ProLex: A Benchmark for Language Proficiency-oriented Lexical Substitution

Anonymous ACL submission

Abstract

Lexical Substitution discovers *appropriate* substitutes for a given target word in a context sentence. However, the task fails to consider substitutes that are of *equal or higher proficiency* than the target, an aspect that could be beneficial for language learners looking to improve their writing. To bridge this gap, we propose a new task — *language proficiency-oriented lexical substitution*. We also introduce ProLex, a novel benchmark designed to assess systems' ability to generate not only appropriate substitutes but also substitutes that demonstrate better language proficiency. Besides the benchmark, we propose models that can automatically perform the new task. We show that our best model, a *Llama2-13B* model fine-tuned with task-specific synthetic data, outperforms *ChatGPT* by an average of 3.2% in F-score and achieves comparable results with *GPT-4* on ProLex[1].

1 Introduction

Nowadays, automatic English learning tools have become widespread across various educational settings. For instance, automatic grammar error correction systems (Omelianchuk et al., 2020; Yasunaga et al., 2021; Tarnavskyi et al., 2022; Cao et al., 2023), simplify grammar correction for learners and help enhance their writing skills. Besides grammar, enhancing vocabulary diversity is another crucial element in improving English writing (Smitherman and Villanueva, 2003; Johnson et al., 2016; González, 2017). Nevertheless, English second-language (L2) learners frequently struggle to use a diverse vocabulary in their writing (Gu and Johnson, 1996; Fan, 2020; Li et al., 2023; Sun et al., 2023). They tend to use the same set of words repeatedly, which may negatively impact their performance on essay writing tests (Johnson et al., 2016).

The beginner L2 learners can leverage existing lexical substitution systems (Zhou et al., 2019; Lacerra et al., 2021; Yang et al., 2022; Wada et al., 2022; Qiang et al., 2023) to enhance their vocabulary breadth. These systems are designed to identify contextually suitable substitutes for a target word, thereby assisting learners in discovering appropriate substitutions. However, merely knowing these contextually appropriate substitutes is insufficient for L2 learners. Prior benchmarks (McCarthy and Navigli, 2007; Kremer et al., 2014; Horn et al., 2014; Lee et al., 2021) for lexical substitution focus solely on the contextual appropriateness of generated substitutes. To enhance the vocabulary diversity and writing proficiency of English learners, the substitutions should also reflect an equal or higher level of language proficiency compared to the target word.

In this work, we present ProLex, a novel lexical substitution benchmark that evaluates system performances on *language proficiency-oriented lexical substitution*, a new task that proposes substitutes that are not only contextually suitable but also demonstrate advanced-level proficiency. To construct ProLex, we begin by selecting target words and sentences according to their frequency in TOEFL-11 (Blanchard et al., 2013), a comprehensive corpus of essays written by non-native English learners. This approach ensures a data distribution that is more representative of L2 English learners. Subsequently, to effectively harness human expertise in annotating advanced proficiency-oriented lexical substitutions, we adopt the methodology recommended by Lee et al. (2021) — directly having humans judge the appropriateness of candidate substitutes instead of generating them spontaneously. To facilitate this, we employ GPT-4 to generate an initial pool of substitute candidates. Then, the human annotators assess these candidates

[1]We will make ProLex publicly available upon publication of the paper.

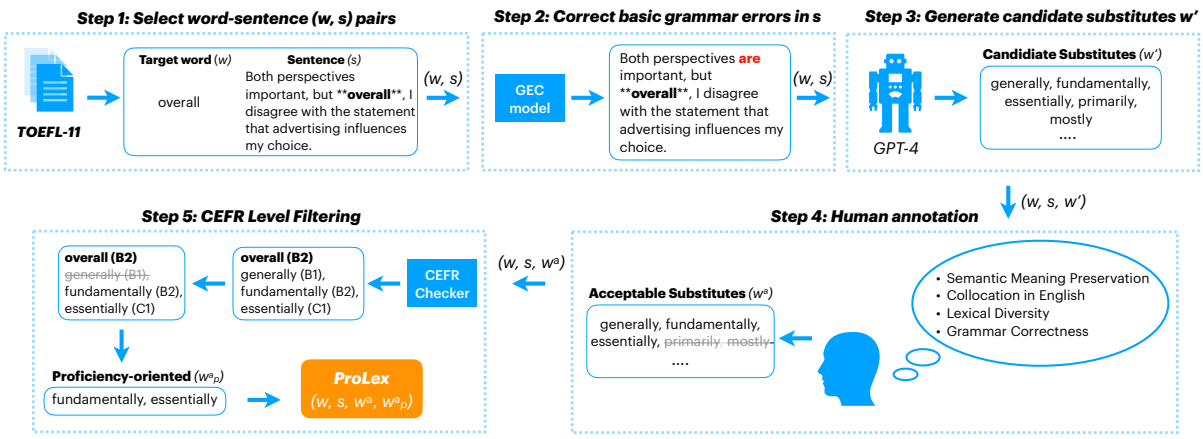


Figure 1: The process of creating ProLex. We start by selecting word-sentence (w, s) pairs from TOEFL-11 based on word frequency. Then we use a fine-tuned Grammar Error Correction (GEC) Model to correct basic grammar errors in the selected sentences. We use *GPT-4* to generate candidate substitutes, each of which is denoted as w'. For each (w, s, w') triple, we ask human expert to assess these w' based on their appropriateness. The resulting list of accetpable substitutes is denoted as w^a. For all substitutes in w^a, we further apply a CEFR Checker (CathovenAI, 2023) to obtain their proficiency levels, and ultimately remove substitutes that demonstrate lower-level proficiency than the target word. This produces our final quadruplets in ProLex, namely (w, s, w^a, w^a_p).

based on their contextual appropriateness using our language-learning-oriented annotation scheme. The proposed scheme takes into account the preservation of semantic meaning, common English collocations, lexical diversity, and grammatical correctness. We specifically recruited two annotators who are pursuing their PhDs in Linguistics, specializing in English language teaching and education. Following human annotation, we use the Common European Framework of Reference (CEFR) (of Europe Education Committee Modern Languages Division Council for Cultural Co-operation, 2001) to remove substitutions that have lower proficiency levels compared to the original target word.

In addition, we build models to perform *language proficiency-oriented lexical substitution* automatically on ProLex. We also evaluate the performance of existing state-of-the-art (SOTA) lexical substitution systems, as well as prompting LLMs in zero-shot and in-context learning settings. The results of our experiments on ProLex reveal that our top-performing model, *Llama2-13B*, which was instruction-tuned using a mix of synthetic and modified data, surpasses all existing SOTA lexical substitution systems by an average of 26.8% in F-score. It also outperforms *ChatGPT* by an average of 3.2% and attains results on par with *GPT-4*. We also demonstrate that, instruction tuning using task-specific synthetic data yields better results compared to zero-shot LLMs on ProLex.

Overall, our contributions are threefold:

- We propose a new task, namely *language proficiency-oriented lexical substitution*, to help L2 English learners enhance their vocabulary diversity and writing proficiency.
- We present a novel benchmark ProLex to assess systems' ability to perform the task by generating substitutes that are not only contextually appropriate but also reflect advanced language proficiency levels.
- We fine-tuned models with task-specific synthetic data and evaluated them with ProLex. The models attain results that are comparable to, or better than, the out-of-the-box LLMs, such as *GPT-4* and *ChatGPT*.

