

Exploiting the English Vocabulary Profile for L2 word-level vocabulary assessment with LLMs

Stefano Bannò, Kate M. Knill, Mark J. F. Gales

ALTA Institute, Department of Engineering, University of Cambridge (UK)
{sb2549, kmk1001, mjfg100}@cam.ac.uk

Abstract

Vocabulary use is a fundamental aspect of second language (L2) proficiency. To date, its assessment by automated systems has typically examined the context-independent, or part-of-speech (PoS) related use of words. This paper introduces a novel approach to enable fine-grained vocabulary evaluation exploiting the precise use of words within a sentence. The scheme combines large language models (LLMs) with the English Vocabulary Profile (EVP). The EVP is a standard lexical resource that enables in-context vocabulary use to be linked with proficiency level. We evaluate the ability of LLMs to assign proficiency levels to individual words as they appear in L2 learner writing, addressing key challenges such as polysemy, contextual variation, and multi-word expressions. We compare LLMs to a PoS-based baseline. LLMs appear to exploit additional semantic information that yields improved performance. We also explore correlations between word-level proficiency and essay-level proficiency. Finally, the approach is applied to examine the consistency of the EVP proficiency levels. Results show that LLMs are well-suited for the task of vocabulary assessment.

1 Introduction

Automated writing evaluation has become an established area of research within natural language processing, playing a key role in language education and computer-assisted language learning (Huawei and Aryadoust, 2023). Within this context, assessing second language (L2) proficiency remains a critical objective, with increasing attention being paid not only to holistic assessment but also to the evaluation of individual aspects of language competence (Weigle, 2002), such as grammar, coherence, and vocabulary. This has proven essential for providing detailed, pedagogically useful feedback to language learners (Hamp-Lyons, 1995). Among these components, vocabulary is the es-

sential building block of language (Schmitt et al., 2001) as words are the main vehicle for expressing meaning (Vermeer, 2001).

Existing vocabulary assessment methods have typically either assigned an overall vocabulary score to learners’ texts at the essay- (Crossley et al., 2023; Bannò et al., 2024) or sentence-level (Arase et al., 2022), extracted vocabulary-related features (e.g., lexical diversity or sophistication) (Kyle and Crossley, 2015; Kyle et al., 2018), or attempted to assess the difficulty or appropriateness of individual words (Bax, 2012; Uchida and Negishi, 2018; Settles et al., 2020; Aleksandrova and Pouliot, 2023). The latter, however, has been largely under-explored and presents significant challenges – particularly in dealing with polysemy and contextual variation as well as handling multi-word expressions.

Our work directly addresses this limitation by focusing on word-level, in-context vocabulary assessment in a fully replicable manner. Specifically, we leverage an open-access reference containing information about which lexical items are used at each level of English learning, the English Vocabulary Profile (see Section 3.1). We combine this resource with state-of-the-art large language models (LLMs) to predict the proficiency level of individual words as they are used in L2 learner writing. To the best of our knowledge, this is the first study to apply LLMs for this task, offering a novel and robust approach to vocabulary assessment that explicitly handles semantic ambiguity and contextual nuance.

In Section 2, we outline the theoretical background of L2 vocabulary assessment and review previous studies that have implemented it in automated systems. Section 3 presents the English Vocabulary Profile and the L2 datasets used in our experiments. Section 4 constitutes the core of our study. The first experiment focuses on identifying the intended meaning of polysemous words from the English Vocabulary Profile through the use of

LLMs, using learner example sentences sourced from the same dataset. In our second experiment, we annotate sentences from a learner corpus by assigning each word a proficiency level based on the English Vocabulary Profile – annotations we plan to make publicly available. We then automatically predict these levels and compare the performance of various LLMs against a random baseline and a part-of-speech (PoS)-based model. The third experiment extends this approach to additional L2 learner datasets annotated only at the essay level, investigating the correlation between predicted word-level proficiency and essay-level proficiency. Additionally, at the end of the same section, we use our LLM-based approach to examine the consistency of the annotations in the English Vocabulary Profile. Finally, in Section 5, we present our conclusions and outline directions for future work.

2 Related Work

2.1 Vocabulary assessment

Despite its fundamental importance, the assessment of vocabulary was only selectively investigated at the beginning of the scientific era of L2 assessment, whereas much more attention was paid to the contrastive analysis of sounds and grammar (Lennon, 2008). When vocabulary knowledge was evaluated, it was primarily tested using the discrete-point approach, an assessment method focused on testing one specific linguistic element – phonology, morphology, syntax, and vocabulary – at a time, generally using multiple-choice questions. This approach to vocabulary testing faced various criticisms, as it offered only a limited view of a learner’s vocabulary knowledge, neglected the role of productive language use, disregarded the importance of context in real-world communication, and failed to consider learners’ use of strategies to cope with unfamiliar words (Read, 2000).

The 1980s represented a watershed in vocabulary assessment since a group of researchers started to publish studies on defined procedures aiming at assessing specific aspects of vocabulary use and knowledge (Anderson and Freebody, 1981, 1983; Nation, 1983; Meara and Buxton, 1987). These seminal works were something of an exception, given that, on the one hand, the field of L2 acquisition was primarily concerned with the investigation of the acquisition by learners of morpho-syntactic features, whereas, on the other hand, the advent of the communicative approach shifted the attention

of language assessment researchers from knowledge of grammatical and lexical elements to the performance of real-world-like tasks (Read, 2013).

For our work, we believe it is important to remember Read’s conceptualisation of vocabulary assessment, who classified it according to 6 dimensions arranged in antonymic pairs: discrete versus embedded, selective versus comprehensive, and context-independent versus context-dependent. The first distinguishes whether vocabulary is assessed as an isolated skill (discrete) or as part of broader language proficiency (embedded). The second refers to the scope of lexical items – either a specific set (selective) or the learner’s full vocabulary range (comprehensive). The third dimension captures whether vocabulary is assessed in isolation or within authentic contexts. Due to the widespread acceptance of the communicative approach (Harding, 2014), it is straightforward to conclude that current trends in language testing and assessment tend to privilege *embedded*, *comprehensive*, and *context-dependent* measures of vocabulary assessment. These three characteristics are central to our approach.

2.2 Lexical sophistication

The Common European Framework of Reference (CEFR) (Council of Europe, 2001, 2020), a key benchmark aligned with communicative language teaching, testing and assessment, distinguishes between vocabulary range and vocabulary control, which have generally been operationalised along the dimensions of lexical diversity and lexical sophistication, respectively.

