Optimizing Chinese Lexical Simplification Across Word Types: A Hybrid Approach

Zihao Xiao[1,4], **Jiefu Gong**[2,3], **Shijin Wang**[2,3] , **Wei Song**[1]
[1]Information Engineering College, Capital Normal University, Beijing, China
[2]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, China
[3]iFLYTEK AI Research (Hebei), Langfang, China
[4]Department of Internet of Things, China Mobile Research, Beijing, China
{zihao, wsong}@cnu.edu.cn, {jfgong, sjwang3}@iflytek.com

Abstract

This paper addresses the task of Chinese Lexical Simplification (CLS). A key challenge in CLS is the scarcity of data resources. We begin by evaluating the performance of various language models at different scales in unsupervised and few-shot settings, finding that their effectiveness is sensitive to word types. Expensive large language models (LLMs), such as GPT-4, outperform small models in simplifying complex content words and Chinese idioms from the dictionary. To take advantage of this, we propose an automatic knowledge distillation framework called **PivotKD** for generating training data to fine-tune small models. In addition, all models face difficulties with out-of-dictionary (OOD) words such as internet slang. To address this, we implement a retrieval-based interpretation augmentation (RIA) strategy, injecting word interpretations from external resources into the context. Experimental results demonstrate that fine-tuned small models outperform GPT-4 in simplifying complex content words and Chinese idioms. Additionally, the RIA strategy enhances the performance of most models, particularly in handling OOD words. Our findings suggest that a hybrid approach could optimize CLS performance while managing inference costs. This would involve configuring choices such as model scale, linguistic resources, and the use of RIA based on specific word types to strike an ideal balance.

1 Introduction

Lexical simplification (LS) is the task of replacing complex words in a sentence with simpler alternatives while preserving the original meaning and structure. LS improves text readability, benefiting a wide range of people, such as students (De Belder and Moens, 2010), nonnative speakers (Paetzold and Specia, 2016), and individuals with cognitive impairments (Saggion, 2017). However, LS is a challenging task that requires both linguistic knowledge and contextual awareness.

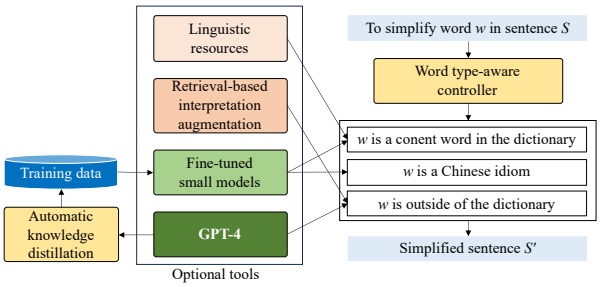

Figure 1: The general framework of the proposed word type-aware Chinese lexical simplification method.

This paper focuses on Chinese Lexical Simplification (CLS). One big barrier for CLS is the scarcity of training data. Consequently, recent research has focused on unsupervised methods that are based on pre-trained language models (PLMs). For instance, BERT-LS (Qiang et al., 2021) utilizes the pre-trained masked language model (MLM) BERT (Devlin et al., 2018), masking target words and generating simpler candidate alternatives for replacement. Despite its simplicity, BERT-LS struggles to fully understand the task, resulting in conservative substitutions, the requirement of complex post-processing and performance bottleneck.

We observe that recent large language models (LLMs), such as GPT-4 (Achiam et al., 2023), exhibit superior task comprehension through task-specific instructions and a few examples. In contrast, small models (defined as ranging from 700M to 7B parameters in our study) struggle to achieve comparable overall performance, indicating the importance of model scale. However, the inference cost associated with GPT-4 is substantial, which presents a trade-off between performance and cost when comparing small and large models.

We aim to improve the performance of small models by learning from and collaborating with large models. Thus we can build a CLS system that delivers competitive performance while significantly reducing inference cost. To achieve this, we

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 15227–15239
November 12-16, 2024

present the following contributions:

- First, we analyze CLS methods based on language models of varying parameter scales to better understand their respective strengths and limitations. We find that the models' performance is largely correlated with word types, such as complex content words, Chinese idioms from the dictionary, and out-of-dictionary (OOD) words. GPT-4 demonstrates superior task comprehension and performance than small models. All models exhibit a need for improvement in handling OOD words.

- Second, we propose a knowledge distillation framework called **PivotKD** for simplifying complex content words and Chinese idioms from the dictionary. PivotKD samples pivot words from the dictionary and employs GPT-4 to create sentences containing these words. These words are then automatically replaced with alternatives of varying complexity. Evaluation results show that small models fine-tuned on the generated dataset outperform GPT-4 in simplifying the targeted word types, demonstrating the effectiveness of PivotKD.

- Third, we propose a retrieval-based interpretation augmentation (RIA) strategy to enhance simplification of OOD words. This strategy involves querying the search engine to acquire an interpretation of the target words, which is then incorporated into the context. Experimental results show that all models exhibit improvements in simplifying OOD words compared to their counterparts without RIA.

Our research suggests that based on the type of complex words, we can configure choices such as model scale, linguistic resources, and the use of RIA to strike an ideal balance between performance and inference efficiency. The codes and resources can be found at https://github.com/cnunlp/hybcls.

2 Related Work

Lexical simplification primarily follows a pipeline consisting of three main stages: the identification of complex words, the generation of candidate substitutes, and the selection and ranking of candidates.

The identification of complex words aims to determine which words are considered complex in a sentence by a specific target population (Shardlow, 2013; Yimam et al., 2018; Dehghan et al., 2022). Aligning with current baseline methods, we do not focus on this stage, since complex words are given in the used dataset.