2 Related Work

We first outline the task of lexical substitution, then introduce recent high-quality benchmark in this field, Swords, as detailed in (Lee et al., 2021). Next, we describe TOEFL-11 (Blanchard et al., 2013), a large-scale corpus of written essays composed by non-native English learners, which we adopt as the corpus for selecting context sentences and target words for ProLex.

Lexical Substitution Lexical substitution is a well-established task, which was originally defined in McCarthy (2002) - given a target word w in a given context c, one needs to generate a list of substitutes w' that can be used to replace w in c.

Specifically, the context c refers to one or more sentences encompassing the target word w. The target word itself, which may be selected manually by human annotators (McCarthy and Navigli, 2007), or automatically based on its part-of-speech as discussed in Kremer et al. (2014); Lee et al. (2021), is a single word within this context. The substitute w' for the target word can be either a single word or a phrase. Although w' can be ungrammatical in previous benchmarks, we want to incorporate only grammatically correct substitutes in ProLex. Moreover, the past benchmarks do not consider the language proficiency of the substitutes w'. In ProLex, our approach includes substitutes that are not only acceptable but also demonstrate equal or higher proficiency level compared to the target word.

Swords Different from other popular benchmarks (McCarthy and Navigli, 2007; Kremer et al., 2014) depending on recall from humans as the only source of data, Swords (Lee et al., 2021) enhanced the collection of data in lexical substitution by treating it as a classification problem. This method was guided by the intuition that it is easier for humans to judge the appropriateness of given substitutes than to create them spontaneously. This approach resulted in a higher-coverage and higher-quality benchmark. In our human annotation process, we follow Swords's established instructions, and propose a language-learning-oriented annotation scheme for the appropriateness judgement.

TOEFL-11 Essay Corpus The large-scale non-native English writing corpus, TOEFL-11 (Blanchard et al., 2013), contains writing samples from non-native TOEFL (the Test of English as a Foreign Language) test takers from 2006 to 2007. In particular, it contains $12,100$ essays, each of which is categorized into one of the three score levels: high, medium, and low. Focusing on enhancing the writing skills of English learning beginners, our methodology involves selecting context sentences and target words primarily from essays with *medium* and *low* scores. This approach is based on the assumption that beginner L2 English learners are likely to possess a more limited vocabulary range (Fan, 2000).

3 ProLex Benchmark

We propose ProLex, a benchmark for language proficiency-oriented lexical substitution. ProLex is composed of quadruplets, each containing a target word, a context sentence, a list of acceptable substitutes, and a list of proficiency-oriented substitutes (w, s, w^a, w^a_p). As indicated by Lee et al. (2021), previous benchmarks (McCarthy and Navigli, 2007; Kremer et al., 2014) gathered substitute words by prompting humans to generate them spontaneously. However, many viable substitutes, challenging for humans to devise, might ultimately be overlooked. Therefore, we follow Lee et al. (2021) by first obtaining a set of candidate substitutes and later asking annotators to judge their appropriateness. Formally, given a context s, target word w, and candidate substitute w', the annotators are asked to judge whether w' is an acceptable substitute for the target word: (w, s, w') $\rightarrow$ {1, -1, 0}, with 1, -1 and 0 indicating "accept", "reject", and "uncertain" respectively. Figure 1 shows the complete benchmark creation process for ProLex. In the following sections, we will first illustrate the data creation process in Section 3.1. Then, in Section 3.2, we will describe our language-learning-oriented annotation scheme. In Section 3.3, we show how we filter substitutes based on CEFR proficiency levels.

3.1 Data Creation

Since we focus on improving the writing proficiency of beginner L2 learners, we turn to TOEFL-11 (Blanchard et al., 2013). We will select contexts, target words, correct basic grammar errors, and generate candidate substitutes through the following three steps.

Step 1: Select contexts, and target words. The TOEFL-11 dataset comprises of $12,100$ essays, each scored as high, medium, or low. Specifically, there are $4,202$ high-scored essays, $6,568$ with medium scores, and $1,330$ categorized as low. Given the limited vocabulary range of low and medium-proficiency L2 English learners, we intend to select the words that are frequently used in their essays as the target words in our benchmark. For each essay of interest, we select mainly the content words (i.e. nouns, verbs, adjectives, and adverbs) that are used *at least three times* in the essay. Then, for each such target word, we randomly choose one sentence from all sentences containing that target word. The general statistics of all target words and context sentences are presented in Appendix C Table 5. In sum, we selected $2,531$ unique word-context (w, s) pairs from low

and medium-level essays from TOEFL-11 corpus.

Step 2: Correct basic grammar errors in selected context sentences. We observed that the chosen sentences frequently contain fundamental grammar mistakes, such as spelling errors, which can obscure the intended meaning of the context. Example grammar errors are demonstrated in appendix D. To correct these errors, we follow Rothe et al. (2021) and fine-tune a GPT-2 model (Radford et al., 2019) on CLang-8 dataset[2]. The experiment details are also shown in appendix D. The resulting model is deployed to correct the basic grammar errors in all of our $2,531$ selected context sentences s. We then sample $1,125$ (w, s) pairs for human annotations[3].

Step 3: Generate candidate substitutes. Powered by recent advances in LLMs (Brown et al., 2020; Wei et al., 2022), for each (w, s) we generate the candidates using GPT-4. Specifically, we prompt GPT-4 with five in-context examples to generate exactly five candidate substitutions for each of our $1,000$ (w, s) pairs — thus resulting in $5,000$ (w, s, w') triples, with w' representing each candidate substitute for each (w, s) pair. Figure 1 demonstrates an example of the generated candidates. The complete prompt for GPT-4 generation is shown in Appendix B.

3.2 ProLex Annotation Scheme

Different from previous schemes (McCarthy and Navigli, 2007; Kremer et al., 2014; Lee et al., 2021), ProLex deems a substitute valid if it maintains the original sentence's meaning, aligns with common English usage, and is grammatically correct. Furthermore, ProLex encourages a diverse set of substitutes as it can expand the vocabulary of language learners. A detailed description of the annotation scheme is in Appendix A. Formally, as described in the previous section, for each (w, s, w') pair, we ask the annotators to judge whether w' is an acceptable substitute for w, by assigning one of the following classes to w', 1 (acceptable), -1 (unacceptable) and 0 (uncertain). Subsequently, for each (w, s) pair, we will combine all w' that are labeled as acceptable to form w^a, leading to (w, s, w^a) triples. In the following, we will discuss the features of the annotation scheme in detail.

[2]https://github.com/ google-research-datasets/clang8

[3]We used 125 pairs for calculating inter-annotator agreement and $1,000$ for the annotation. We will include more pairs to expand the benchmark in future work.

3.2.1 Semantic Meaning Preservation

Given a triple (w, s, w'), namely a target word w, and the context sentence s where w is situated, the word substitute w' should preserve the semantic meaning of the original context sentence s. Concretely, there are three cases of acceptable substitutes (i.e. labeled as "1") for meaning preservation in ProLex. Firstly, the annotators accept a substitute of w if they convey exactly the same meaning in s. Besides, a substitute is also acceptable if it is an entailment from the sense of the target word. For example, *trust a person* entails *rely on a person* in the following sentence: *"I **trust** a person who has more knowledge than I do."* Hence, *rely on* is a valid substitute to target word *trust* in the example. Lastly, a substitute can be used figuratively and convey a similar meaning as the target word w. In the following sentence: *"Once the undergraduate studies are **pursued** by a student, the student is more aware of different subjects and the knowledge he has gained in the period of his studies."*, *embarked on* is an acceptable substitute as it is a figurative use case, preserving the meaning of the target word *pursued*.