Lexical diversity, concerning the breadth of vocabulary used by learners (Yu, 2010; Lu, 2012), is typically measured through metrics like type-token ratio or number of unique words. Its relationship with L2 writing proficiency has been widely studied (Crossley and McNamara, 2012; Gebril and Plakans, 2016; Treffers-Daller et al., 2018; Woods et al., 2023). While important, lexical diversity is not the primary focus of this paper.

Our work is more closely related to the idea of lexical sophistication. Its focus is the depth of lexical knowledge and is frequently characterised by the presence of relatively rare or uncommon words within a given language sample (Baese-Berk et al., 2021). It is generally operationalised using features related to word frequency and familiarity, such as the Lexical Frequency Profile, which reflects the proportion of a learner’s vocabulary falling within

various frequency bands derived from a reference corpus (Laufer and Nation, 1995). The English Vocabulary Profile (EVP) (Capel, 2015), adopted in our work (see Section 3), is a resource that describes words, phrases, idioms, and collocations used by English learners at different CEFR levels. The study by Leńko-Szymańska (2015), which represents an important precedent for our work, employed it to assign proficiency bands to 90 essays, finding a strong correlation between the clusters obtained using the vocabulary profile and the human-assigned CEFR levels.

Similarly to the EVP, Dürlich and François (2018) created an online database called EFLLex, which presents the distribution of English words across CEFR levels (A1 to C1), mainly derived from the analysis of textbook corpora. The CEFR-J, a Japan-specific adaptation of the CEFR, also features a word list with proficiency levels (Tono, 2013).

2.3 Automated approaches

Yoon et al. (2012) investigated the use of a vocabulary profile to extract features of lexical sophistication for proficiency assessment of spontaneous speech and found interesting correlations with oral proficiency scores. Kyle and Crossley (2015) introduced the Tool for Automatic Analysis of Lexical Sophistication (TAALES), which computes 135 lexical indices. They found that 5 measures of lexical complexity accounted for more than 50% of the variance in the human ratings of the spoken and written datasets considered in their study. Text Inspector (Bax, 2012) is an online tool that appears to use the EVP to assess writing proficiency; however, its implementation is not publicly available and is most likely rule-based, as it presents all possible meanings (and corresponding proficiency levels) of a word to the user when ambiguities arise. The calculation method used by the CVLA (CEFR-based Vocabulary Level Analyzer) is openly available (Uchida and Negishi, 2018); however, the strictly vocabulary-related part of this tool is also rule-based and, like Text Inspector, does not appear to address issues related to polysemous or ambiguous words.

Duolingo developed and released a tool called the CEFR checker, which is now discontinued. This tool allowed users to assess the difficulty level of English and Spanish words and texts. The lexical complexity component of the tool is described in Settles et al. (2020) and features a vocabulary

scale model based on the CEFR framework and a database of 6,823 English words, partly obtained from the EVP. The authors introduce two regression models trained on lexical representations using surface-level features designed to approximate word frequency. However, these models do not appear to account for multi-word expressions or words with multiple meanings.

Garí Soler and Apidianaki (2021) demonstrated that BERT could effectively generate contextual embeddings for polysemous words, laying the groundwork for further research in lexical complexity assessment. Building on this, Aleksandrova and Pouliot (2023) explored how the most frequent senses of polysemous words appear in language learner essays. Their study leverages BERT to develop a CEFR-aligned classifier aimed at evaluating the lexical complexity of both single-word and multi-word expressions in English and French. It is important to note that their classifier is trained to predict the CEFR level of an item in context, but not to explicitly identify or disambiguate the meaning or sense of that expression.¹

While the aforementioned studies partially or entirely fall short in handling polysemy and ambiguity, our study directly addresses this challenge by targeting word-level, in-context vocabulary assessment in a fully replicable manner. Additionally, to the best of our knowledge, ours is the first study to leverage LLMs for this purpose.

3 Data

3.1 English Vocabulary Profile

The English Vocabulary Profile (EVP) (Capel, 2015) is a publicly available reference² that contains information about which words, phrases, idioms, and collocations are used at each level of English learning. It is grounded in extensive research using the Cambridge Learner Corpus (Nicholls, 2003), a growing collection of exam scripts written by learners worldwide.

For our experiments, we only considered the British English section of the profile, which includes 15671 entries corresponding to 6747 unique words. The difference arises because some words have multiple meanings or grammatical functions, resulting in several entries comprising multiple op-

¹Other works have targeted word-level vocabulary assessment, but not for English (Gala et al., 2014; Alfter et al., 2016; Alfter and Volodina, 2018).

²englishprofile.org/

tions. Additional details about the degree of polysemy are provided in Table 6 in Appendix A.

Base Word	Guideword	Level	Part of Speech	Topic	Details
if/when push comes to shove	IDIOM	C2	phrase		
push	MOVE SOMEONE/SOME THING	A2	verb	people: actions	
push	PRESS	B1	verb	people: actions	
push	MOVE YOURSELF	B1	verb	people: actions	
push	ENCOURAGE	C1	verb	people: actions	
push	ENCOURAGEMENT	C1	noun		
push	PRESS	B1	noun	people: actions	
push (sb) for sth/v to do sth		B2	phrase	communication	
pushy		C2	adjective	people: personality	
push yourself		B2	phrase	people: actions	

Figure 1: Example for the word *push* from the EVP.

Each EVP entry includes a base word, a guide-word, its CEFR level(s), manually assigned PoS, topic, and additional details such as learner and dictionary examples in context (see Figure 1). Among all entries, when distinguishing by PoS, 62.24% have more than one CEFR level. However, when considering only unique base words, this proportion drops to 29.21%. In this work, we refer to words that, within a single PoS, have multiple CEFR levels as *ambiguous* words; the rest are considered *non-ambiguous*. For example, the noun *aim* is classified solely as B1, making it non-ambiguous, whereas the verb *aim* spans levels A2, B2, C1, and C2 depending on context, making it ambiguous. About 95% of the section of the EVP we considered in our work has examples taken from L2 learner writing, many of which also contain grammatical errors. For the remaining 5%, we relied on dictionary examples in our experiments.

3.2 L2 learner datasets

3.2.1 OneStopEnglish

OneStopEnglish (Vajjala and Lučić, 2018), is a publicly available corpus³ for readability assessment and text simplification including 189 parallel compositions across three readability levels: Elementary, Intermediate, and Advanced. We extracted 293 triplets of parallel sentences from the corpus, spanning these three readability levels (see Appendix B for an example), and annotated them at the word level with CEFR levels from the EVP, excluding stopwords,⁴ punctuation, and words not

³github.com/nishkalavallabhi/OneStopEnglishCorpus

⁴Stopwords include those from the NLTK list (nltk.org) plus *across*, *among*, and *away*.

featured in the EVP (see Figure 2). These annotations will be made publicly available.

