From Prompting to Fine-Tuning: LLM Strategies for CEFR-Based Readability Assessment

De prompting a fine-tuning: estrategias para clasificación de legibilidad basada en CEFR

Elisabet Comelles Pujadas[1], **Laura Alonso Alemany**[2], **Juan Cruz Oviedo Ferreyra**[2]
[1]Universitat de Barcelona, [2]Universidad Nacional de Córdoba
elicomelles@ub.edu, lauraalonsoalemany@unc.edu.ar, juan.oviedo.869@unc.edu.ar

Abstract: This paper explores the use of generative open-source large language models (LLMs) to identify the CEFR level of texts for learners of English, a critical step in readability-controlled text adaptation for pedagogical purposes. To enable a systematic and cost-effective evaluation, we assess model performance on a manually annotated CEFR classification dataset, rather than relying on the more complex and less stable evaluation of generated adaptations. We evaluate several prompting strategies, including prompts enriched with CEFR descriptors or vocabulary lists, across multiple LLM families, and compare their performance to traditional readability tools and proprietary LLMs. Results show that prompting alone yields moderate performance and high variability, with models below 8B parameters behaving inconsistently, while fine-tuning markedly improves classification accuracy. Our experiments also show the importance of balanced, manually annotated training data, particularly for low-resource CEFR levels such as A1.
Keywords: CEFR classification, Automatic Readability Assessment, Generative Large Language Models, Language Learning Technology.

Resumen: Este artículo investiga el uso de grandes modelos de lenguaje generativos y de código abierto (LLMs) para identificar el nivel de competencia de MCER de textos para estudiantes de inglés, un paso fundamental en la adaptación automática de textos con fines pedagógicos. Buscando una evaluación sistemática con los recursos disponibles, analizamos el rendimiento de los modelos sobre un conjunto de datos notados manualmente con los niveles del MCER, en lugar de recurrir a la evaluación de textos adaptados generados automáticamente, que es más compleja y menos estable. Evaluamos varias estrategias de prompting, incluidos prompts enriquecidos con descriptores del MCER o listas de vocabulario, en múltiples familias de LLMs, y comparamos los resultados con herramientas tradicionales de evaluación de la legibilidad de textos y modelos propietarios. Los resultados muestran que el prompting por sí solo produce un rendimiento moderado y una alta variabilidad, con modelos por debajo de 8B comportándose de manera inconsistente. El fine-tuning mejora notablemente la precisión de la clasificación.También se observa la importancia de contar con datos de entrenamiento equilibrados y anotados manualmente, especialmente para niveles con pocos recursos, como A1.
Palabras clave: Clasificación CEFR, Evaluación automática de legibilidad, LLMs generativos, Tecnología para el aprendizaje de idiomas.

1 Introduction and Motivation

Personalizing learning trajectories is a central goal in modern language education, as it allows learners to engage with materials that match both their proficiency level and individual interests. Automated generation of teaching materials offers an efficient way to achieve this personalization at scale, especially in diverse and heterogeneous learning contexts such as rural schools or adult education programs.

Within this scope, the *taskGen* project aims to support teachers of second languages by providing them with a tool that includes language technologies to design communicative tasks. A key component of this system is the automatic adaptation of texts to the different proficiency levels established by the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2020). By automatically adjusting the linguistic complexity of texts, the system enables teachers to provide each learner with suitable and engaging content, thus promoting more effective and inclusive language learning.

To achieve such automatic adaptation, it is first necessary to reliably determine the proficiency level associated with a given text. This task falls within the domain of Automatic Readability Assessment (ARA), which estimates the linguistic complexity of texts to align them with the skills of target learners (Vajjala, 2022). Although ARA has long been a field of research within educational NLP, advances in large language models (LLMs) offer new opportunities to improve CEFR-based systems.

In the context of adapting texts to a target CEFR level, readability classification is not an end in itself, but rather a necessary intermediate step. That is, the accurate identification of the level of a text enables more faithful and pedagogically-sound adaptations to a target CEFR level. Evaluating the actual adaptation of texts to a given level is a costly task that ideally requires input from human experts. When automated, adapted texts are often evaluated by comparing them with only one (or few) of many possible reference adaptations with a similarity metric (Alva-Manchego et al., 2025). This approach allows to reduce the cost of evaluating adaptations, but it arguably falls short to assess the adequacy of adaptations that differ from reference texts. In contrast, evaluating the ability of a model to identify the level of a text is a task that can be done automatically, resorting to manually labelled corpora. Thus, performance in level classification can serve as a proxy for adaptation quality, allowing the more resource-intensive evaluation of adapted texts to be reserved for the most promising approaches.