3.2.2 Common Collocations in English

Learners can improve their general levels of writing proficiency in the language through collocation knowledge (Howarth, 1998; Gitsaki, 1999; Boers and Webb, 2018). In ProLex, annotators accept substitutes only if they are common English collocations within the context sentence. For instance, consider the following example where the target word is *defying*: *"Successful people are not only seeking new experiences, **defying** obstacles and hardship but also trying to be creative."* Suppose we have a candidate substitute *conquering*, even though *conquering obstacles* conveys a similar meaning as *defying obstacles*, it will be rejected since it is not a common expression and collocation in English.

In addition, in cases where the annotators are unsure about the collocation of any expression, we encourage them to refer to an external collocation knowledge base, such as COCA, the Corpus of Contemporary American English (Davies, 2010), which contains 400 million words from various genres (e.g. spoken, fiction, newspapers and etc.) distributed from 1990 to present time. In particular, we ask the annotator to accept certain expressions if

the frequency[4] of it queried from COCA is greater than 5. Take *conquer obstacles* from the previous example, since there is only one instance of *conquer obstacles* in COCA, it is rejected from the candidate list.

3.2.3 Lexical Diversity

A richer and more diverse vocabulary can improve both the quality and effectiveness of communication in English (Yu, 2010). To enhance lexical diversity in ProLex, annotators are asked to mark a substitute as acceptable if, in general, it matches at least one connotation of the target word in the context sentence. For instance, in the sentence *"In fact they not only give the possibility to move in a short time but sometimes they are also **precious** goods."*, *precious* conveys a connotation of high value in price or importance. Hence, the following candidates are all acceptable: *valuable, prized, cherished, invaluable* and *treasured*, as they express either "high value in price" or "importance", matching at least one connotation of the target word *precious*.

3.2.4 Grammar Correctness

Although previous benchmarks (McCarthy and Navigli, 2007; Kremer et al., 2014; Lee et al., 2021) do not involve grammar correction, ProLex considers the substitutes to be acceptable only if they are grammatically correct. For instance, in the sentence *"Nevertheless, who are mostly responsible for these **research**?"*, all of the following candidates are rejected because none of them are in plural forms: *study, investigation, analysis, examination* and *inquiry*.

3.3 Substitutes Filtering based on CEFR Proficiency Levels

Taking the triples (w, s, w^a) from human annotations in section 3.2, for each target word w in a context sentence s, and the list of acceptable substitutes w^a, we perform filtering to select substitutes with equal or higher proficiency compared to the target word to form w_p^a — resulting in the final quadruplets of ProLex (w, s, w^a, w_p^a). Concretely, we refer to the CEFR Checker (CathovenAI, 2023) to automatically determine the CEFR level of each target word w and its associated acceptable substitutes w^a. An example filtering process is shown in Figure 1. Note that *generally* is removed from the set since its CEFR level (i.e. B1) is lower than that of *overall* (i.e. B2).

[4]https://www.english-corpora.org//coca/

Essay Proficiency	# (w, s)	Avg s length	# w^a	# w_p^a
low	169	18.8	489	419
medium	511	22.7	1466	1140
all	680	21.8	1955	1559

Table 1: Statistics of ProLex, separated according to essay proficiency. "all" denotes the combinations of both "low" and "medium"-level essays.

4 Dataset annotation process and statistics

Considering the necessity for substantial semantic understanding and knowledge of English collocations in the annotation process, we recruited two annotators currently pursuing PhDs in Linguistics, with specializations in English language teaching and education. On the 125 (w, s) pairs, namely 625 (w, s, w') triples, the two annotators reached an inter-annotator agreement of $\kappa = 0.6$, indicating substantial agreement[5]. Then, they annotated the complete test set separately, with each annotating $2,500$ (w, s, w') triples sampled from section 3.1. In total, among all $5,000$ candidate substitutes w', 39% is labeled as 1 (accept), while 51% and 10% are annotated as -1 (reject) and 0 (uncertain), respectively. This entails that GPT-4 generated candidates may not always be valid in ProLex, indicating future research opportunities in using LLMs for generating semantically correct, collocationally appropriate English lexical substitutions.

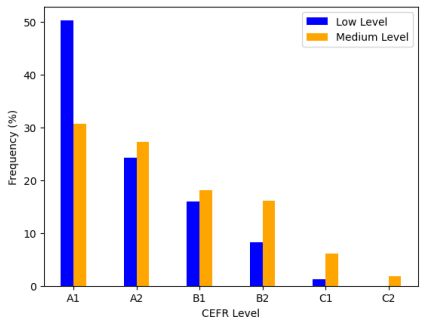
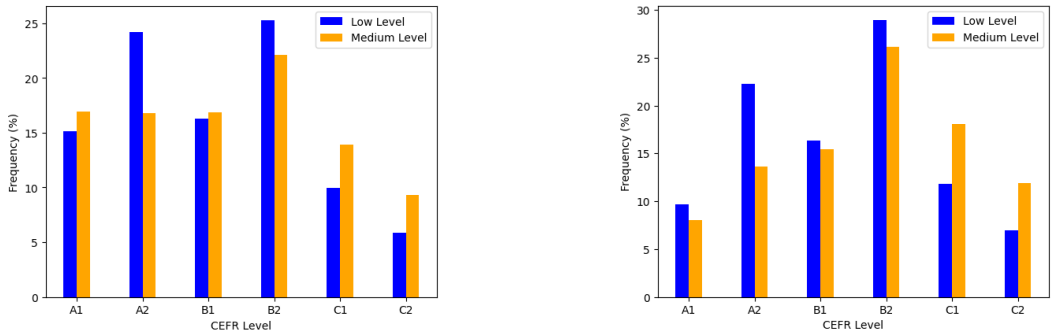


Figure 2: Distribution of CEFR levels of target words in low and medium sentences in ProLex.

We further process the annotations to compose the final quadruplets (w, s, w^a, w_p^a) in ProLex,

[5]We perform Cohen's Kappa (Cohen, 1960) to calculate the inter-annotator agreement. We exclude the labels with 0 to consider only the contexts that are conceivable to both annotators (i.e. labeled as 1 or -1).



(a) Acceptable substitutes w^a.

(b) Proficiency-oriented substitutes w_p^a.

Figure 3: Distribution of CEFR levels for substitutes in w^a and substitutes in w_p^a in ProLex. In low-level sentences, more than 65% of the proficiency-oriented substitutes are from B1 level or higher; similarly, in medium-level sentences, over 75% of these substitutes are sourced from B1 level or above.

where w^a denotes a list of acceptable substitutes for w, and w_p^a denotes the list of acceptable substitutes after filtering based on CEFR levels. We exclude the target words that have zero acceptable substitutes. Table 1 provides an overview of the statistics for the resulting data, grouped by low and medium-level proficiency. In total, there are 680 word-sentence (w, s) pairs in ProLex with at least one acceptable substitute [6]. On average, for each (w, s) pair, there are 2.9 acceptable substitutes w^a and 2.3 proficiency-oriented substitutes w_p^a. Furthermore, we provide the distribution of CEFR levels of the target words in both low and medium sentences, and the result is shown in Figure 2. As evident, a significant proportion of the chosen target words are from A1 and A2 levels, comprising 75% in low-level sentences and 55% in medium-level sentences, respectively. Moreover, in Figure 3, we also present the CEFR level distributions for acceptable substitutes w^a and proficiency-oriented substitutes w_p^a. The distribution resonates closely with our research objective of ProLex, which aims to propose acceptable substitutions of high-proficiency level words for commonly used low-proficiency level words in L2 English sentences.