3.2.2 EFCAMDAT

Arguably the largest publicly available L2 learner corpus,⁵ the second release of the EF-Cambridge Open Language Database (EFCAMDAT) (Geertzen et al., 2013; Huang et al., 2017) consists of 1,180,310 scripts written by 174,743 L2 learners as assignments for Englishtown, an online English language school. The corpus includes 128 distinct writing tasks covering a range of topics, such as describing the rules of a game, reporting a news story, explaining a homemade remedy for fever, and writing to a pen pal. Each composition is annotated with a proficiency level from 1 to 16, corresponding to CEFR levels A1 to C2.⁶ Learners' first languages are not directly available but can be inferred from their nationalities (approximately 200 in total). For our experiments, we randomly selected 1,000 essays for each CEFR level from the corpus, resulting in a total of 6,000 essays.

3.2.3 ELLIPSE

The English Language Learner Insight, Proficiency, and Skills Evaluation (ELLIPSE) Corpus (Crossley et al., 2023) is a publicly available collection⁷ of approximately 6,500 writing samples from L2 learners of English. Each sample is annotated with both an overall holistic proficiency score and detailed analytic scores covering aspects such as cohesion, syntax, vocabulary, phraseology, grammar, and punctuation conventions. Proficiency scores are on scale from 1 to 5. For our experiments, we extracted 359 essays from the training set and 175 from the test set. Further details about the selected essays can be found in Table 7 in Appendix A.

4 Experiments

4.1 Semantic understanding

Our first experiment serves as a proof of concept and focuses on semantic understanding, with the goal of identifying the intended meaning of a unique word from the EVP when considering the EVP learner examples. Specifically, an LLM is provided with an EVP learner-produced example sentence containing a given word along with

⁵ef-lab.mml.cam.ac.uk/EFCAMDAT.html

⁶The official EF-CEFR mapping can be found at: myenglishlive.ef.com/help-me-article?articleID=55&Vote=Up

⁷github.com/scrosseye/ELLIPSE-Corpus

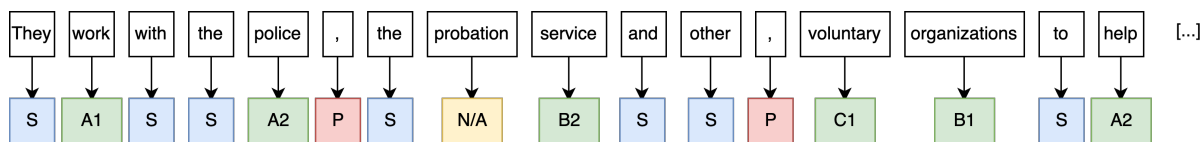


Figure 2: Annotation example on OneStopEnglish. S: stopword; P: punctuation; N/A: word not featured in the EVP.

all possible EVP entries for that given word. The model’s task is to select the most contextually appropriate meaning from the options provided. Note that multi-word expressions are also included by feeding their reference words, e.g., the word *push* for the expression *push (sb) for sth/to do sth* (see Figure 1).

The prompt used for this task and a relevant example are reported in Appendix C (“Prompt for semantic understanding”). For this experiment, we focused on words with 3, 4, 5, and 6 possible meanings. To avoid positional bias (Liusie et al., 2024), all possible permutations were considered for words with 3 options.⁸ For words with 4, 5, and 6 options, however, only ten permutations were considered due to time and resource constraints. For each permutation, we extract the logits for each option, apply a softmax function, then compute the average probability across permutations and select the option with the highest average probability.

For this experiment, we compared the performance of two proprietary LLMs, i.e., GPT-4o, GPT-4o-mini (OpenAI, 2023), and three open-source LLMs, i.e., Llama 3.1 8B, Llama 3.1 70B (4-bit quantised) (Llama Team, 2024), and Qwen 2.5 32B (4-bit quantised) (Yang et al., 2024). These models were selected to ensure a representative range in terms of both model size and the open-source versus proprietary distinction.

Results are evaluated in terms of Accuracy.

4.2 Word-level proficiency prediction

In the second part of our work, we perform CEFR proficiency level prediction at the word level on the annotated sentences extracted from the OneStopEnglish dataset (see Section 3.2 and Figure 2). We implement the approach proposed in Figure 3 in order to extract the proficiency level for each word in a given composition.

⁸In this context, a permutation refers to a unique ordering of the multiple-choice options associated with each question. That is, while the content of the options remains the same, their positions (e.g., labeled A, B, and C) are shuffled across different permutations. This ensures that the model’s predictions are not influenced by the fixed position of any particular option, thereby reducing positional bias.

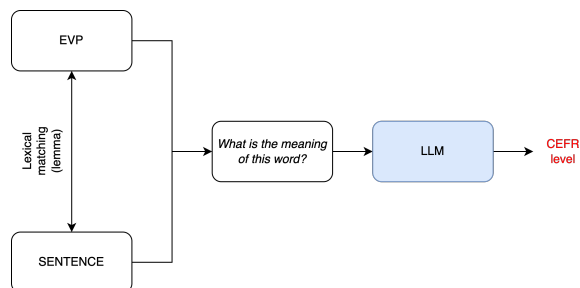


Figure 3: High-level diagram of the approach proposed for word-level CEFR prediction.

Since it is not feasible to feed all EVP entries for each word in a sentence into the LLM at once, we use spaCy⁹ to lemmatise each word and map the resulting lemma to its corresponding entries in the EVP. For example, for the word *work* in the sentence shown in Figure 2, the lemma remains *work*, which we then match to all corresponding *work* entries in the EVP. This allows us to filter out irrelevant EVP words in each LLM run.

[They] work with the police [...]
 They [work] with the police [...]
 They work [with] the police [...]
 They work with [the] police [...]
 They work with the [police] [...]

Figure 4: Highlighting method.

For each sentence we iterate through each word by highlighting it in square brackets, as shown in Figure 4. If there is no match between a given lemmatised word in the sentence and the EVP, we automatically assign *N/A* to the word (see Figure 2). In all the other cases, we add a “None of the other options” choice to the other available options and ask the LLM to choose the most suitable option. Then, we extract the logits for each option, apply a softmax function, and select the option with the highest average probability. Finally, we select the

⁹spacy.io

CEFR level assigned to this option. In this set of experiments, we also feed the manually assigned PoS information and a brief definition associated to each option.