With this motivation, we explore the ability of generative LLMs to identify CEFR levels in texts for learners of English. Our main research questions are:

1. *Can LLMs reliably identify CEFR levels?*
2. *Do generative LLMs (decoder-only) have a level of performance comparable to classifier LLMs (encoder-only) for CEFR level detection?*
3. *Does linguistic knowledge impact on the performance of classifiers?*
4. *How does labelled data impact the performance of LLMs to classify texts into CEFR levels?*

The rest of the paper is organized as follows. In the next section, we provide an overview of relevant work concerning CEFR and ARA. Then, we introduce our LLM-based approach to identify CEFR levels in texts for learners of English. Later, we describe our experimental setup: the dataset, the metrics and the different experiments that we carried out. Next, we report some quantitative and qualitative analysis of results, including a comparison with state-of-the-art approaches and off-the-shelf tools. We conclude our paper with some remarks on the ability of generative LLMs to identify CEFR levels, the impact of fine-tuning and future work.

2 Related Work

2.1 The Common European Framework of Reference

The Common European Framework of Reference (CEFR) (Council of Europe, 2020) is an international standard that aims to organise language proficiency in any language into 6 proficiency levels that range from A1 (beginner) to C2 (mastery). The levels are structured into three broad categories, that is Basic User (A1-A2), Independent User (B1-B2) and Proficient User (C1-C2), which are defined through 'can-do' descriptors that can be applied to reading, writing, speaking and listening skills. Although the CEFR is used across Europe and has extended globally, one of its main drawbacks is the vagueness of its descriptors and the need for fine-tuning them (Díez-Bedmar and Luque-Agulló, 2023; Simons and Colpaert, 2015).

Distinguishing between far-apart levels (e.g. between A2 and C1) is easy; however, when it comes to closer levels (e.g. between

B1 and B2), distinguishing whether a text falls within one or the other is not always an easy task. Such a difficulty has a direct effect on both human and automatic assessment of texts.

2.2 Readability Assessment

ARA refers to "the task of modeling the reading and comprehension difficulty of a given piece of text, for a given target audience" (Vajjala, 2022). It is an interdisciplinary field where researchers from NLP, education and psychology interact. Readability and ARA have been widely studied (Vajjala, 2015; Meng et al., 2020; Lee and Vajjala, 2022; Deutsch, Jasbi, and Shieber, 2020). Most of the research on ARA has focused on predicting reading difficulty for native speakers, although since 2007, several efforts in developing ARA tools that take L2 learners into account have emerged (Heilman et al., 2007; Vajjala and Meurers, 2012). Thus, one of the applications of ARA is deciding whether a text is appropriate for a specific L2 level.

ARA has been approached from different perspectives. Some of the most common have been readability formulae, which try to measure a text readability through simple measures, such as number/length of words/sentences in a text, percentage of difficult words, etc. Some of these traditional formulae are Flesch-Kincaid Grade level (Kincaid et al., 1975), SMOG Grading (McLaughlin, 1969) and Coleman-Liau Index (Coleman and Liau, 1975), among others.

In the last two decades, NLP researchers started taking an interest in this task and new methods have been used since then. These methods range from statistical language models to machine learning approaches based on feature engineering (Hancke, Vajjala, and Meurers, 2012; Vajjala and Meurers, 2012; Chen and Meurers, 2018; Deutsch, Jasbi, and Shieber, 2020; Lee, Jang, and Lee, 2021; Lee and Vajjala, 2022). Some of the features used are sentence length, word lists, morphological variation and complexity, coreference resolution, discourse relations or word embeddings.

The methods above, however, rely on the availability of well-labelled data annotated with the readability levels for L2 learners. In fact, the need of such data has become even more urgent with the advent of generative AI, which can be an excellent tool to adapt a text and make it simpler or more complex, or to generate text that matches a user's proficiency level (Alva-Manchego et al., 2025; Tran et al., 2025; Huang, Wei, and Huang, 2024; Farajidizaji, Raina, and Gales, 2024; Ribeiro, Bansal, and Dreyer, 2023). A natural first step towards enabling these applications is to explore how well LLMs can perform ARA, a question that has received limited attention to date (Benedetto et al., 2025; Malik et al., 2024).

The TSAR 2025 shared task on Readability-Controlled Text Simplification (Alva-Manchego et al., 2025) included identification of the CEFR level of the automatically adapted texts via a ModernBert-based classifier, the TSAR CEFR Evaluator Model. However, this was a classifier model, an encoder-only LLM, independent from the adaptation task. In contrast, the main objective of our work is to assess the ability of generative LLMs to identify CEFR levels, as a first approach to assess their general ability for readability-controlled adaptation, and as a way to identify successful approaches to improve these models for the adaptation task.