Knowledge-based methods Early research on lexical simplification relied on lexical knowledge databases to generate substitutes (Carroll et al., 1998; Drndarevic and Saggion, 2012). However, databases are not only expensive to develop and maintain, but also limited in word coverage.

Word embedding-based methods With the advent of deep learning, semantic similarity computation based on word embeddings has become a popular method for substitute generation and ranking (Paetzold and Specia, 2017). The cost of training word embedding models is lower than that of constructing knowledge databases, and word embeddings also alleviate the problem of word coverage.

PLM-based methods Subsequently, pre-trained language models show strong ability in capturing contextual semantic information, and have been used for lexical simplification. For example, BERT-LS (Qiang et al., 2020) introduced an unsupervised method that employs BERT to generate substitutions for complex words based on the encoding of surrounding context. PromptLS (Vásquez-Rodríguez et al., 2022) found that fine-tuning PLMs can achieve better performance compared to unsupervised approaches. ConLS (Sheang et al., 2022) fine-tuned an encoder-decoder model for substitute generation, which naturally predicts simple words with multiple tokens. One major challenge in fine-tuning is the limited availability of supervised training data for certain languages, like Chinese. Xiao et al. (2024) proposed a prompt tuning-based method to alleviate this.

LLM-based methods Recent study shows that GPT-3 is capable of comprehending the task and learning task instructions with a few demonstrations, achieving good performance in English lexical Simplification (Aumiller and Gertz, 2022). GPT-4 has also been shown to be effective in assessing lexical complexity and simplifying complex words in a high-quality multilingual context despite being costly (Enomoto et al., 2024).

Knowledge distillation Knowledge distillation (KD) aims to enhance the performance of a smaller student model by leveraging the knowledge from a larger teacher model (Kim and Rush, 2016). In this study, we adopt a black-box KD method to utilize GPT-4 to obtain high-quality training data for CLS.

3 Task, Data and Preliminary Analysis

In this section, we provide a brief introduction to task formulation, dataset, and evaluation metrics, followed by an analysis of representative baselines that use BERT and LLMs.

3.1 Lexical Simplification Settings

Considering that the identification of complex words depends on a target population, we assume that a sentence and a target complex word are given following previous work (Qiang et al., 2021) and focus on substitute generation.

Formally, given a sentence S and a complex word w in S, the task is to generate a simpler alternative v, a word or a phrase, to form a simpler sentence S', which is expected to be smooth, clear and maintain the same meaning as S.

3.2 Dataset and Metrics

3.2.1 Dataset

We use the publicly available Chinese lexical simplification dataset HanLS (Qiang et al., 2021). HanLS includes 524 sentences, each containing a complex word from the advanced level of the Chinese Proficiency Test (Hanyu Shuiping Kaoshi, HSK), and each complex word has 8.51 annotated simple substitutes on average as reference answers.

3.2.2 Evaluation Metrics

Following previous work (Paetzold and Specia, 2016; Qiang et al., 2021), we use precision and accuracy as metrics.

Precision (PRE): The proportion of predicted substitutes that are the original complex word itself or appear in the reference answers.

Accuracy (ACC): The proportion of predicted substitutes that are different from the original complex word and appear in the reference answers.

A higher PRE indicates a lower probability of predicting misleading or incorrect words, reflecting the system's robustness. Considering a conservative system may retain a large number of original words to achieve high PRE, thus ACC is involved to measure its simplification ability.

HanLS uses individual words as reference points; however, LLMs may generate phrases even when instructed to generate a word, leading to potential inconsistencies and unfair comparisons. To address this, we also introduce **fuzzy-PRE (f-PRE)** and **fuzzy-ACC (f-ACC)** as complementary metrics. A prediction is considered correct if it contains any of the words in the reference answer list as a substring.

Task instruction		请你根据任务要求执行词汇简化任务，以下是任务示例 Please perform the lexical simplification task according to the task requirements. Examples are as follows.
Demonstration (few-shot)	Q	句子：她下班时我们为她端上一杯浓浓的咖啡，说句#温馨#的话语。 要求：针对句子中的#温馨#，请你给出一个能在句子中将其流畅替换且含义相同的词。 Sentence: When she finishes work, we serve her a cup of strong coffee and say a # cordial # word. Requirement: For the word # cordial #, please provide a simpler word that can be smoothly replaced in the sentence with the same meaning.
	R	回答：#温暖# Response: # warm #
		[Provide the test Q and let LLM generate R]

Figure 2: An example of instruction and demonstration design for prompting LLMs for CLS.

3.3 Baseline Systems

We adopt BERT-LS (Qiang et al., 2021) along with several dialogue models of varying scales using few-shot learning methods as baselines.

3.3.1 BERT-LS

The input of BERT-LS is formed by concatenating the original sentence with its copy, where the target complex word is replaced by the [MASK] token. BERT then predicts potential substitutes for the masked positions.

Since a Chinese word often consists of multiple Chinese characters and BERT's tokenizer operates at character level, BERT-LS accommodates predictions with varying numbers of [MASK] tokens (e.g., one to four). If the complex word is listed in the Chinese synonymy thesaurus (Mei, 1983), its synonyms are used as substitutes. Finally, BERT-LS ranks these substitutes by leveraging multiple sources of evidence, including similarity based on word embeddings, fluency measured by BERT scores, and word frequencies.