5 Language Proficiency-oriented Lexical Substitution

In this section, we outline recommended evaluation practices for ProLex, propose models that can automatically perform the task on ProLex, and assess the performance of out-of-the-box LLMs and top lexical substitution systems.

[6]The Part-Of-Speech distribution for the target word w in ProLex is shown in Table 6.

5.1 Evaluation Settings

For the task of lexical substitution, there are two mainstream evaluation settings: generative setting (McCarthy and Navigli, 2007) and ranking setting (Thater et al., 2010). In the generative context, systems produce a sequence of potential substitutes, and there are no limits on the number of candidate substitutes they can generate. In the ranking scenario, systems receive the entire set of substitute options provided by the benchmark. The task then is to order these substitutes according to their appropriateness. In this work, we mainly focus on the generative setting and defer the ranking setting to future research.

5.2 Evaluation Metrics

We evaluate the system performance in terms of its substitute *appropriateness* and *proficiency*. For assessing *appropriateness*, we compare the system's predictions with the acceptable substitutes, denoted as w^a; similarly, for evaluating *proficiency*, we contrast the system's predictions with proficiency-oriented substitutes, represented as w_p^a. Inspired by Lee et al. (2021), we consider the evaluation metrics that measure the quality and coverage of the predicted substitutes from a system. Specifically, for *appropriateness*, we compute precision (P^k), recall (R^k) and F-score (F^k)[7] at k:

$$P^k = \frac{\#\text{ acceptable subs } w^a \text{ in system top-}k}{\#\text{ substitutes in system top-}k}$$

$$R^k = \frac{\#\text{ acceptable subs } w^a \text{ in system top-}k}{\min(k, \#\text{ acceptable subs } w^a)}$$

[7]F^k is calculated as the harmonic mean of P^k and R^k.

Target word (w)	promotion (B2)
Sentence (s)	This **promotion** has a beautiful and effective visual part, but they miss the real point: the product.
Acceptable (w^a)	advertising (A2), marketing (B1), publicity (B2), campaign (B1), advertisement (A2)
Proficiency-oriented (w_p^a)	publicity (B2)
GPT-4 (32-shot)	**advertising**, **marketing**, **publicity**, hype, announcement
ChatGPT (32-shot)	**advertising**, **marketing**, **publicity**, **campaign**, endorsement
Vicuna-1.5-13B-D_L	advancement, **publicity**, **marketing**, endorsement
Llama2-13B-D_{LS}	advancement, progression, elevation
Para-LS	**advertising**, **advertisement**, **marketing**, **campaign**, **publicity**, ad, presentation, show, event, production

Table 2: Example data point and predictions from the top systems. In the outputs, the acceptable substitutes generated from each systems are bolded, while proficiency-oriented ones are both bolded and blue.

Similarly, we evaluate *proficiency* against the list of proficiency-oriented substitutes w_p^a. P_p^k and R_p^k represent the precision and recall of system outputs against this smaller candidate list, with F_p^k as their harmonic mean. Also, we follow previous work (McCarthy and Navigli, 2007; Lee et al., 2021) to mainly examine performance for $k = 10$, and implement a *soft* evaluation setting, where all substitutes generated by a system are lemmatized before comparison. In addition, given that ProLex considers grammatically correct substitutes, we also apply a *hard* evaluation setting, where predictions should *exactly match* the reference without lemmatization.

5.3 Baselines

We evaluate systems in the following settings: 1) existing top lexical substitution systems, 2) instruction-tuning language models on task-specific synthetic data, and 3) zero-shot prompting and in-context learning with LLMs.

Existing Top Lexical Substitution Systems Past lexical substitution systems were proposed and evaluated on three widely-known benchmarks: LS07 (McCarthy and Navigli, 2007), CoInCo (Kremer et al., 2014) and Swords (Lee et al., 2021). We evaluate the following two representative systems on ProLex:

- *BERT-LS* (Zhou et al., 2019): a BERT-based lexical substitution system. It once reached SOTA results on LS07 and CoInCo dataset.
- *Para-LS* (Qiang et al., 2023): a paraphraser-based system, achieving SOTA performance on all three benchmarks: LS07, CoInCo, and Swords.

Instruction Tuning with Task-specific Synthetic Data As supervised instruction tuning has proliferated in building powerful instruction following models (Ouyang et al., 2022; Bai et al., 2022; OpenAI, 2023; Chiang et al., 2023), we experiment with two instruction-tuned large-scale language models, namely *Vicuna*[8] (Chiang et al., 2023) and *Llama-2* (Touvron et al., 2023). Note that we mainly focus on the smaller variants: *Vicuna* 7B/13B and *Llama-2* 7B/13B. Since there is no existing training data for our language proficiency-oriented lexical substitution task, we adopt two approaches to synthesize the data: 1) directly generate data with GPT-4 through prompting, and 2) modify the existing lexical substitution benchmark, Swords. The prompts we employed to synthesize the data with GPT-4 are shown in section E.2.1. As for the second approach, we considered only the acceptable substitutes (i.e. score greater than 50%) from the Swords dataset and filtered the substitutes based on their CEFR levels, by only keeping the ones that have equal or higher proficiency levels compared to the target word. As a result, we denote the training data synthesized with GPT-4 as D_L, the data modified from Swords as D_S, and the aggregate of these two datasets as D_{LS}.

Zero-shot and In-context Learning with LLMs Leveraging the potent zero-shot capabilities of Large Language Models (LLMs), we evaluate their performance on ProLex by prompting two widely recognized LLMs, *GPT-4* and *ChatGPT* [9], to generate potential substitution candidates. Moreover, we also perform in-context learning with examples selected from D_L[10]. See appendix E.1 for more details.

5.4 Results and Analysis

Table 3 demonstrates the results of our experiments on ProLex. We take the top 10 ($k = 10$) predictions from each system to evaluate. In general, the *appropriateness* scores F^{10} for both *soft* and *hard* settings are greater than the *proficiency* scores F_p^{10},

[8] We used Vicuna v1.5 provided by (Zheng et al., 2023).

[9] Note that we used the version `gpt-3.5-turbo-1106`

[10] We randomly select 32 examples from D_L for in-context learning with LLMs

Models		BERT-LS	Para-LS	Llama-2				Vicuna-1.5				ChatGPT		GPT-4	
				7B		13B		7B		13B					
				D_S	D_L	D_S	D_L	D_S	D_L	D_S	D_L	*zero*	32	*zero*	32
Soft	F^{10}	25.2	30.2	44.1	48.6	44.4	51.1	43.5	48.0	46.2	52.1	50.9	51.5	50.5	**54.7**
	F_p^{10}	20.3	25.3	39.1	43.2	38.6	46.2	38.6	43.1	40.7	46.8	45.2	45.9	44.6	**48.8**
Hard	F^{10}	20.0	20.2	31.2	46.5	27.1	49.0	31.4	45.6	32.2	49.6	47.8	48.3	48.8	**52.7**
	F_p^{10}	15.8	17.2	27.8	41.2	23.9	44.1	27.7	40.9	28.6	44.5	42.5	43.0	43.3	**46.8**

Table 3: Evaluation of systems on ProLex under *soft* and *hard* settings. We fine-tuned *Vicuna* and *Llama-2* model variants on both D_L, the synthesized data generated by LLM, and D_S, the filtered dataset based on Swords (Lee et al., 2021). We also conducted zero-shot (i.e. $zero$) and in-context learning (i.e. 32 shots) with LLMs.