3 Decoder-only LLMs to identify the CEFR level of texts

In this paper, we explore the ability of generative decoder-only LLMs to identify the CEFR level of texts. This focus is motivated by our broader objective of assessing the potential of such models for the automatic adaptation of texts to specific CEFR levels (Oviedo et al., 2025). In this context, examining their performance on classification tasks provides an indirect yet informative proxy for their adaptation capabilities, as direct evaluation of text adaptation remains methodologically challenging and resource-intensive. In contrast, classification offers a more manageable and cost-effective evaluation framework. Furthermore, this approach enables us to systematically analyze the impact of different strategies aimed at improving the CEFR-related performance of generative models.

We assess open-weight models that can be fine-tuned and we evaluate both prompting and fine-tuning approaches.

3.1 Families of decoder-only LLMs

Our aim is for ARA to be fully integrated into an independent platform to help EFL teachers design communicative tasks (Castellví Vives, Gilabert Guerrero, and Comelles Pujadas, 2024); thus, our experiments focused on open-source language models.

Different versions of three families of LLMs were explored: Llama (3.1 8B, 3.2 3B, 3.2 11B), Mistral (8B and 12B) and Gemma2 (9B). Finally, after some experimental results, we kept Llama 3.1 8B as most promising, and, due to hardware restrictions, we experimented on fine-tuning this model only.

3.2 Prompting approaches

Assuming that generalistic LLMs already have some knowledge about CEFR, extensive prompting seemed the simplest and easiest way to start our set of experiments. Several ways to build prompts for proficiency level identification were explored. Each approach builds up in complexity by adding more useful information on the CEFR levels. Prompts can be found in Appendix A.

3.2.1 Baseline

LLMs already have some information regarding CEFR, thus the most straightforward way to start prompting is by asking the LLM to identify the CEFR level of a text, so that it can draw on the knowledge that it already has.

We used 9 different prompts (see prompts 0 to 8 in Appendix A). In all of them we specified that the text had to be classified according to the CEFR levels and that only the CEFR level had to be provided, no introduction or explanation was required.

In addition, some prompts included that the text was aimed to be used as a reading task, whereas others also specified that the person making the query was an EFL teacher.

3.2.2 In-context learning

A very common technique in prompt engineering is the use of examples (Malik et al., 2024) to improve the LLMs' performance. To this aim, we used the baseline prompts in combination with 1 or 2 examples of each level, as shown in prompts 9 to 20 in Appendix A. The examples were taken from the training partition dataset described in Section 4.1, ensuring that the examples did not belong to the holdout (test) partition.

In preliminary explorations, we found that LLMs had a tendency to overpredict the B2 level. To address this bias, we created some prompts where examples were provided for all levels except B2 (prompts 13, 14, 19 and 20).

Some of the prompts with examples could get very long, which is known to impact the performance of LLMs. In some cases, the longest prompts reached the limit of tokens that could be processed by a given LLM, thus causing the approach to fail.

3.2.3 Including CEFR descriptors

As already mentioned, LLMs already have some knowledge of what the CEFR is. However, we wanted to test whether a reminder of the CEFR descriptors for Reading Comprehension could help improve their performance. Thus, we instructed the LLM to classify the given text using the CEFR Reading Descriptors in (Council of Europe, 2020) (see Appendix A for the corresponding prompts, 21 and 22).

To address the bias towards B2, we applied prompt 21, omitting B2 descriptors to check whether that resulted in any improvement (prompt 22).

3.2.4 Including CEFR vocabulary

English is a highly lexical language, that is, the meaning of a sentence relies heavily on the meaning of the words within it, not just on how the words are arranged. In addition, several studies have shown that the size and mastering of vocabulary is a clear indicator of a student's CEFR level (Milton and Alexiou, 2009; Milton and Alexiou, 2020). Therefore, we tried adding to the prompt a list of 8000 words organised by CEFR level as provided by (Oxford University Press, n.d.), instead of the CEFR descriptors (prompt 23).

3.3 Fine-tuning approach

Fine-tuning is also another strategy that has proved effective to improve the performance of LLMs to identify the CEFR level of texts and adapt them to the corresponding level (Malik et al., 2024).

After assessing the performance of different LLMs with prompting using the training partition, we selected Llama 3.1 8B (Meta, 2024) to fine-tune, because it showed the best tradeoff between performance and

size of the model, and provided consistent results. We fine-tuned it on the training partition of the manually labelled corpus described in 4.1, and used the validation partition for hyperparameter search. We explored the impact of training with CERD+A1 dataset only, with EDIA dataset only, and with both of them.

Fine-tuning was implemented by adding a 6-neuron lineal layer to the model, with each neuron corresponding to a CEFR class. The input was a text from the training partition of the manually annotated dataset, and the expected output was the CEFR level for that text. Partial fine-tuning was applied, freezing all model parameters except the classification layer. This strategy allowed to significantly reduce computational costs of fine-tuning, to avoid overfitting and to preserve the base linguistic knowledge previously encoded in the model. The hyperparameters for tuning can be seen in Appendix B.