3.3.2 Dialogue models

We use GPT-4 and three open-source small Chinese dialogue models: Qwen1.5-7B-Chat (Qwen-Chat for short) (QwenTeam, 2024), ChatGLM2-6B (ChatGLM for short) (Du et al., 2022) and ChatYuan-large-v2 (700m parameters, ChatYuan for short) (Xuanwei Zhang and Zhao, 2022) for comparison. As in the example shown in Figure 2, we explore their performance through few-shot learning, incorporating task instructions and two

Models	PRE	ACC	f-PRE	f-ACC
BERT-LS	80.7	70.4	81.9	71.6
ChatYuan (0.7B)	59.7	31.1	63.2	34.5
ChatGLM (6B)	47.1	46.9	58.6	58.4
Qwen-Chat (7B)	65.5	65.5	74.0	74.0
GPT-4	**73.1**	**73.1**	**78.4**	**78.4**

Table 1: The overall results of various models.

demonstrations within the context. We extract predictions directly from their responses.

3.4 Analysis and Discussion

3.4.1 Overall Results

Table 1 shows the overall results. ChatYuan does not perform much simplification, as shown by its high PRE scores and low ACC scores, indicating a poor understanding of the task. ChatGLM, Qwen-Chat, and GPT-4 demonstrate a better understanding of the task. Their performance is in correlation with the model scales. BERT-LS also achieves impressive results, however, it heavily depends on external linguistic resources to rank substitutes. In general, GPT-4 demonstrates the best capabilities in task understanding and prediction.

3.4.2 Analysis

We examine the relationship between model performance and the types of complex words. We categorize complex words into three types:

- **Content words from the dictionary**: Refer to complex content words listed in the Chinese Xinhua dictionary[1].
- **Chinese idioms**: Idioms or Chengyu, an crucial component of the Chinese language, typically composed of four Chinese characters that convey a moral or lesson in a concise and elegant manner. Chinese idioms are also included in the Xinhua dictionary.
- **Out-of-dictionary (OOD) words**: Refer to words not included in the Xinhua dictionary, primarily consisting of new vocabulary such as internet slang.

The three types of complex words can be easily distinguished on the basis of the Xinhua dictionary.

Table 2 presents the performance of different models on these types of complex words. GPT-4 surpasses other models in simplifying complex

[1]Digital resource of the Xinhua dictionary is available at https://github.com/pwxcoo/chinese-xinhua, which covers more than 320k words.

Original Sentence	提起这个题目我又要提起一段往事——后来每每回想起来就让我有些#黯然神伤#的往事。 Bringing up this topic, I am reminded of a past event—a memory that, whenever I recall it, leaves me feeling somewhat # melancholic #.
BERT-LS	提起这个题目我又要提起一段往事——后来每每回想起来就让我有些#痛苦#的往事。 # painful #.
GPT-4	提起这个题目我又要提起一段往事——后来每每回想起来就让我有些#心情低落#的往事。 # feeling low #

Figure 3: The outputs of BERT-LS and GPT-4 on simplifying a Chinese idiom.

content words and demonstrates a remarkable advantage in simplifying OOD words, but lags behind BERT-LS in handling Chinese idioms measured by PRE and ACC. This discrepancy arises because GPT-4 tends to generate more phrases, whereas BERT-LS predominantly predicts single words, aligning with the format of annotated gold answers. Figure 3 shows an example. The simplified sentence containing the GPT-4-generated phrase is also reasonable, with minimal semantic loss. Consequently, when evaluated using the fuzzy-ACC metric, GPT-4 and BERT-LS have comparable performance.

The performance of the three small dialogue models falls short compared to both GPT-4 and BERT-LS in simplifying content words and idioms from the dictionary. Given that these models should theoretically surpass BERT in capability, the linguistic resources employed by BERT-LS likely play an important role in boosting its performance. To further assist these models, we incorporate the Chinese synonymy thesaurus. Each model generates the top 10 candidates using beam search, and the highest-ranked synonym of the complex word, if available, is chosen as the substitute. Table 2 presents the results of this modification (denoted as +synonymy). The performance in simplifying complex content words and Chinese idioms covered by the thesaurus is markedly improved. However, the advantage may be overrated since a considerable portion of the words in HanLS can be covered by the thesaurus.

Figure 4 shows an challenging example of simplifying OOD words. The term "magnesium-aluminum" is a Chinese internet slang that sounds like "beauty" and refers to beautiful women. Both GPT-4 and small models fail to provide a suitable answer, probably because they have limited knowledge of the OOD word. GPT-4 and Qwen-Chat out-

Models	Content Words				Chinese Idioms				OOD Words			
	PRE	ACC	f-PRE	f-ACC	PRE	ACC	f-PRE	f-ACC	PRE	ACC	f-PRE	f-ACC
BERT-LS	**86.8**	77.0	**87.9**	78.1	**70.8**	**41.7**	**70.8**	**41.7**	34.0	28.3	35.8	30.2
ChatYuan	61.3	33.8	64.7	37.1	58.3	16.7	66.7	25.0	47.2	15.1	49.1	17.0
+synonymy	74.5	65.1	78.1	68.7	54.2	25.0	62.5	33.3	47.2	15.1	49.1	17.0
ChatGLM	49.4	49.2	60.4	60.2	20.8	20.8	41.7	**41.7**	39.6	39.6	50.9	50.9
+synonymy	66.4	66.4	73.8	73.8	25.0	25.0	45.8	45.8	39.6	39.6	50.9	50.9
Qwen-Chat	68.9	68.9	77.2	77.2	20.8	20.8	29.2	29.2	56.6	56.6	**67.9**	**67.9**
+synonymy	82.1	82.1	87.7	87.7	29.2	29.2	37.5	37.5	56.6	56.6	67.9	67.9
GPT-4	77.4	**77.4**	81.7	**81.7**	12.5	12.5	41.7	**41.7**	**64.2**	**64.2**	**67.9**	**67.9**

Table 2: Detailed results of simplifying three types of complex words by various models. +synonymy indicates using the Chinese synonymy thesaurus as an external resource for substitute ranking. Bold figures reflect the best inherent capability of the models, and underlines indicate overall best performance.