Models		Vicuna-1.5		Llama-2	
		7B	13B	7B	13B
Soft	F^{10}	49.9	52.4	49.7	**54.2**
	F_p^{10}	45.0	47.0	44.6	**48.4**
Hard	F^{10}	48.1	50.5	48.0	**51.6**
	F_p^{10}	43.0	45.2	42.8	**45.9**

Table 4: Evaluation results on ProLex after fine-tuning *Vicuna-1.5* and *Llama-2* on the aggregate dataset D_{LS}.

indicating the challenges of generating proficiency-oriented substitutes. Additionally, all systems tend to achieve higher scores in *soft* setting, given that all predictions are lemmatized before comparison. This indicates that these systems might produce substitutes that are grammatically incorrect yet suitable and demonstrate advanced language proficiency.

Overall, in-context learning with *GPT-4* achieved the best results across all systems in all evaluation settings. Note that the 32 in-context examples were randomly selected from our task-specific synthetic dataset D_L. This underscores the effectiveness of our data synthesis method, which can be applied in situations with limited data resources. Using D_L, *Vicuna-1.5-13B* achieved the best performance among all fine-tuned model variants on single dataset and previous top-performing lexical substitution systems, even surpassing *ChatGPT* on both zero-shot and in-context learning scenarios, and *GPT-4* on zero-shot setting.

In addition, in our instruction-tuning setup, training with dataset D_S resulted in lower performance compared to training with the synthetic dataset D_L. This occurred because D_S is derived from filtering Swords, which comprises context sentences not authored by L2 English learners, thus leading to a distribution shift. On the other hand, all context sentences in dataset D_L originate from the TOEFL-11 corpus, the same source used to develop ProLex. We also conducted fine-tuning experiments using the aggregate dataset D_{LS}. Table 4 demonstrates that combining D_S and D_L enhanced the performance of all fine-tuned models. Notably, *Llama-2-13B* achieved scores nearly comparable to the best result of *GPT-4*.

As for previous lexical substitution systems, *Para-LS* consistently surpasses *BERT-LS* but lags far behind all other larger-scale systems. This highlights the limitations of these earlier methods and points towards a promising research direction in developing more efficient systems that can perform better in generating language proficiency-oriented substitutions.

For error analysis, Table 2 shows an example and the outputs generated by top systems. *GPT-4* and *ChatGPT* produce words that have different semantics from the target (e.g. "hype" for **promotion**). Our fine-tuned models produce substitutes covering multiple connotations of the target (e.g. "advancement" and "marketing"), leading to lower precision. *Para-LS* generates words that have a more general meaning than the target (e.g. "event" and "show").

6 Conclusion and Future Work

We propose a new task *language proficiency-oriented lexical substitution* to improve the vocabulary diversity and writing proficiency of L2 English learners. We introduce ProLex, a novel benchmark designed to assess systems' ability to generate *appropriate* and *language proficiency-oriented* substitutes for given target words in the context. Besides, we also fine-tuned open-source language models on task-specific synthetic data, achieving results that are on par with, or better than GPT-4 and ChatGPT. In future, we will expand ProLex by leveraging larger L2 English writing corpora and incorporating more comprehensive sets of substitutes.

7 Limitations

Benchmark Size and Coverage Compared to previous Lexical Substitution benchmarks (McCarthy and Navigli, 2007; Kremer et al., 2014; Lee et al., 2021), the size of ProLex is relatively small. However, considering that ProLex draws its data from the TOEFL-11 corpus, it could be significantly extended by incorporating sentences from additional sources authored by L2 speakers.

In addition, the systems may produce valid substitutes that are not present in ProLex during evaluation, indicating that ProLex has limitations in coverage. By employing the annotation scheme described in Section 3.2, we can iteratively update and enhance ProLex with additional gold standard substitutes, thereby expanding its coverage in future versions.

Limits of CEFR Checker The CEFR Checker (CathovenAI, 2023) we used in this work is capable of assigning CEFR levels only at the word level. However, we noticed that in some cases, phrases could also serve as appropriate substitutes. Hence, extending the CEFR Checker's functionality to include CEFR-level assignments at both the word and phrase levels would offer a more holistic evaluation in future versions of ProLex.

8 Ethics Statement

Reproducibility In this work, our data creation process utilized GPT-4. We also used ChatGPT for evaluation purposes. Although they are not open-sourced language models, to facilitate the reproducibility of our results, we demonstrated all of prompts used in our paper. In addition, all the other models used in this research, are publicly available in peer-reviewed articles and referenced in this paper. All datasets, including our synthetic fine-tuning dataset and all annotated test data, will be released.

Biases Our models are built over *Vicuna* and *Llama-2*. In this work, we did not explicitly handle any bias that exists in the two pre-trained language models.

Human Annotators Both annotators were recruited from the doctoral programs in the linguistics department, and they specialize in English language teaching and education. They were paid at a rate of $13 per hour. To ensure the privacy and anonymity of all contributors, no personal or demographic information was collected.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Frank Boers and Stuart Webb. 2018. Teaching and learning collocation in adult second and foreign language learning. *Language Teaching*, 51(1):77–89.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hannan Cao, Wenmian Yang, and Hwee Tou Ng. 2023. Mitigating exposure bias in grammatical error correction with data augmentation and reweighting. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2115–2127.

CathovenAI. 2023. Cefr checker (version 1.1.0) [web app]. ado language hub.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Mark Davies. 2010. The corpus of contemporary american english as the first reliable monitor corpus of english. *Literary and linguistic computing*, 25(4):447–464.

May Fan. 2000. How big is the gap and how to narrow it? an investigation into the active and passive vocabulary knowledge of l2 learners. *Relc journal*, 31(2):105–119.

Na Fan. 2020. Strategy use in second language vocabulary learning and its relationships with the breadth and depth of vocabulary knowledge: A structural equation modeling study. *Frontiers in psychology*, 11:752.

Christina Gitsaki. 1999. Second language lexical acquisition: A study of the development of collocational knowledge.

Melanie C González. 2017. The contribution of lexical diversity to college-level writing. *TESOL Journal*, 8(4):899–919.

Yongqi Gu and Robert Keith Johnson. 1996. Vocabulary learning strategies and language learning outcomes. *Language learning*, 46(4):643–679.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. Learning a lexical simplifier using wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 458–463.

Peter Howarth. 1998. Phraseology and second language proficiency. *Applied linguistics*, 19(1):24–44.

Mark D Johnson, Anthony Acevedo, and Leonardo Mercado. 2016. Vocabulary knowledge and vocabulary use in second language writing. *TESOL Journal*, 7(3):700–715.

Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. 2014. What substitutes tell us-analysis of an "all-words" lexical substitution corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 540–549.

Caterina Lacerra, Rocco Tripodi, and Roberto Navigli. 2021. Genesis: a generative approach to substitutes in context. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10810–10823.

Mina Lee, Chris Donahue, Robin Jia, Alexander Iyabor, and Percy Liang. 2021. Swords: A benchmark for lexical substitution with improved data coverage and quality. *arXiv preprint arXiv:2106.04102*.

Mo Li, Xiaotian Zhang, and Barry Lee Reynolds. 2023. Exploring lexical bundles in low proficiency level l2 learners' english writing: an ets corpus study. *Applied Linguistics Review*, 14(4):847–873.