The prompt used for these experiments as well as a relevant example can be found in Appendix C (“Prompt for word-level CEFR prediction”).¹⁰

It is important to note that our approach can effectively handle multi-word expressions. For example, when encountering the phrasal verb *take off* in the sense of “becoming airborne”, the method processes *take* and *off* separately. The word *off* is treated as a stopword and thus excluded from the final assessment. However, *take* is still evaluated in context, allowing the LLM to infer the intended meaning from the surrounding sentence or essay, in addition to the PoS information (i.e., *verb*) and the EVP definition (i.e., *If an aircraft takes off, it leaves the ground and begins to fly.*). This ensures that the CEFR level assigned reflects the actual usage rather than the isolated form. Similarly, for the expression *take advantage of sth*, both *take* and *advantage* are mapped to the same corresponding multi-word expression in the EVP, while the stopword *of* is excluded from the evaluation.

For this part of our work, we try the same LLMs as Section 4.1 except for GPT-4o-mini and Llama 3.1 8B. We compare the performance of these LLMs with a PoS-based system, which only relies on PoS tags extracted using spaCy. Specifically, if a given lemma in the sentence has only one entry in the EVP then its corresponding level is automatically assigned to the word (e.g., *voluntary* in Appendix B has only one EVP entry, hence one CEFR level). Otherwise the relevant PoS tag is matched with the corresponding one in the EVP. If the EVP only specifies one CEFR level for this lemma/PoS pair then this level is automatically assigned to the word (e.g., *criminal* in Appendix B has two entries – noun and adjective – at two different CEFR levels). Where the pairing is assigned to multiple CEFR levels in the EVP, the lowest of these CEFR levels is assigned to the word (e.g., the word *service* in the example in Appendix B has 6 noun entries with 3 different CEFR levels). Additionally, we consider a random baseline that relies solely on lexical matching. In this case, for words with a single entry, such as *voluntary*, after lemmatisation, we simply assign the associated CEFR

level. For words with multiple levels, we randomly pick a level from the available EVP entries.

Results are evaluated in terms of overall Accuracy. Additionally, performance on individual CEFR levels is reported using F_1 score.

4.3 Essay-level proficiency prediction

The third part of our work applies the approach presented in the previous section to additional L2 learner datasets, namely EFCAMDAT and ELLIPSE (see Sections 3.2). Unlike the On-StopEnglish data, these datasets are annotated only at the essay level and lack word-level CEFR labels. We therefore use our approach to predict the CEFR level of each word and leverage this information as features for essay-level proficiency prediction. For this part of the work, we only used Qwen 2.5 32B¹¹ and the PoS-based model to extract vocabulary-related information. In other words, we aim to explore the predictive power of vocabulary-related features extracted through our models in the context of holistic proficiency assessment. Additionally, since the ELLIPSE data include analytic scores targeting specific aspects of proficiency, we further investigate the relationship between these features and scores related to dimensions such as cohesion, syntax, vocabulary, phraseology, grammar, and punctuation conventions.

To do this, for EFCAMDAT, we employ a naive classifier that uses the distribution of predicted CEFR levels within each essay. Specifically, we compute the proportion of words at each predicted CEFR level (i.e., the count of words at each level divided by the total essay length), weight these proportions by their corresponding CEFR levels (i.e., 1 for A1, 2 for A2, 3 for B1, etc.), and sum them to obtain a composite score. We then assess the correlation between this score and the human-assigned holistic score. In addition to the naive classifier, we also employ a simple Support Vector Regression (SVR) model with default parameters (i.e., $\epsilon = 0.1$, $C = 1$, and a radial basis function kernel). The model is trained using the proportions of words at each predicted CEFR level as input features and evaluated using 5-fold cross-validation.

The SVR is also employed for the ELLIPSE dataset. The model is trained on the training split to predict the holistic proficiency scores. Subse-

¹⁰We also tried this task by prompting the LLMs without EVP information, but results were significantly worse.

¹¹The prompt is the same reported in Appendix C (“Prompt for word-level CEFR prediction”) with the only difference that “sentence” is replaced with “essay”.

quently, it is evaluated on the test set for both holistic and analytic scores prediction (see Section 3.2).

Results are evaluated in terms of Pearson’s correlation coefficient (PCC) and Spearman’s rank coefficient (SRC).

Experimental results

4.4 Semantic understanding results

Table 1 reports the results in terms of Accuracy for the semantic understanding task. As can be seen, GPT-4o achieves the best performance, followed by Llama 3.1 70B, with Qwen 2.5 32B performing nearly on par. While model size appears to play a significant role in the performance gap between GPT-4o and the other models, this pattern does not hold when comparing Qwen 2.5 32B to Llama 3.1 70B, despite the latter being more than twice as large. All models show a remarkable performance for this task with the exception of Llama 3.1 8B. As expected, Accuracy decreases as the number of options available for a given word increases.

A reasonable question to ask is whether these LLMs have been exposed to the EVP during training, given that it is publicly available. This consideration, among others, motivated us to extend our experiments to additional L2 learner datasets.

Model	No. of options				avg.
	3	4	5	6	
GPT-4o	89.0	86.2	83.1	79.5	84.4
GPT-4o_{mini}	84.4	78.9	75.4	71.2	77.5
Llama3.1_{70B}	85.1	82.0	78.9	73.1	79.8
Llama3.1_{8B}	77.4	70.0	64.3	64.4	69.0
Qwen2.5_{32B}	85.6	80.8	76.4	75.4	79.6

Table 1: Accuracy (%) of EVP classification results.

4.5 Word-level proficiency prediction results

Model	Ambig.	Non-amb.	All
Random	29.1	88.7	61.6
PoS-based	66.7	93.4	80.7
GPT-4o	75.1	90.5	83.3
Llama 3.1 70B	76.9	91.5	84.6
Qwen 2.5 32B	80.5	92.8	87.0

Table 2: Accuracy (%) results for word-level CEFR prediction (OneStopEnglish).

Table 2 shows the results in terms of Accuracy for the task of word-level CEFR level prediction on the OneStopEnglish data. As mentioned in

Section 3.1, we refer to words that have multiple CEFR levels within a single PoS as *ambiguous* words, whereas the rest are considered *non-ambiguous*. We report the results for ambiguous, non-ambiguous, and all words. As expected, the Random classifier performs extremely poorly when predicting the level of ambiguous words. Incorporating PoS information leads to noticeable improvements, as demonstrated by the PoS-based model. However, the best performance on both ambiguous and overall cases is achieved – in increasing order – by GPT-4o, Llama 3.1 70B, and Qwen 2.5 32B. Remarkably, Qwen, despite being the smallest of the three LLMs, outperforms all others on this task. The improved performance of LLMs is likely due to their ability to leverage semantic information in addition to grammatical and syntactic knowledge.

For non-ambiguous words, as expected, the Random classifier achieves relatively decent performance. In this setting, the PoS-based model yields the highest accuracy, followed closely by Qwen 2.5 32B. One might expect the PoS-based model to achieve perfect accuracy; however, this is not the case, as the data include instances where words are used with meanings that are not covered in the EVP, leading to misclassifications by the PoS-based model. To address such cases, we included a “None of the other options” choice among the LLM’s available options (see Section 4.2) – though this option appears to be over-selected by the model, potentially affecting its accuracy.