Original sentence	我最近网上冲浪的时候总能刷到好多#镁铝#哦！ I always see a lot of # magnesium aluminum # when I surf the internet recently!
BERT-LS	我最近网上冲浪的时候总能刷到好多#金属#哦！ I always see a lot of # metal # when I surf the internet recently!
GPT-4	我最近网上冲浪的时候总能刷到好多#热门话题#哦！ I always see a lot of # hot topics # when I surf the internet recently!

Figure 4: The outputs of BERT-LS and GPT-4 on simplifying an OOD word.

perform other models in simplifying OOD words, likely because they are trained on larger, more recent datasets and benefit from having more parameters. However, there is still significant room for improvement in simplifying OOD words.

In summary, our analysis reveals the following observations: 1) Small models like ChatYuan and BERT struggle with task understanding when not given sufficient supervision. Although ChatGLM and Qwen-Chat demonstrate better task understanding, their simplification capability in few-shot settings remains limited. 2) Linguistic resources, such as the synonymy thesaurus, are crucial for enhancing the performance of models with limited capacity, provided they are applicable. 3) Model performance is strongly correlated with scale, with GPT-4 having the best inherent capability. However, simplifying Chinese idioms and out-of-dictionary (OOD) words remains a challenge for all models.

This analysis raises several key questions for further exploration: 1) How can we leverage GPT-4's capabilities to enhance the performance of smaller models? 2) What strategies can be employed to improve models' ability to simplify OOD words effectively? 3) How can we optimize the configuration of models and settings to strike the best balance between performance and inference cost?

4 The Proposed Method

We propose a framework that consists of three key modules: automatic knowledge distillation, retrieval-based interpretation augmentation, and a word type-aware controller.

4.1 Automatic Knowledge Distillation

GPT-4 demonstrates proficiency in simplifying complex words from the dictionary, suggesting a strong understanding of lexical complexity. Our objective is to develop a high-quality training dataset for CLS by distilling the knowledge of GPT-4.

Specifically, we propose an automatic knowledge distillation strategy named **PivotKD**, which does not require human intervention. Figure 5 illustrates its main workflow.

4.1.1 Pivot Word Sampling

We sample words from the Xinhua dictionary and refer to these sampled words as *pivot words*. To enhance diversity, each word can be sampled only once.

4.1.2 Pivot Sentence Generation

We instruct GPT-4 to generate a sentence containing a pivot word, which leverages the strengths of GPT-4 in the following aspects: 1) GPT-4 is capable of generating correct and fluent sentences, thereby avoiding the spelling and grammar errors commonly found in sentences collected from the web or existing corpus; 2) GPT-4 can generate sentences covering diverse topics since we do not constrain topics during the sampling of pivot words and sentence generation. We can assume that the generated dataset is topic-independent.

4.1.3 Multi-level Lexical Substitution

Following the generation of a pivot sentence with a pivot word, we then direct GPT-4 to generate sub-

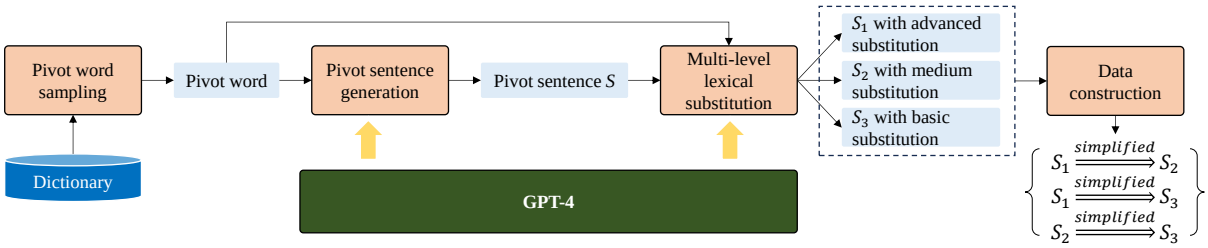

Figure 5: The main workflow of the PivotKD framework for generating CLS data based on GPT-4.

Instruction	[Provide the standard of word complexity levels] 请根据上述词汇难度等级划分标准，完成以下任务： 1. 给出#[中心词]#的词汇难度等级。 2. 任意生成一个包含#[中心词]#的句子。 3. 针对#[中心词]#，为每个词汇难度等级给出n个不重复的替换词，满足替换后的句子语义一致且流畅通顺。 Please complete the following tasks according to the above word complexity standard: 1. Evaluate the word complexity level for the # [Pivot Word] #. 2. Generate a sentence containing the # [Pivot Word] #. 3. Generate n unique substitutes of the # [Pivot Word] # for each word complexity level, ensuring each of the replaced sentence maintains semantic consistency and flows smoothly.
Response	[Let LLM generate the response]

Figure 6: An instruction for 3-level lexical substitution.

stitutes at three distinct levels of word complexity (*basic*, *medium*, and *advanced*) to replace the pivot word in the sentence. The advantage is that there are no constraints on the complexity of the pivot word, as it primarily serves to provide semantic guidance. We convey the requirements to GPT-4 through the instruction as shown in Figure 6. The detailed standard of words complexity levels are as follows:

- **Basic:** Refer to words that are frequently used, with simple and clear meanings that are easy to understand, usually encountered by children and beginners.
- **Medium:** Refer to words that are moderately used for general communication and writing.
- **Advanced:** Refer to words that are infrequently used, which are more complex and precise, and suitable for professional fields, literary works, or advanced writing.