Diana McCarthy. 2002. Lexical substitution as a task for wsd evaluation. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions*, pages 89–115.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 48–53.

Council of Europe Education Committee Modern Languages Division Council for Cultural Co-operation. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. Gector–grammatical error correction: tag, not rewrite. *arXiv preprint arXiv:2005.12592*.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Jipeng Qiang, Kang Liu, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. Parals: Lexical substitution via pretrained paraphraser. *arXiv preprint arXiv:2305.08146*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction. *arXiv preprint arXiv:2106.03830*.

Geneva Smitherman and Victor Villanueva. 2003. *Language diversity in the classroom: From intention to practice*. SIU Press.

Danning Sun, Zihan Chen, and Shanhua Zhu. 2023. What affects second language vocabulary learning? evidence from multivariate analysis. In *Frontiers in Education*, volume 8, page 1210640. Frontiers.

Maksym Tarnavskyi, Artem Chernodub, and Kostiantyn Omelianchuk. 2022. Ensembling and knowledge distilling of large sequence taggers for grammatical error correction. *arXiv preprint arXiv:2203.13064*.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2010. Contextualizing semantic representations using syntactically enriched vector models. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 948–957.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Takashi Wada, Timothy Baldwin, Yuji Matsumoto, and Jey Han Lau. 2022. Unsupervised lexical substitution with decontextualised embeddings. *arXiv preprint arXiv:2209.08236*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Xi Yang, Jie Zhang, Kejiang Chen, Weiming Zhang, Zehua Ma, Feng Wang, and Nenghai Yu. 2022. Tracing text provenance via context-aware lexical substitution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11613–11621.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2021. Lm-critic: Language models for unsupervised grammatical error correction. *arXiv preprint arXiv:2109.06822*.

Guoxing Yu. 2010. Lexical diversity in writing and speaking task performances. *Applied linguistics*, 31(2):236–259.

Xun Yuan, Derek Pham, Sam Davidson, and Zhou Yu. 2022. ErAConD: Error annotated conversational dialog dataset for grammatical error correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 76–84, Seattle, United States. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. Bert-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373.

A Detailed ProLex Annotation Scheme

In this section, we present the detailed annotation scheme for ProLex. In the annotation process, each annotator is presented with $2,500$ (w, s, w') triplets. For each (w, s, w'), we ask the annotators to judge whether w' is an acceptable substitute for w in s. In particular, they will choose among three labels: 1 (acceptable), -1 (unacceptable), and 0 (uncertain). The general guidelines are shown in Table 9. A candidate substitute w' is labeled as 1 only if it satisfies all criteria in the table. It is labeled as -1 once it violates any one of them. We also provide label 0 to exclude the cases when either the target word w or context sentence s is extremely hard to understand as they are written by beginner L2 English learners. Besides the general guidelines, we also provide a detailed specification along with some examples for all three labels. See Table 10 for more illustration examples. The detailed specification is as follows:

- To determine if a substitute is a common collocation or expression in actual English use, annotators should refer to COCA to search for certain expressions:
  - Mark 1 if the frequency of the expression is greater than 5
  - Mark -1 otherwise
- Mark 1 if the substitute is an **entailment** from the sense of the target word.
- Mark 1 if the substitute can be used **figuratively** and also convey the same semantics as the target word does.
- Mark 0 if the substitute is a hyponym of the target word.
- A sentence can contain some grammar errors (e.g. spelling errors). If it is written in a way that it is hard for you to understand, mark it 0; otherwise, if you can understand, mark the substitute as 1 or -1.
- Mark 1 if the substitute is good even though with incorrect determiner like "a" vs "an".
- Mark -1 if the substitute is good but not quite grammatically correct.
- Grammar of a substitute matters, please keep a high bar for making a substitute as 1.
- Mark 1 if the target word itself appears as a substitute.
- Mark -1 if the target word is a proper noun or part of a fixed expression/phrase.
- Mark 0 if you need more contexts of the original sentence to help you judge the acceptability.

B Generate Candidate Substitutions from GPT-4

We generate the candidate substitutes for human annotation by prompting GPT-4. The prompt we used during generation is demonstrated in Table 11. In particular, we selected five instances from Swords (Lee et al., 2021) as the in-context examples to guide the generation process. In these examples, the target words consist exclusively of content words — nouns, verbs, adjectives, and adverbs — which represent the primary focus of ProLex.

C Dataset Statistics

In this section, we illustrate the statistics of all target words and context sentences extracted from Section 3.1. The overall statistics are shown in Table 5 below. In general, there are $2,531$ unique (w, s) pairs combing sentences from low and medium-level essays. Note that on average, sentences from medium-level essays are longer than the ones from low-level essays. Furthermore, nouns comprise the majority of the target words, accounting for 57% of all selected words. In contrast, verbs, adjectives, and adverbs make up 23%, 16%, and 4% of the selection, respectively.

Note that we sample $1,000$ (w, s) pairs from the global dataset, generate candidate substitutes from GPT-4, and ask human annotators to judge the contextual appropriateness of these candidates. After filtering the ones labeled with 0, we end up with 680 (w, s) pairs in the final ProLex dataset. We also present the statistics of these (w, s) pairs in ProLex in Table 6. Among all 680 pairs, 52% of them are nouns, 25% are verbs, 18% are adjectives, and 5% are adverbs, closely mirroring the distribution found in the overall dataset described above.

D Grammar Error Correction on TOEFL-11

Because the sentences in ProLex are extracted from low and medium-level essays from TOEFL-11 cor-

Essay Proficiency	# of (w, s)	# of unique s	Avg s length	# of nouns	# of verbs	# of adj	# of adv
low	630	609	19	345	169	94	22
medium	1901	1824	64	1090	415	320	76
all	2531	2433	21	1435	584	414	98

Table 5: Statistics of all target words w and context sentences s. (w, s) indicates a pair of target words and contexts. "all" means the combinations of both "low" and "medium" essays.

Essay Proficiency	# of (w, s)	# of nouns	# of verbs	# of adj	# of adv
low	169	83	50	27	9
medium	511	268	122	95	26
all	680	351	172	122	35

Table 6: Statistics of target words w in ProLex in terms of the part-of-speech tags.

pus, these sentences frequently contain basic grammar errors that obscure the intended meaning of the context. To alleviate this problem, we first follow Rothe et al. (2021) and tested a T5 base model trained on cLang-8 on the CoNLL-14 test set. Finding its performance unsatisfactory, we further fine-tuned the model on ErAConD (Yuan et al., 2022) which slightly improved it. At last, we chose to fine-tune a GPT-2 model (Radford et al., 2019) on the proposed cLang-8 dataset, which provided a satisfactory $F_{0.5}$-score as shown in table 7, allowing us to correct most grammatical errors in the TOEFL-11 essays.

In Table 8, we provide some examples to show the original sentences from TOEFL-11 and the corresponding corrected sentences output from our grammar model.

E Details of Baseline Experiments

The details of the baseline experiments are presented in this section. We measure the performance of the following baseline systems: 1) zero-shot LLMs prompting, 2) instruction tuning with task-specific synthesized data, and 3) current state-of-the-art lexical substitution systems. In the following, the setups for each of these baselines are illustrated in great detail.