Another reason why the LLMs do not outperform the PoS-based model for non-ambiguous words lies in the way options are selected. When multiple PoS entries exist for a given lemma, the PoS-based model uses only the relevant, non-ambiguous entry matching the POS tag. In contrast, all available options – including those with irrelevant PoS tags – are fed into the LLMs. For instance, as illustrated in Section 3.1, if a sentence contains the word *aim* used as a noun (which appears in the EVP only at the B1 level), the PoS-based model considers only this entry. However, the LLM is presented with both the noun and verb entries (at A2, B2, C1, and C2 in the EVP) for *aim*.¹² These findings are further supported by the performance figures in Table 9 in Appendix D, which reports the breakdown by word-level CEFR level in terms of F_1 score.

¹²We deliberately chose not to filter out these entries to avoid assuming perfect accuracy from the spaCy tagger.

Overall, these results suggest that the most effective solution may lie in a hybrid approach, where an LLM is used to handle ambiguous words, while a PoS-based model deals with non-ambiguous ones.

4.6 Essay-level proficiency prediction results

We use the same approach to predict word-level CEFR levels on additional L2 learner data, which are only annotated at the essay level. Starting from EFCAMDAT, Figure 5 shows a cumulative plot of the predicted normalised distribution of words for each word-level CEFR level across essay-level CEFR levels. We intentionally reverse the order of CEFR levels on the x-axis to emphasise that vocabulary usage is indicative of a certain proficiency level or higher, rather than exclusive to that level. For example, B1-level items are also commonly used by more advanced learners.

Focusing on a specific element of the figure, if we observe B1 level essays (represented by the green line), we observe that vocabulary from the C2 and C1 levels is used very little, while B2 level vocabulary appears to some extent – more than in A1 and A2 level essays, but less than in B2, C1, and C2 level essays. Vocabulary from B1 and lower levels is used even more frequently.

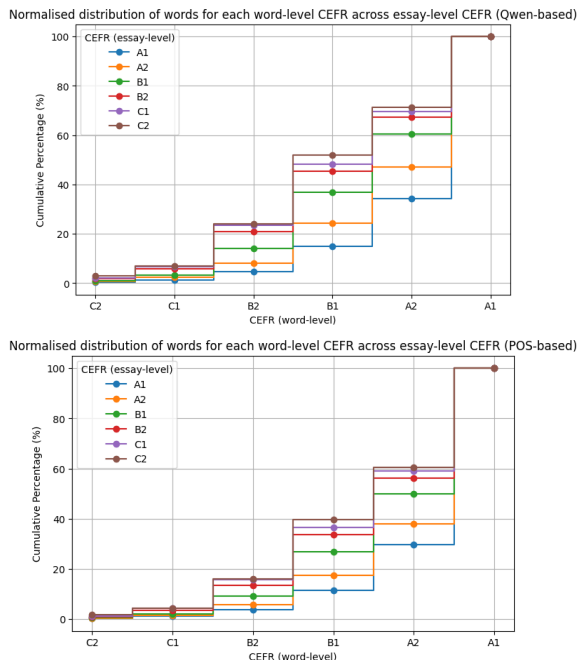


Figure 5: Predicted normalised distribution of words for each word-level CEFR level across essay-level CEFR levels. Qwen 2.5 32B vs PoS-based (EFCAMDAT).

When considering the overall trends, words from A2 to B2 show a steady increase in usage as essay-

level proficiency progresses. Interestingly, higher-level words (i.e., C1 and C2) are not as frequently used even in essays written at higher proficiency levels. Nonetheless, a moderate distinction across essay-level proficiency bands can still be observed. To this end, we present Figure 6, which displays the empirical cumulative distribution function (eCDF) of the AUC (Area Under the Curve) values computed from Figure 5. In both figures, we observe larger gaps – indicating better differentiation across levels – when using Qwen compared to the PoS-based model.

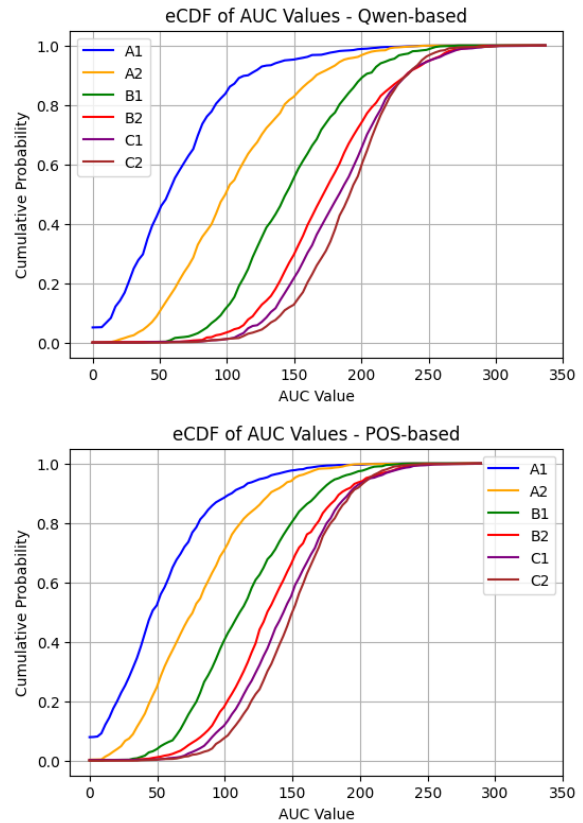


Figure 6: eCDF of AUC values. Qwen 2.5 32B vs PoS-based (EFCAMDAT).

These findings are further supported by the essay assessment results reported in Table 3. When used as features for predicting holistic scores, the vocabulary information extracted with Qwen consistently outperforms that derived from the PoS-based model both in the naive classifier and by the SVR.¹³

Finally, we report the results of our experiments on ELLIPSE using Qwen in combination with an SVR trained on holistic scores (see Section 4.3). As shown in Table 4, despite the high inter-correlation

¹³These experiments aim to demonstrate the predictive power of the extracted features, not to achieve state-of-the-art essay scoring.

Features	Classifier	PCC	SRC
PoS	Naive	0.580	0.603
Qwen		0.636	0.656
PoS	SVR	0.734	0.713
Qwen		0.771	0.749

Table 3: Results for essay-level holistic proficiency prediction (EFCAMDAT).

among all analytic scores (see Table 8 in Appendix A), the SVR (trained on holistic scores) predictions correlate most strongly with the Vocabulary scores, followed by Phraseology, suggesting our features are effectively targeting lexical aspects of language.