We allow GPT-4 to generate a word as a substitute for complex content words from the dictionary while a word or a phrase as a substitute for Chinese idioms and OOD words.

4.1.4 Data Construction

For one pivot word, we employ GPT-4 to generate n substitutes across each level of word complexity. Then we construct a set of sentence pairs based on the *complex-to-simple* criteria. Specifically, a pair of sentences (S, S') is chosen if the level of complexity of the substitute in S is higher than that of S'.

4.1.5 Instruction Fine-tuning

We conduct instruction fine-tuning with small models. For ChatYuan, all parameters are fine-tuned, while LoRA (Hu et al., 2021) is used to fine-tune Qwen-Chat and ChatGLM. The training data, derived from the constructed sentence pairs $\{(S, S')\}$, is transformed into a question-response format $\{Q \rightarrow R\}$.

Specifically, the input question is in the form of $Q = (I, S, w)$, where I is a task-specific instruction; S is the target sentence, and w is the target complex word with the tag #. The output response R is the simplified substitute w' instead of the whole sentence S', since other words in the original sentence are not necessarily changed.

4.2 Retrieval-based Interpretation Augmentation

OOD words present a huge challenge for simplification. The majority of OOD words are internet slang. New internet slang emerge continuously, while pre-trained models remain static and may lack relevant knowledge. Motivated by recent work on retrieval-augmented LLMs (Lewis et al., 2020; Nakano et al., 2021), we propose a retrieval-based interpretation augmentation (RIA) approach that dynamically collect word interpretations from the web to alleviate the knowledge gap.

Retrieving Interpretation from Web Online interpretations for new vocabulary may be accessible. We retrieve the search results from the Google search engine using the query "[*complex word*] *meaning*", and extract the content of the top k snippets containing the word as its interpretation.

Retrieving Interpretation from Dictionary For words from the dictionary, we can also apply the

RIA strategy by fetching their precise definitions, as these definitions may provide useful information.

We fine-tune small models with the interpretation augmented data, where the obtained interpretation M is injected into the input question during both training and inference, i.e., changing $Q = (I, S, w)$ to $Q = (I, S, w, M)$. For GPT-4, we use the same interpretation augmented question during inference. Therefore, RIA offers flexibility and ease of integration as a plug-in.

4.3 Word-type aware Controlled Inference

Currently, we have several options for CLS, including GPT-4 or small models, and the incorporation of linguistic resources and retrieval-based interpretation augmentation. We aim to find an effective and efficient way for integrating these components to maintain a balance between performance and inference cost.

Our solution is based on the observation that the performance of CLS is sensitive to the type of complex words. We suggest using small models as the basic model, supplemented by GPT-4 or retrieval-based interpretation augmentation to handle Chinese idioms or OOD words. The following evaluation section will discuss optimal strategies for different types of complex words.

5 Evaluation

5.1 Experimental Settings

The Distilled Dataset For PivotKD, we sampled 5,000 pivot words from the Xinhua dictionary and avoided using complex words in HanLS as pivot words to prevent data leakage. Since the sentences are entirely generated by GPT-4, there is no overlap with HanLS, which comprises 447 complex content words, 24 Chinese idioms, and 53 OOD words.

The multi-level lexical substitution module generates $n = 1$ substitution for each level of word complexity. Finally, we established a training dataset comprising 12,478 distinct complex-to-simple substitution pairs. More details of the data can be found in Appendix A.

Additionally, we conducted a human evaluation on 500 samples from the dataset. For each sentence pair, we establish criteria to evaluate the relative word complexity, categorizing it as *clearly reasonable*, *hard to distinguish*, or *contradiction or irrelevant*. The annotation results are 80%, 17%, and 3%, respectively, indicating that the quality of the dataset produced by PivotKD is acceptable.

Model Fine-tuning We fine-tuned ChatYuan for 1 epoch, and fine-tuned Qwen-Chat and ChatGLM with LoRA for 3 epochs. Detailed settings can be found in Appendix B. We report the average performance of three different seeds to reduce randomness.

RIA We used top $k = 1$ snippet from the Google search engine for OOD words since it already provides satisfactory performance. We used word definitions from the Xinhua dictionary for complex content words and Chinese idioms.

5.2 Experimental Results

5.2.1 Auto-Evaluation

Table 3 presents the overall results of various models and variants, specified with the separated results on three types of complex words. We report the ACC and f-ACC scores, as these metrics best reflect the effectiveness of actual simplifications. We observe several trends:

1) **The effect of PivotKD** Let us see the results of the fine-tuned small models without the use of external resources, which reflect their inherent simplification capability. The small models greatly benefit from supervised instruction fine-tuning based on the training data provided by PivotKD. The smallest model ChatYuan exhibits better understanding of the task and achieves a 33% improvement in ACC and f-ACC compared to its unsupervised frozen counterpart. ChatGLM and Qwen-Chat also obtain significant improvements and outperform BERT-LS. Qwen-Chat even surpasses GPT-4 by a wide margin in simplifying complex content words and Chinese idioms from the dictionary. Overall, larger-scale fine-tuned small models perform better, but they still struggle with OOD words.

2) **The effect of the Chinese synonymy thesaurus** The Chinese synonymy thesaurus continues to aid in simplifying complex content words and Chinese idioms, though the improvement is smaller compared to that seen in the frozen models. This implies that supervised fine-tuning facilitates models in acquiring knowledge pertaining to synonymy relationships.