E.1 Zero-shot and In-context Learning with GPT-4 and ChatGPT

We evaluate the zero-shot performance of GPT-4 and ChatGPT on ProLex. The prompt we used in the experiment is shown in Table 12. Specifically, we used the model provided by OpenAI platform, namely `gpt-4` and `gpt-3.5-turbo-1106`, to evaluate the performance of GPT-4 and ChatGPT, respectively. For each (w, s) pair, GPT-4 and ChatGPT outputs a list of substitutes that are separated by commas. In our evaluation, we take the top-k ($k = 10$) substitutes from the outputs and compare them with our labels in ProLex. Similarly, we also measure the performance of GPT-4 and ChatGPT under in-context learning settings. We randomly select 32 examples from the synthetic dataset D_L and use these as in-context examples for both GPT-4 and ChatGPT. The prompt we used for this experiment is shown in Table 14.

E.2 Instruction tuning experiments

In our instruction tuning experiments, we conducted evaluations on the 7B and 13B variants of *Vicuna* and *Llama-2*. To synthesize task-specific training data, we applied two approaches: 1) generate data with GPT-4 and 2) modify the existing lexical substitution benchmark Swords. In the following, we will present detailed descriptions of the two approaches, along with the specific experimental settings used in the fine-tuning process.

E.2.1 Synthesizing task-specific training data

Generate data from GPT-4 We start by selecting the context sentences from the TOEFL-11 dataset. Concretely, we randomly select 500 sentences from each proficiency level[11], rendering a total of $1,500$ sentences. We also make sure that the selected sentences are not in ProLex. Similar to section 3.1, we perform grammar error correction to the sentences extracted from low and medium-level essays, while sentences from high-level essays are retained in their original form. Subsequently, for each sentence, we prompt GPT-4 to select a target word based on a randomly pre-defined

[11] Note that there are three proficiency levels in the TOEFL-11 dataset, namely high, medium and low.

Model	Fine-tuning Data	Precision	Recall	$F_{0.5}$-score
T5-base	cLang-8	59.7	26.1	47.5
T5-base	cLang-8 + ErAConD	60.6	36.3	53.4
GPT2-large	cLang-8	**66.8**	**48.1**	**61.9**

Table 7: Performance of different GEC systems on the CoNLL-14 Shared Task test dataset, as measured by the official M^2Scorer (Dahlmeier and Ng, 2012).

Original Sentence	Corrected Sentence
Meanwhile when you have a tour guide, that means you are safer especially if he from the area which you will go that will be better.	Meanwhile when you have a tour guide, that means you are safer especially if **he is** from the area which you will **go to** that will be better.
but if you are a man who isn't open new ideas and you are't produce new things you are not be a good employee or a good boss.	but if you are a man who isn't **open to** new ideas and you **don't** produce new things you are **not a** good employee or a good boss.
There are two reasons for this statements: time and knowridge.	There are two reasons for this **statement**: time and **knowledge**.

Table 8: Examples of basic grammar error correction for sentences selected from TOEFL-11. The erroneous parts are marked in red, and the corresponding corrected portions are highlighted in yellow.

part-of-speech tag and generate the proficiency-oriented substitutes. The complete prompt for this process is shown in Table 13. Specifically, we provide five in-context examples to guide the generation process. As a result, we take the final substitutes generated for each sentence as the proficiency-oriented substitutes. The resulting synthesized data is post-processed to fine-tune *Vicuna* 7B/13B and *Llama-2* 7B/13B models. In total, after post-processing[12], the synthesized dataset contains $1,383$ unique (w, s) pairs, with $6,375$ candidate substitutes and $4,982$ final substitutes. The part-of-speech tags for all selected target words are uniformly distributed, with 361 nouns, 335 verbs, 357 adjectives, and 330 adverbs.

Modify existing benchmark Another way to synthesize training data is to modify the dataset provided by the Swords benchmark, filtering out lower proficiency level substitutes. Concretely, we only consider acceptable substitutes (i.e. score greater than 50%) in Swords, and perform filtering based on their CEFR levels.

The Swords dataset, with 1132 target word and context pairs, have an average of 60.7 substitutes per target, of which 4.1 are acceptable on average. After filtering out substitutes with a lower CEFR proficiency level than the target word, the dataset now contains an average of 3 substitutes per target. The average CEFR level of targets is A2, whereas the average level of substitutes after filtering is B1, indicating improvements on proficiency-level. The CEFR level distributions of the target words and proficiency-oriented substitutes from the filtered dataset D_S are shown in Figure 4.

E.2.2 Experiment setup

We established our training pipeline based on the platform developed by (Zheng et al., 2023). Concretely, for the 7B models, we fine-tuned both *Vicuna* and *Llama-2* for a maximum of 10 epochs each, using a single NVIDIA A100-80G GPU. For the 13B models, each variant was trained for up to 5 epochs on two NVIDIA A100-80G GPUs. In all cases, we configured the training batch size per device to 1 and established the initial learning rate at `1e-5`, employing a linear learning rate scheduler. The best checkpoints were selected based on the performance on a separate development set of ProLex.

[12]The post-processing involves dropping null values and duplicated rows.

Label	Criteria
1	· doesn't change the meaning and semantics of the sentence, **and** · is a common collocation or expression in actual English use, **and** · in general, matches at least one connotation of the target word in the context, **and** · grammatically correct
-1	· changes the meaning and semantics of the sentence, **or** · is not a common collocation or expression in actual English use **or** · doesn't match any connotation of the target word in the context **or** · is grammatically incorrect
0	· I do not know the definition of the substitute or target word · I do not know the meaning of the context sentence

Table 9: General annotation guidelines to assign labels to a given triplet (w, s, w'). Note that the sentence s can be hard to understand since it is written by beginner L2 English learners.

Sentence	Substitute	Label	Reason
For example, I **trust** a person who has more knowledge than I do	rely on	1	"I trust a person" entails "I rely on a person"
For example, in **private**, an airplane costs 3000 dollars in America from Japan.	personally	0	Unknown meaning of the sentence
In academic institutions they are introducing new technologies and topics which are going to be used full to the students that are going to change rapidly. Take as an example a **computer**.	laptop	0	Substitute is a hyponym of target word
If one wants to live a **better** and more plentiful life, I find it basic to experiment, dare a little, risk it a bit and try new things.	superior	1	Good substitute - fulfills all criteria
In **addition**, I will argue over the belief in the following reasons.	furthermore	-1	Ungrammatical substitute
The cow **jumped** over the moon.	leap	-1	Ungrammatical ("leap" vs "lept"), verb tense does not match.
He **grew** up in the town.	evolved	-1	Highlighed word is part of a phrase ("grow up")
It's just impossible for him to additionally work **voluntarily**.	willingly	0	Additional context is needed to judge if "voluntarily" refers to "working without pay" or "working freely"
Once the undergraduate studies are **pursued** by a student, the student is more aware of different subjects and the knowledge he has gained in the period of his studies.	embarked on	1	Figurative use case of "embarked on", which conveys the same meaning as "pursued".

Table 10: Example annotations and the reasons behind. Note that for each sentence, the target word is encompassed with double asterisks (**).