	PCC	SRC
Overall	0.650	0.624
Vocabulary	0.637	0.627
Phraseology	0.630	0.614
Grammar	0.577	0.556
Syntax	0.584	0.547
Cohesion	0.595	0.569
Conventions	0.613	0.578

Table 4: Results for essay-level holistic and analytic proficiency prediction using Qwen-SVR (ELLIPSE).

4.7 Word-level analysis

Finally, to evaluate the consistency of the EVP, we reverse the approach used thus far. Specifically, we select the two most common words with multiple meanings in the EFCAMDAT data (i.e., *work* and *like*) and examine their word-level CEFR level distribution (predicted with Qwen) across essay-level CEFR levels, as shown in Figures 7 and 8. The captions report the number of essays containing these words for each essay-level CEFR level. Figures 7

	<i>work</i>	<i>like</i>
$\geq B2$	88.6	-
$\geq B1$	90.6	89.6
$\geq A2$	96.8	99.4

Table 5: EVP consistency in terms of Accuracy (%) for words *work* and *like* (EFCAMDAT).

and 8 are summarised in Table 5, where we assess the consistency of the EVP for these two words in terms of Accuracy. As discussed in Section 4.6, vocabulary use typically reflects a certain proficiency level or higher, rather than being exclusive to a specific level. Accordingly, we compute Accuracy

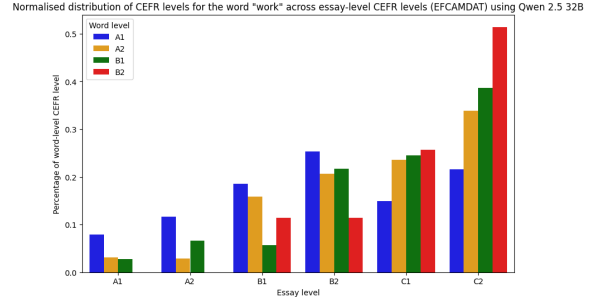


Figure 7: Distribution for word *work* (A1: 191; A2: 278; B1: 482; B2: 667; C1: 457; C2: 667).

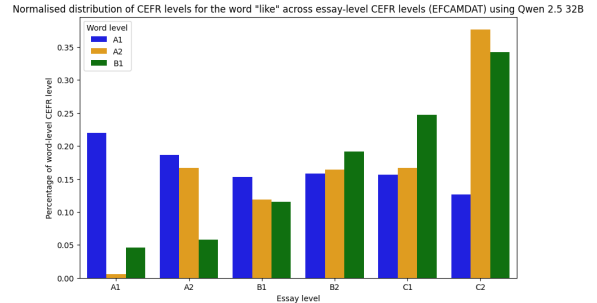


Figure 8: Distribution for word *like* (A1: 401 ; A2: 411; B1: 380; B2: 459; C1: 497; C2: 591).

by checking whether a word appears in essays at its assigned level or above. It is also important to note that we are comparing word-level *vocabulary* levels with essay-level *holistic* scores, where vocabulary represents only one component of the overall assessment. The results show high accuracy and suggest a strong degree of consistency in the CEFR classification provided by the EVP.

5 Conclusions and future work

In this work, we introduced a novel approach to in-context, word-level L2 vocabulary assessment by leveraging LLMs in combination with the English Vocabulary Profile. We compared the performance of several open-source and proprietary LLMs to a PoS-based model and showed that the former are particularly effective for this task, especially in handling polysemy and lexical ambiguity.

We plan to integrate this approach into an automatic essay grading system, where it could enrich holistic scoring with fine-grained feedback on vocabulary use, for example, by identifying lexical gaps relative to the learner’s proficiency or highlighting advanced vocabulary usage as a strength. Additionally, we plan to extend our experiments to spoken transcriptions in order to further evaluate the robustness of LLMs and assess their effectiveness across different modalities.

Limitations

One limitation of this study is the lack of a systematic investigation into how learner errors affect our approach. In particular, lemmatisation and matching with EVP entries may be hindered by such errors.¹⁴ While spelling mistakes can be addressed with a spellchecker, grammatical or lexical errors pose a greater challenge. In this case, it would be interesting to test our approach on pairs of original and grammatically corrected (manually and/or automatically) sentences or essays and analyse shifts in the LLM's probability distributions in the presence of learner errors.

Another limitation of our approach lies in the way word-level proficiency labels are assigned. In our experiments, each word in the OneStopEnglish data was annotated with its CEFR level, not its specific sense. As a result, polysemous words with multiple meanings but identical CEFR levels could be matched to an incorrect sense, even though the level remains technically accurate. However, it is important to note that the LLMs are prompted to identify the intended meaning of each word based on its context. The CEFR level is then assigned post hoc by mapping that selected meaning to the corresponding entry in the English Vocabulary Profile, hence to a CEFR level.

Finally, due to space constraints, we did not conduct a focused analysis on multi-word expressions such as idioms and phrasal verbs. Nonetheless, given the strong overall results and the fact that multi-word expressions are included in our data, it is reasonable to assume that our approach also performs well on these cases.

Acknowledgments

This paper reports on research supported by Cambridge University Press & Assessment, a department of The Chancellor, Masters, and Scholars of the University of Cambridge. The authors would like to thank the ALTA Spoken Language Processing Technology Project Team for general discussions and contributions to the evaluation infrastructure.

¹⁴In our first experiment specifically, learner example sentences from the EVP may contain errors, but not involving the target word. Additionally, this experiment does not require lemmatisation.

References

- Desislava Aleksandrova and Vincent Pouliot. 2023. [CEFR-based contextual lexical complexity classifier in English and French](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 518–527, Toronto, Canada. Association for Computational Linguistics.
- David Alfter, Yuri Bizzoni, Anders Agebjörn, Elena Volodina, and Ildikó Pilán. 2016. [From distributions to labels: A lexical proficiency analysis using learner corpora](#). In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, pages 1–7, Umeå, Sweden. LiU Electronic Press.
- David Alfter and Elena Volodina. 2018. [Towards single word lexical complexity prediction](#). In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 79–88, New Orleans, Louisiana. Association for Computational Linguistics.
- Richard C. Anderson and Peter Freebody. 1981. Vocabulary knowledge. In J. P. Guthrie, editor, *Comprehension and teaching: Research reviews*, pages 77–117. International Reading Association, Newark.
- Richard C. Anderson and Peter Freebody. 1983. Reading comprehension and the assessment and acquisition of word knowledge. In B. Hutson, editor, *Advances in reading / language research*, volume 2. JAI Press, Greenwich.
- Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwar. 2022. [CEFR-based sentence difficulty annotation and assessment](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Melissa M. Baese-Berk, Shiloh Drake, Kurtis Foster, Dae-yong Lee, Cecelia Staggs, and Jonathan M. Wright. 2021. Lexical diversity, lexical sophistication, and predictability for speech in multiple listening conditions. *Frontiers in psychology*, 12:661415.
- Stefano Bannò, Hari Krishna Vydana, Kate Knill, and Mark Gales. 2024. [Can GPT-4 do L2 analytic assessment?](#) In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 149–164, Mexico City, Mexico. Association for Computational Linguistics.
- Stephen Bax. 2012. [Text inspector](#). *Online text analysis tool*.
- Annette Capel. 2015. The English Vocabulary Profile. In J. Harrison and F. Barker, editors, *English Profile in Practice*. Cambridge University Press, Cambridge.

- Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, Cambridge.
- Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment - Companion volume*. Council of Europe, Strasbourg.
- Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. 2023. The English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) Corpus. *International Journal of Learner Corpus Research*, 9(2):248–269.
- Scott A Crossley and Danielle S McNamara. 2012. Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2):115–135.
- Luise Dürlich and Thomas François. 2018. [EFLLex: A graded lexical resource for learners of English as a foreign language](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Núria Gala, Thomas François, Delphine Bernhard, and Cédric Fairon. 2014. Un modèle pour prédire la complexité lexicale et graduer les mots. In *TALN'2014*, pages 91–102.
- Aina Garí Soler and Marianna Apidianaki. 2021. Let's play monopoly: BERT can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.
- Atta Gebril and Lia Plakans. 2016. [Source-based tasks in academic writing assessment: Lexical diversity, textual borrowing and proficiency](#). *Journal of English for Academic Purposes*, 24:78–88.
- Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2013. [Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database \(EFCAMDAT\)](#). In *Proceedings of the 31st Second Language Research Forum*, pages 240–254, Somerville. Cascadilla Proceedings Project.
- Liz Hamp-Lyons. 1995. [Rating nonnative writing: The trouble with holistic scoring](#). *TESOL Quarterly*, 29(4):759–762.
- Luke Harding. 2014. [Communicative language testing: Current issues and future research](#). *Language Assessment Quarterly*, 11(2):186–197.
- Yan Huang, Jeroen Geertzen, Rachel Baker, Anna Korhonen, and Theodora Alexopoulou. 2017. The EF Cambridge Open Language Database (EFCAMDAT): Information for users.
- Shi Huawei and Vahid Aryadoust. 2023. A systematic review of automated writing evaluation systems. *Education and Information Technologies*, 28(1):771–795.
- Kristopher Kyle, Scott Crossley, and Cynthia Berger. 2018. The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior research methods*, 50(3):1030–1046.
- Kristopher Kyle and Scott A Crossley. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4):757–786.
- Batia Laufer and Paul Nation. 1995. Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3):307–322.
- Paul Lennon. 2008. Contrastive analysis, error analysis, interlanguage. In S. Gramley and V. Gramley, editors, *Bielefeld Introduction to Applied Linguistics. A Course Book. Bielefeld*, pages 51–60. Aisthesis, Bielefeld.
- Agnieszka Leńko-Szymańska. 2015. The English Vocabulary Profile as a benchmark for assigning levels to learner corpus data. In M. Callies and S. Götz, editors, *Learner Corpora in Language Testing and Assessment*. John Benjamins, Amsterdam; Philadelphia.
- Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. [LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.
- Llama Team. 2024. [The Llama 3 herd of models](#).
- Xiaofei Lu. 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.
- Paul Meara and Barbara Buxton. 1987. An alternative to multiple choice vocabulary tests. *Language Testing*, 4(2):142–154.
- Paul Nation. 1983. Testing and teaching vocabulary. *Guidelines*, 5(1):12–25.
- Diane Nicholls. 2003. The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 Conference*, pages 572–581.
- OpenAI. 2023. [GPT-4 Technical Report](#). *Preprint*, arXiv:2303.08774.
- John Read. 2000. *Assessing vocabulary*. Cambridge University Press, Cambridge.
- John Read. 2013. Second language vocabulary assessment. *Language Teaching*, 46(1):41–52.

- Norbert Schmitt, Diane Schmitt, and Caroline Clapham. 2001. Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, 18(1):55–88.
- Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for computational Linguistics*, 8:247–263.
- Yukio Tono. 2013. The CEFR-J handbook: A resource book for using CAN-DO descriptors for English language teaching. *Tokyo: Taishukan*.
- Jeanine Treffers-Daller, Patrick Parslow, and Shirley Williams. 2018. Back to basics: How measures of lexical diversity can help discriminate between cefr levels. *Applied Linguistics*, 39(3):302–327.
- Satoru Uchida and Masashi Negishi. 2018. Assigning CEFR-J levels to English texts based on textual features. In *Proceedings of Asia Pacific Corpus Linguistics Conference*, volume 4, pages 463–467.
- Sowmya Vajjala and Ivana Lučić. 2018. [OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification](#). pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.
- Anne Vermeer. 2001. Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, 22(2):217–234.
- Sara Cushing Weigle. 2002. *Assessing Writing*. Cambridge Language Assessment. Cambridge University Press.
- Kelly Woods, Brett Hashimoto, and Earl K. Brown. 2023. [A multi-measure approach for lexical diversity in writing assessments: Considerations in measurement and timing](#). *Assessing Writing*, 55:100688.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.
- Su-Youn Yoon, Suma Bhat, and Klaus Zechner. 2012. Vocabulary profile as a measure of vocabulary sophistication. In *Proceedings of the seventh workshop on building educational applications using NLP*, pages 180–189.
- Guoxing Yu. 2010. Lexical diversity in writing and speaking task performances. *Applied Linguistics*, 31(2):236–259.

A Appendix A: Other stats

Table 6 shows the degree of polysemy in the EVP.

Table 7 shows the scores distribution in the selected ELLIPSE essays. Note that, in the original

No. of options	count (%)
1	56.96
2	21.13
3	8.82
4	3.99
5	2.59
6	1.53
> 6	4.98

Table 6: Degree of polysemy in the EVP.

corpus, higher (i.e., 4.5 and 5) and lower (i.e., 1 and 1.5) proficiency levels are underrepresented compared to intermediate levels.

Overall Score	# Essays	
	Train	Test
5	28	12
4.5	13	7
4	87	43
3.5	39	23
3	61	27
2.5	62	31
2	38	19
1.5	22	11
1	9	2

Table 7: Distribution of ELLIPSE essays by overall score in the considered training and test sets.

Table 8 shows the Correlation between ground truth analytic scores and ground truth overall scores in terms of SRC for the ELLIPSE test split considered in this work.