3) **The effect of RIA** The application of RIA significantly boosts the performance of all models in simplifying OOD words, demonstrating that the retrieved word interpretations provide valuable in-

Models	Content Words		Chinese Idioms		OOD Words		All	
	ACC	f-ACC	ACC	f-ACC	ACC	f-ACC	ACC	f-ACC
BERT-LS	77.0	78.1	41.7	47.1	28.3	30.2	70.4	71.6
GPT-4 (frozen)	77.4	81.7	12.5	41.7	64.2	67.9	73.1	78.4
+RIA	77.0	86.4	25.0	45.8	**73.6**	**84.9**	74.2	84.4
ChatYuan (frozen)	33.8	37.1	16.7	25.0	15.1	17.0	31.1	34.5
ChatYuan (full-tuning)	70.0	72.7	33.3	54.2	44.0	45.9	65.6	69.2
+RIA	74.1	77.1	45.8	63.9	57.9	58.5	71.1	74.6
+synonymy	84.9	87.3	37.5	54.2	44.0	45.9	78.6	81.6
+synonymy+ RIA	87.2	89.3	**52.8**	**65.3**	57.9	58.5	82.6	85.1
ChatGLM (frozen)	49.2	60.2	20.8	41.7	39.6	50.9	46.9	58.4
ChatGLM (LoRA)	74.3	77.6	29.2	58.3	52.8	54.7	70.0	74.4
+RIA	73.4	77.0	27.8	54.2	68.6	72.3	71.1	75.5
+synonymy	85.9	88.1	41.7	62.5	52.8	54.7	80.5	83.6
+synonymy+ RIA	86.7	88.2	33.3	52.8	68.6	72.3	82.4	84.9
Qwen-Chat (frozen)	68.9	77.2	20.8	29.2	56.6	67.9	65.5	74.0
Qwen-Chat (LoRA)	79.1	82.8	29.2	52.8	59.1	66.7	74.8	79.8
+RIA	78.7	81.8	26.4	50.0	62.9	71.1	74.7	79.3
+synonymy	**89.3**	**91.9**	36.1	54.2	59.1	66.7	83.8	**87.6**
+synonymy+ RIA	89.1	90.9	40.3	57.0	62.9	71.1	**84.2**	87.4

Table 3: System comparisons on HanLS. RIA indicates utilizing the retrieval-based interpretation augmentation strategy. The results with the highest accuracy are bolded, and the best results obtained by each model on different types of complex words are marked with underlines.

formation that models can effectively utilize. With RIA, GPT-4 excels in simplifying OOD words, highlighting its strong ability to integrate external information into context. For simplifying complex content words and Chinese idioms from the dictionary, ChatYuan benefits from RIA, whereas larger models like ChatGLM and Qwen-Chat see no further improvement, likely because they already possess a strong understanding of word definitions.

The combination of the synonymy thesaurus and RIA proves to be more effective in simplifying Chinese idioms than it is for complex content words.

A hybrid approach Based on the results and analysis, we suggest optimizing CLS by selecting the best configuration in terms of model choice, and the use of the synonymy thesaurus and RIA, tailored to the type of complex words.

1) **For simplifying complex content words and Chinese idioms from the dictionary, fine-tuned small models are preferred**, but the use of linguistic resources and RIA depends on the choice of specific small models.

Models with approximately 6-7 billion parameters, such as ChatGLM and Qwen-Chat, can achieve competitive, and sometimes even superior, performance compared to GPT-4. When the target complex words are included in the synonymy thesaurus, this resource should be utilized to effectively select the best candidates. While RIA is beneficial for simplifying Chinese idioms, it may be less necessary for complex content words, as it still incurs inference costs.

For small models with fewer than 1 billion parameters, such as ChatYuan, both linguistic resources and RIA should be applied.

2) **For simplifying OOD words, GPT-4 with RIA is the best choice**. Fine-tuned small models with approximately 6-7 billion parameters can obtain moderate performance with the aid of RIA but still leg behind GPT-4 for about 5%-10% in ACC and f-ACC.

The best configuration for each model on each type of complex words are marked with underlines in Table 3.

5.2.2 Human Evaluation

The system outputs may be reasonable but outside the reference answers. So we conduct a human evaluation for GPT-4 and the small models with the best configuration shown in Table 3. We sample 20 complex content words, 20 Chinese idioms, and 20 OOD words from HanLS, rating the mixed outputs of different systems according to the following three evaluation criteria:

- **Complexity reduction**: The substitute is simpler than the complex word.
- **Fluency**: The sentence, when replaced with the substitute, remains fluent and natural.

Models	C	F	S	Overall
BERT-LS	1.50	1.53	1.32	4.35
GPT-4	1.62	**1.95**	**1.88**	**5.45**
ChatYuan	1.60	1.87	1.73	5.20
ChatGLM	**1.79**	1.73	1.63	5.15
Qwen-Chat	**1.80**	1.89	1.78	**5.47**

Table 4: Human evaluation of the models in optimal configurations. The overall rating ranges from 0 (worst) to 6 (best). C, F, S refers to Complexity reduction, Fluency and Semantic consistency, respectively.

- **Semantic consistency**: The substitute expresses the same meaning as the complex word in the sentence.

The score range of each aspect is 0, 1 and 2, with higher scores indicating greater alignment with the evaluation criteria.

Table 4 shows the averaged human evaluation results by two annotators. GPT-4 leads in terms of fluency and semantic consistency, while Qwen-Chat and ChatGLM excel in reducing complexity. Overall, Qwen-Chat achieves competitive performance to GPT-4, making it a strong alternative.