You are a helpful assistant to perform a lexical substitution task. Specifically, you will be given a tuple of text consisting of 1) context with target word indicated using asterisks, and a 2) natural language query.
You should generate exactly five substitutes separated by commas. Do not generate the same word as the target word.
Here are some examples:

I have completed the invoices for April, May and June and we owe Pasadena each month for a **total** of $3,615,910.62.
I am waiting to hear back from Patti on May and June to make sure they are okay with her.
Q: What are appropriate substitutes for **total** in the above text?
A: amount, sum, price, balance, gross

. . . I thought as much. Now leave, before I **call** the rats on you." We left.
Q: What are appropriate substitutes for **call** in the above text?
A: summon, order, rally, send, sic

The e-commerce free **zone** is situated in north Dubai, near the industrial free **zone** in Hebel Ali.
Q: What are appropriate substitutes for **zone** in the above text?
A: sector, district, area, region, section

The state's action, the first in the nation, has the blessing of the American Psychological Association (APA), which considers prescriptive authority a **logical** extension of psychologists' role as health-care providers.
Q: What are appropriate substitutes for **logical** in the above text?
A: rational, reasonable, sensible, justifiable, relevant

They nodded. "Oh, **totally**," said the hunchback. "I get that all the time around here."
Q: What are appropriate substitutes for **totally** in the above text?
A: absolutely, for sure, surely, completely, definitely

{Insert CONTEXT with **TARGET** here}
Q: What are appropriate substitutes for **{TARGET}** in the above text?
A:

Table 11: The prompt for GPT-4 to generate candidate substitutes for human annotation. Note that we incorporate five in-context examples from Swords (Lee et al., 2021).

You are about to perform a lexical substitution task, considering the proficiency level of the substitute compared to the target word in a sentence. The task is to generate a set of candidate substitutes separated by commas for a target word in a given sentence. The target word is highlighted in the sentence, encompassed by two double asterisks.

Target word: [TARGET]
Sentence: [SENTENE]
Substitutes:

Table 12: Prompt for zero-shot evaluation of GPT-4 and ChatGPT on ProLex. The prompt takes a given target word and a context sentence, and outputs a list of substitutes separated by commas.

You are about to synthesize data for a lexical substitution task, considering the proficiency level of the substitute compared to the target word in a sentence.
Concretely, for each data point, I will first give you an original sentence. Then, you should follow the following steps to create a complete data point:
1) Based on the queried Part of Speech tag, select as unique as possible a content word as the target word to be substituted from the sentence (Content words include nouns, verbs, adjectives and adverbs). Do only select single word as the target;
2) generate at least 10 candidate substitutes (separated by commas) for the selected content word from Step 2;
3) final candidate substitutes after excluding candidates from Step 3 that are not common expressions in actual English use.

You should make sure each of the generated substitutes follows exactly the following characteristics:
a) does NOT change the meaning and semantics of the sentence, and
b) is a common collocation or expression in actual English use (appears at least five times in Corpus of Contemporary American English), and
c) in general, matches at least one connotation of the target word in the context sentence, and
d) is grammatically correct, and
e) has an equal or higher language proficiency level compared to the target word. Please use CEFR (Common European Framework of Reference for Languages) standard to describe the language proficiency of a word. The specification of CEFR levels (from the least proficient to the most proficient) is defined as follows: A1 (beginner), A2 (Elementary), B1 (Intermediate), B2 (Upper Intermediate), C1 (Advanced), C2 (Proficient).

Here are some examples (Tagged sentence denotes sentence where the target word is surrounded by two double asterisks). Do not change the original sentence:
Query Part of Speech tag: adverb
Original Sentence: Students can learn to study independently from understanding ideas and concepts.
Tagged Sentence: Students can learn to study **independently** from understanding ideas and concepts.
Target word: independently (B2 - Upper Intermediate)
Candidate Substitutes: autonomously (C2 - Proficient), individually (C1 - Advanced), solo (B2 - Upper Intermediate)
Final Substitutes: autonomously, individually, solo

Query Part of Speech tag: adjective
Original Sentence: It is because of this that various kinds of people with special knowledge can complement each other.
Tagged Sentence: It is because of this that various kinds of people with **special** knowledge can complement each other.
Target word: special (AI - beginner)
Candidate Substitutes: specific (A2 - Elementary), distinctive (C1 - Advanced), exclusive (B2 - Upper Intermediate), unique (B2 - Upper Intermediate), particular (A2 - Elementary)
Final Substitutes: specific, distinctive, unique, particular

Query Part of Speech tag: noun
Original Sentence: At the start of a life a person doesn't have success yet, only during life does your action make your success.
Tagged Sentence: At the start of a life a person doesn't have success yet, only during life does your **action** make your success.
Target word: action (A1 - beginner)
Candidate Substitutes: behavior (A2 - Elementary), conduct (B2 - Upper Intermediate), operation (B1 - Intermediate), undertaking (C1 - Advanced), activity (A1 - beginner)
Final Substitutes: behavior, conduct, activity

Query Part of Speech tag: verb
Original Sentence: It has no arguments to support it and is terribly broad.
Tagged Sentence: It has no arguments to **support** it and is terribly broad.
Target word: support (A2 - Elementary)
Candidate Substitutes: back (B2 - Upper Intermediate), substantiate (C2 - Proficient), uphold (C1 - Advanced), justify (B2 - Upper Intermediate)
Final Substitutes: back, substantiate, justify

Query Part of Speech tag: adverb
Original Sentence: many old people have diseases that rob them of their health and make them unable.
->POS does not exist in the sentence.

Please note that if there are no words that correspond to the queried part of speech tag in the original sentence, simply generate "POS does not exist in the sentence".
Do only select single word as the target.

Now, please generate:

Query Part of Speech tag: [Query_POS]
Original sentence: [Sentence]->

Table 13: Prompt used to synthesize training data for fine-tuning: it generates 1) the target word and tags the target word in the sentence; 2) an initial set of candidate substitutes along with their CEFR proficiency level; and 3) the final proficiency-oriented substitutes.

You are about to perform a lexical substitution task, considering the proficiency level of the substitute compared to the target word in a sentence.
The task is to generate a set of candidate substitutes separated by commas for a target word in a given sentence. The target word is highlighted in the sentence, encompassed by two double asterisks.
Here are some examples:
Target word: honest
Sentence: This is not right as people think that the message sent to them is **honest**, thus they believe whatever they hear.
Substitutes: truthful, sincere, genuine

Target word: fully
Sentence: that provides them with access to enjoy their life **fully**.
Substitutes: completely, totally, absolutely, wholly

Target word: roads
Sentence: the government may build more **roads** so that they provide translation.
Substitutes: highways, tracks, routes, lanes

Target word: backs
Sentence: Few year **backs**, music industry was dominated by the Walkman from Sony.
Substitutes: ago, previously, earlier, formerly

Target word: For example
Sentence: **For example**, let' s talk about the French Revolution.
Substitutes: For instance, To illustrate

Target word: better
Sentence: But if you look into it, why do pairs of shoes coast so much, because it is an advertisement that Nike makes their shoes look **better** than everyone else's.
Substitutes: superior, improved, exceptional

Target word: any
Sentence: Learning facts in a scientific field, like for example the resulting speed of an object, does not give **any** clue about the result of an experiment in another context.
Substitutes: whatsoever, at all, absolutely

Target word: rather
Sentence: When they achieve one particular goal they proceed with a new goal **rather** than doing what they already know.
Substitutes: preferably, instead, alternatively, ideally

Target word: Balancing
Sentence: **Balancing** between these methods is actually what we need.
Substitutes: Harmonizing

Target word: come
Sentence: we leave like one day will **come**
Substitutes: arrive, appear, emerge

... (22 more examples)

Target word: [TARGET]
Sentence: [SENTENCE]
Substitutes:

Table 14: Prompt for in-context learning with GPT-4 and ChatGPT.

(a) Target word in D_S.

(b) Proficiency-oriented substitutes in D_S.

Figure 4: Distribution of CEFR levels for target words and proficiency-oriented substitutions in D_S.