B Appendix B: OneStopEnglish examples

The following is an example of three parallel sentences across the three proficiency levels drawn from OneStopEnglish:

Elementary: *They work with the police, the probation service and other, voluntary organizations to help members of the violent criminal gangs of London.*

Intermediate: *They work with the police, the probation service and other, voluntary organizations to help people who feel trapped and frightened in the violent criminal gangs of London.*

Advanced: *They work with the police, the probation service and other, voluntary organizations to help those who feel trapped and frightened in the*

	Overall	Vocabulary	Phraseology	Cohesion	Grammar	Syntax	Conventions
Overall	1.000	0.869	0.897	0.880	0.878	0.911	0.873
Vocabulary	0.869	1.000	0.863	0.776	0.776	0.798	0.748
Phraseology	0.897	0.863	1.000	0.828	0.835	0.847	0.811
Cohesion	0.880	0.776	0.828	1.000	0.781	0.825	0.816
Grammar	0.878	0.776	0.835	0.781	1.000	0.840	0.794
Syntax	0.911	0.798	0.847	0.825	0.840	1.000	0.816
Conventions	0.873	0.748	0.811	0.816	0.794	0.816	1.000

Table 8: Correlation between ground truth analytic scores and ground truth overall scores in terms of SRC.

violent criminal gangs that operate across London.

C Appendix C: Prompts

Prompt for semantic understanding

Read this sentence: **[EVP SENTENCE]**
Choose the correct meaning of **[WORD]**
by selecting the most suitable among the
following options A, B, C, D, E, or F. No
other answer is allowed. Only output the
respective option letter without any addi-
tional comments, notes, or explanations.

A) **[DEFINITION A]**

B) **[DEFINITION B]**

C) **[DEFINITION C]**

D) **[DEFINITION D]**

E) **[DEFINITION E]**

F) **[DEFINITION F]**

Example

Read this sentence: **“It was tough on the worn out employees.”** Choose the correct meaning of **“tough”** by selecting the most suitable among the following options A, B, C, D, E, or F. No other answer is allowed. Only output the respective option letter without any additional comments, notes, or explanations.

A) not easy to break or damage

B) describes food that is difficult to cut or eat

C) Tough people are mentally strong and not afraid of difficult situations.

D) difficult

E) Tough rules are severe.

F) unfair or unlucky

Prompt for word-level CEFR prediction

Read this L2 learner sentence: **[SENTENCE]**

Choose the correct meaning of **[WORD]**
(in square brackets) by selecting the most
suitable among the following options.
Also consider the additional information
and the PoS of each option. No other
answer is allowed. Only output the re-
spective option number without any addi-
tional comments, notes, or explanations.

1. **[DEFINITION 1]** - Additional infor-
mation: **[INFO]** (PoS)

2. **[DEFINITION 2]** - Additional infor-
mation: **[INFO]** (PoS)

3. **[DEFINITION 3]** - Additional infor-
mation: **[INFO]** (PoS)

[...]

n. None of the other options.

where the additional information consists of a brief definition of the word. In round brackets, we also feed the information related to the manually assigned PoS as described in the EVP. See the example below for further information.

The prompt for the experiments conducted on essay-level proficiency prediction is the same. The only difference is that “sentence” is replaced with “essay”.

Example

Read this L2 learner sentence: **They [work] with the police , the probation service and other , voluntary organizations to help those who feel trapped and frightened in the violent criminal gangs that operate across London .**

Choose the correct meaning of **“work”**
(in square brackets) by selecting the most
suitable among the following options.

Also consider the additional information and the part of speech of each option. No other answer is allowed. Only output the respective option number without any additional comments, notes, or explanations.

1) the place where you go to do your job - Additional information: work (PLACE) (Part of speech: noun)

2) something you do as a job to earn money - Additional information: work (JOB) (Part of speech: noun)

3) to do a job, especially the job you do to earn money - Additional information: work (DO JOB) (Part of speech: verb)

4) the activities that you have to do at school, for your job, etc. - Additional information: work (ACTIVITY) (Part of speech: noun)

5) If a machine or piece of equipment works, it is not broken. - Additional information: work (OPERATE) (Part of speech: verb)

6) when you use physical or mental effort to do something - Additional information: work (EFFORT) (Part of speech: noun)

7) If something works, it is effective or successful. - Additional information: work (SUCCEED) (Part of speech: verb)

8) to exercise in order to improve the strength or appearance of your body - Additional information: work out (EXERCISE) (Part of speech: verb)

9) a painting, book, piece of music, etc. - Additional information: work (CREATION) (Part of speech: noun)

10) to try hard to achieve something - Additional information: work at sth (Part of speech: verb)

11) to spend time repairing or improving something - Additional information: work on sth (Part of speech: verb)

12) to do a calculation to get an answer to a mathematical question - Additional information: work sth out or work out sth (Part of speech: verb)

13) If a problem or a complicated situation works out, it ends in a success-

ful way. - Additional information: work out (BECOME BETTER) (Part of speech: verb)

14) to know how to use a machine or piece of equipment - Additional information: can work sth; know how to work sth (Part of speech: verb)

15) to understand something or to find the answer to something by thinking about it - Additional information: work sth out or work out sth (UNDERSTAND) (Part of speech: verb)

16) None of the other options

D Appendix D: Other results

Table 9 reports the breakdown by word-level CEFR level in terms of F_1 score. For ambiguous cases, the PoS-based model performs reasonably well only at the A1 level. However, this result is partly influenced by the rule we set, i.e., assigning the lowest available CEFR level when multiple options are present (see Section 4.2), which inherently favours lower-level classifications. For both the N/A and C2 levels, the model yields an F_1 score of 0, highlighting its complete inability to handle these cases. For non-ambiguous cases, the PoS-based achieves the best performance on all the CEFR levels for the reasons explained in Section 4.5, with the exception of the C2 level. For this specific CEFR level, the PoS-based model obtains a high Recall (99.28%) but a low Precision (73.02%), while Qwen shows more balanced results, with a Recall of 90.65% and a Precision of 89.36%.

	Ambiguous		Non-ambiguous		All	
	Qwen 2.5 32B	PoS	Qwen 2.5 32B	PoS	Qwen 2.5 32B	PoS
N/A	50.0	0.0	94.3	94.4	92.0	91.6
A1	84.4	83.1	91.3	94.0	86.5	85.8
A2	82.1	67.4	89.8	91.8	85.1	75.6
B1	80.6	56.7	94.3	95.1	85.7	72.9
B2	78.5	49.3	90.5	90.8	84.2	72.0
C1	75.5	20.0	91.9	92.3	82.9	61.5
C2	67.2	0.0	90.0	84.2	78.9	59.7

Table 9: Breakdown of classification results by word-level CEFR level in terms of F_1 (OneStopEnglish).