We conducted an error analysis on the top-performing small model, Qwen-Chat with the optimal configuration. We found that most of these errors or less-than-perfect examples are often associated with semantically intricate or emotionally intense words.

Original sentence	人们#大喜过望#，纷纷奔向船只。 People were **# ecstatic #** and rushed toward the ships.
Qwen-Chat	人们#欣喜若狂#，纷纷奔向船只。 People were **# overjoyed #** and rushed toward the ships.

Figure 7: An example of complexity reduction failure.

Figure 7 illustrates an error in complexity reduction where an idiom is replaced with another idiom. The original idiom conveys strong emotional intensity, making it challenging to find a single simple word that fully captures its meaning. As a result, the model selects another idiom with similar meaning and emotional depth as a substitute.

In some instances, the model successfully replaces a word rich in meaning with a simpler one; however, this can lead to semantic loss and negatively impact the fluency of the simplified sentence. An example of this issue is shown in Figure 8, where substitution, although simpler, results in a less fluent sentence. In this case, a phrase with a degree adverb would be a more suitable choice.

Original sentence	这头狮子大约有四五岁，很明显已经#饥肠辘辘#。 The lion is about four or five years old and is **# evidently ravenous #**.
Qwen-Chat	这头狮子大约有四五岁，很明显已经#饥饿#。 The lion is about four or five years old and is **# hungry #**.

Figure 8: An example of fluency degradation.

Original sentence	我已经到达#足球国#了。 I have arrived in the **# land of football #**.
Qwen-Chat	我已经到达#中国#了。 I have arrived in the **# China #**.

Figure 9: An example of hallucination of OOD words.

Additionally, there are instances where complex linguistic phenomena such as metonymy are present. Figure 9 illustrates a case where the OOD phrase "land of football" is colloquially associated with *Brazil* in Chinese internet; however, the model inaccurately associated it with *China*. Comprehending the semantics of this OOD word requires contextual background knowledge, and the top search results returned for it do not provide the correct interpretation.

These findings may suggest that the conventional word-to-word substitution approach for lexical simplification can be enhanced by incorporating more flexible substitutes. Additionally, the evaluation of semantic loss should be considered.

6 Conclusion

This paper presents a word-type aware approach for Chinese lexical simplification. We find that in unsupervised and few-shot settings, GPT-4 can obtain good performance in simplifying complex words from the dictionary, largely outperforming the small models, but still struggle to tackle OOD words. Considering that GPT-4 is costly, we propose an automatic knowledge distillation framework, PivotKD, to generate training data using GPT-4 for fine-tuning small models, which can outperform GPT-4 in simplifying complex words from the dictionary. For addressing the issue of OOD words, we propose a retrieval-based interpretation augmentation strategy, which effectively improves the performance on OOD words. Consequently, we are able to configure the choice in terms of model scale, the use of linguistic resources and the interpretation augmentation strategy according to the type of complex words to strike an optimal balance between performance and cost.

Limitations

There are three possible limitations of this work. First, the HanLS dataset used in our evaluation is limited in size and coverage. We plan to extend the dataset to more realistically and objectively reflect the ability of CLS methods. Second, we assume that GPT-4 understand the lexical difficulty levels, but we verify this assumption by analyzing the relative lexical difficulty between a pair of words in the generated data. More detailed and specially designed probing analysis can be conducted. Third, this paper focuses on Chinese lexical simplification, but the proposed method can be potentially applied to other languages. We plan to address these limitations in the future work.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.62376166). Wei Song is the corresponding author.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Dennis Aumiller and Michael Gertz. 2022. Unihd at tsar-2022 shared task: Is compute all we need for lexical simplification? In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 251–258.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10. Citeseer.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Prroceedings of the SIGIR workshop on accessible search systems*, pages 19–26. ACM; New York.

Mohammad Dehghan, Dhruv Kumar, and Lukasz Golab. 2022. GRS: Combining generation and revision in unsupervised sentence simplification. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 949–960, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Biljana Drndarevic and Horacio Saggion. 2012. Towards automatic lexical simplification in spanish: An empirical study. In *PITR@ NAACL-HLT*, pages 8–16.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Taisei Enomoto, Hwichan Kim, Tosho Hirasawa, Yoshinari Nagai, Ayako Sato, Kyotaro Nakajima, and Mamoru Komachi. 2024. Tmu-hit at mlsp 2024: How well can gpt-4 tackle multilingual lexical simplification? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 590–598.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jiaju Mei. 1983. *Synonymy Thesaurus of Chinese Words*. Shanghai Lexicographical Publishing House, Shanghai.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Gustavo Paetzold and Lucia Specia. 2016. Unsupervised lexical simplification for non-native speakers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Gustavo Paetzold and Lucia Specia. 2017. Lexical simplification with neural ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 34–40.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. Lexical simplification with pretrained encoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8649–8656.

Jipeng Qiang, Xinyu Lu, Yun Li, Yunhao Yuan, and Xindong Wu. 2021. Chinese lexical simplification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1819–1828.

QwenTeam. 2024. Introducing qwen1.5. https://qwenlm.github.io/blog/qwen1.5/.

Horacio Saggion. 2017. Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, 10(1):1–137.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109, Sofia, Bulgaria. Association for Computational Linguistics.

Kim Cheng Sheang, Daniel Ferrés, and Horacio Saggion. 2022. Controllable lexical simplification for english. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 199–206.

Laura Vásquez-Rodríguez, Nhung Nguyen, Matthew Shardlow, and Sophia Ananiadou. 2022. Uom&mmu at tsar-2022 shared task: Prompt learning for lexical simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 218–224.

