provided below. The English version is a direct, literal translation of the primarily used Chinese version.

standard measure word, and approximate measure word. It was hand-annotated by 3 native Chinese speakers with research background in Chinese linguistics.

To assess the reliability and the objectivity of this annotation, we calculated the Inter-Annotator Agreement (IAA) through the application of Cohen ’ s kappa to a randomly selected sample of 500 classifier tokens. We thereby reached the IAA score of 82.68%, the existence of which indicates high annotator agreement and the robustness of our scheme of classifier categorization. For example, Event Classifier had been used in the counting of action or event occurrences. It does classify events and not physical objects. Therefore, the diction example like "场", "次", "趟" can be categorized into Event Classifier. Finally, we got the statistics about the the Count and Frequency for each classifier type, which can be seen in Table 7. It presents the frequencies of the six classifier types on the test set in numbers. The extreme imbalance of frequency often (e.g., Individual classifiers as the majority, Approximate/Standard classifiers as the minority) offers invaluable background in understanding the fine-grained model performance investigation by classifier type. It indicates the degree of difficulty and availability of data for each of the classifier types during the evaluation process.

Classifier Type	Count	Frequency (%)
Individual classifier	1173	62.79
Kind Classifier	247	13.22
Event Classifier	180	9.64
Container Measure	134	7.17
Standard	57	3.05
Approximate Measure	77	4.12

Table 7: The count and frequency of 6 different types of classifiers in test set

C Chinese Classifier Categories and Explanation

In Chinese, classifiers are a crucial grammatical category used in noun phrases to mark noun classes or quantify nouns. They can be broadly divided into two main types: sortal classifiers and measure words.

Sortal classifiers primarily serve to highlight inherent characteristics of the head noun, categorizing it based on shape, function, or other salient properties. They are further classified into three subtypes:

Individual classifiers classify concrete or abstract objects based on their natural units, such as 只 for animals (e.g., 一只猫 "a cat") or 张 for flat objects (e.g., 一张纸 "a piece of paper").

Event classifiers enumerate occurrences of events, like 次 for instances of actions (e.g., 一次会议 "one meeting") or 场 for performances (e.g., 一场比赛 "a match").

Kind classifiers categorize nouns by type rather than individual instances, such as 种 for kinds (e.g., 三种动物 "three types of animals").

On the other hand, measure words focus on quantifying the noun rather than classifying its inherent features. They include:

Container measure words, which denote quantity based on containers (e.g., 碗 "bowl" in 一碗饭 "a bowl of rice").

Standard measure words, which use fixed units of measurement (e.g.,米 "meter" in 三米长 "three meters long").

Approximation measure words, which indicate vague quantities (e.g.,些 "some" in 一些问题 "some problems").

D Relationship of classifier frequency and accuracy

Group	NoC	Count	Proportion (%)
Group 1	1	677	34.95
Group 2	1	113	5.83
Group 3	1	102	5.27
Group 4	4	120	6.20
Group 5	10	116	5.99
Group 6	15	120	6.20
Group 7	15	120	6.20
Group 8	15	120	6.20
Group 9	15	120	6.20
Group 10	21	118	6.09
Group 11	40	120	6.20
Group 12	34	91	4.70

Table 8: This table demonstrates the rank based on the frequency of classifiers from the highest frequency of group 1 to lowest frequency group 12 and distribution of the classifiers. The NoC(number of categories) demonstrate how many categories of classifiers in this group.

To explore the relationship between classifiers' frequency and accuracy, we sorted and grouped the classifiers that appeared in Chinese Classifier Dataset (Peinelt et al., 2017) based on their frequency of occurrence. With the grouping results summarized in Table 8, we first assign the classifier exhibiting the highest frequency to Group 1. The classifiers with the second and third highest frequencies are then placed into Group 2 and Group 3, respectively. To achieve a more balanced distribution across categories, and considering that Groups 2 and 3 are of appropriate scale to accommodate additional classifiers, we establish a count threshold of 120 for group assignment. Any remaining classifiers that meet this threshold are sequentially allocated to subsequent groups in the processing order. We then separately aggregated the results of all applied models on the test set according to these groupings, as shown in Figure 4. It can be observed that, overall, the accuracy of language models in predicting quantifiers decreases as the frequency of quantifiers in the dataset declines, indicating the prediction of classifier (or LogProb ranking) are influenced by the word frequency.

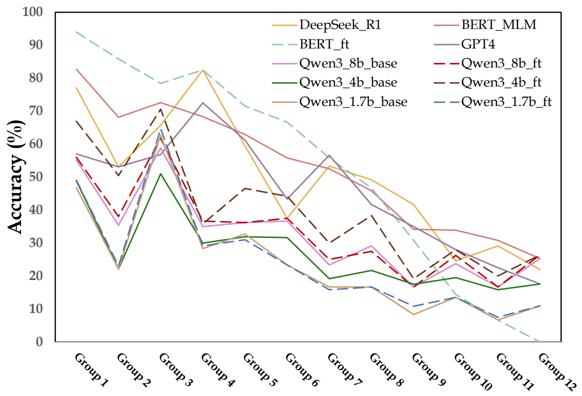

Figure 4: Line chart of accuracy for used models in groups divided by classifier frequency from high to low. The dashed line represents the fine-tuned model, and the solid line represents the original model.