Zihao Xiao, Miaomiao Cheng, Jiefu Gong, Xu Han, Shijin Wang, and Wei Song. 2024. Chinese lexical simplification based on prompt-tuning. *Journal of Chinese Information Processing*, 38(8):34–43.

Liang Xu Xuanwei Zhang and Kangkang Zhao. 2022. Chatyuan: A large language model for dialogue in chinese and english. https://github.com/clue-ai/ChatYuan.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

Attribute	Value
Complex content words	10,593
Chinese idioms	2,541
Distinct Sentences	4,087
Avg. length of sentences	19.14
Distinct substitutes	6,329
Avg. length of substitutes	2.43

Table 6: Basic statistics of the automatically generated dataset via PivotKD.

A sentence pair	
Complex	他#悄无声息#地走进房间，以免吵醒熟睡中的孩子。 He slipped into the room #stealthily#, so as not to wake thesleeping child.
Simple	他#偷偷#地走进房间，以免吵醒熟睡中的孩子。 #quietly#
A training sample	
Prompt	句子： 他#悄无声息#地走进房间，以免吵醒熟睡中的孩子。 要求： 针对句子中的#悄无声息#，请你给出一个能在句子中将其流畅替换且含义相同的简单词或短语。 Sentence: Same as the sentence above Requirement: For the word #stealthily#, please provide a simpler word or phrase that can be smoothly replaced in the sentence with the same meaning.
Response	偷偷 quietly.

Table 5: An example of a constructed sentence pair and the corresponding training sample.

A More Details in Dataset Construction

We sampled 5,000 pivot words for knowledge distillation of GPT-4 via PivotKD. After constructing sentence pairs according to the complexity levels of the substitutions, we use some rules to further reduce noise.

First, we excluded substitutions which are complex words in HanLS, thus there is no overlap between the generated data and the test data. Second, we ensure that the complex word in each constructed sentence pair is sourced from the Xinhua Dictionary, guaranteeing that GPT-4 has a solid understanding of it.

Table 5 shows a constructed sentence pairs and the corresponding training sample. Some basic statistics of the training dataset for fine-tuning the small models are shown in Table 6.

B Parameter Settings

Table 7 shows the infrastructure for conducting our experiments. The hyper-parameters used for fine-tuning ChatYuan-large-v2, ChatGLM2-6B and Qwen1.5-7B-Chat are listed in Table 8. We run each model three times with seed 2022, 2023 and 2024. Besides, we set the value of the temperature of GPT-4 API to 0 to prioritize generation quality, while controlling diversity through the use of pivot words.

Settings	Value
GPU	Nvidia A6000
GPU memory	48 GB
CPU	AMD EPYC 7542
OS	Ubuntu 20.04.5 LTS
Pytorch version	1.31.1
CUDA version	11.6

Table 7: Infrastructure for conducting our experiments.

C More Experimental Results

C.1 The Effect of the Number of Training Samples for Fine-tuning Small Models

The impact of increasing the number of training samples on the performance of the fine-tuned small models is depicted in Figure 10. In general, model performance improves as the number of training samples increases, particularly up to 2000 samples. This suggests that the constructed training dataset is both high-quality and data-efficient.

Furthermore, models with larger parameter scales have a higher potential for simplification capabilities. However, when only 500 training

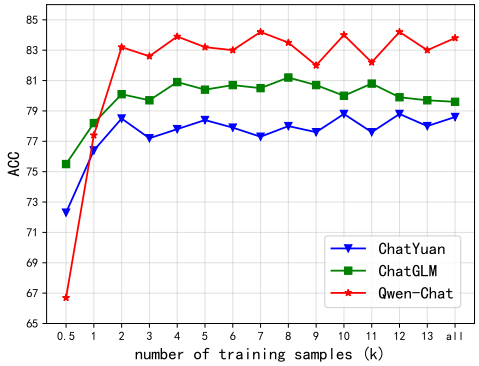

Figure 10: The effects of the number of training samples for fine-tuning small models.

Models	RIA(infer)		RIA(train & infer)	
	ACC	f-ACC	ACC	f-ACC
Chatyuan	82.0	84.2	82.6	85.1
ChatGLM	81.2	84.3	82.4	84.9
Qwen-Chat	83.4	87.1	84.2	87.4

Table 9: Results of overall performance of various models. RIA(infer): use interpretations merely for inference; RIA(train & infer): use interpretations for both training and inference.

samples are available, larger models tend to underperform compared to smaller models. This indicates that smaller models are more advantageous when the training data is very limited.

C.2 The Effect of Fine-tuning with RIA

To verify the effectiveness of incorporating interpretation for fine-tuning small models, we conduct a comparative experiment. For small models that trained without interpretation, we incorporate interpretation during inference and compare its performance with the standard RIA methods that utilize interpretation for both training and inference.

The results are shown in Table 9, which indicate that for all small models, the alignment usage of interpretation for both training and inference leads to a slight performance improvement.

Hyper-parameters	Value
ChatYuan	
max_seq_length	512
num_epoch	1
learning_rate	5e-5
scheduler	cosine
batch_size	16
ChatGLM	
max_seq_length	512
num_epoch	3
learning_rate	5e-5
scheduler	cosine
batch_size	16
lora_rank	8
lora_alpha	32
lora_dropout	0.1
Qwen	
max_seq_length	512
num_epoch	3
learning_rate	5e-5
scheduler	cosine
batch_size	4
lora_rank	8
lora_alpha	16
lora_dropout	0

Table 8: Hyper-parameter settings for fine-tuning ChatYuan, ChatGLM-6B and Qwen1.5-7B models.