We conducted experiments adding specific instructions to match the numeric fine-tuning classification levels (0, 1, 2, 3, 4 and 5) with the CEFR levels (A1, A2, B1, B2, C1 and C2). To this aim, we added the sentence "*Keep in mind that the label 0 is A1, label 1 is A2, label 2 is B1, label 3 is B2, label 4 is C1, label 5 is C2.*" to each of the prompts in Appendix A[1]. Additionally, we also performed experiments without adding this specific instruction.

Finally, the fine-tuned model was also fed the texts without an engineered prompt, to assess the ability of the model to do the task by virtue of the information acquired via fine-tuning only.

4 Experimental setup

4.1 Manually labelled dataset

A corpus of texts for learners of English that had been manually classified according to their CEFR level was used to carry out the current study.

The first dataset we used was the Cambridge English Readability Dataset (CERD) (Xia, Kochmar, and Briscoe, 2016), which consists of reading passages from past exams Readings Tasks ranging from A2 to C2. The dataset includes 331 texts, that is 68 texts per level approximately.

[1]The phrase "*Keep in mind*", seems to be taken as an idiom, and not literally referring to a mind, thus it does not seem to introduce a confusor

dataset	training	tuning	holdout
CERD+A1	316	39	41
EDIA	569	71	72
Combined	885	110	113

Table 1: Number of texts in the partitions of the manually labelled corpora for training, tuning (validation) and holdout (test).

To complement this dataset for level A1, we added 65 texts from the Kaggle CEFR levelled English texts dataset (Montgomerie, 2021), constituting the CERD+A1 corpus up to a total of 396 texts (see Table 1). Some of the texts in (Montgomerie, 2021)'s dataset were pre-assigned a CEFR level, whereas others were classified automatically with Text Inspector (Bax, 2025). Each of the 65 A1 added texts was revised manually by two experienced EFL teachers familiar with CEFR to confirm the level of the texts.

The second dataset used was the dataset created by EDIA (Breuker, 2023), which includes 712 texts manually classified with their CEFR level.

Finally, both datasets (i.e., CERD+A1 and EDIA) were combined to evaluate prompting strategies and to fine-tune the Llama 3.1 8B model. To do that, we created training, tuning (validation) and holdout (test) partitions, shown in Table 1.

The evaluation of both prompting and fine-tuning strategies was carried out in the holdout (test) partition of this dataset.

4.2 Metrics

To measure the performance of LLMs with greater nuance, we used strict accuracy and approximate accuracy, which accounts for near-miss predictions.

Strict accuracy determines the proportion of correct predictions, that is, if the prediction matches the manually assigned level exactly. However, this metric is too strict to account for the fuzzy nature of CEFR levels. Indeed, professional teachers of English state that some texts may show an overlapping of features between adjacent levels. To capture these intutions from practitioners, we used approximate accuracy to complement strict accuracy, representing how well the prediction approximates the manually assigned level, close in spirit to the RMSE metric used in the TSAR 2025 Shared Task (Alva-Manchego et al., 2025). Thus, if

a model's prediction is an exact match, it is given a score of 1, if the predicted level is one level above or below the manually-assigned level, it is given a score of 0.5. The rest of predictions are scored as fully incorrect with a score of 0. This metric reflects the relative severity of prediction errors, rewarding the model for approximate correctness, while penalizing larger deviations.

4.3 Comparison with state of the art

To evaluate the performance of our approaches, we compared their performance to those of other well-known systems: the CEFR Evaluator Model used in TSAR 2025 (Alva-Manchego et al., 2025), Text Inspector (Bax, 2025), CVLA (Uchida and Negishi, 2025) and ChatGPT. The holdout portion of our dataset was used and both accuracy and approximate accuracy were calculated.

The evaluator model used in TSAR 2025 was obtained by fine-tuning three different classifier LLMs (encoder-only) of the ModernBert-Base LLM (Warner et al., 2025) with three different portions of the Universal-CEFR dataset (Imperial et al., 2025), and then choosing the final level to predict as the majority vote or the most confident vote. We applied this Evaluator Model to the texts in the holdout portion of our dataset, and obtained both accuracy and approximate accuracy. It is important to note that the Universal-CEFR dataset contains the corpus that we used as holdout corpus to evaluate the performance of our different approaches. Therefore, the TSAR Evaluator Model is quite likely to perform exceedingly well in that corpus, as it may have memorized some of the particular features of the specific texts being evaluated.

Other widely-used analysers were evaluated for the sake of comparison. The first one was Text Inspector, one of the most popular text analysers among teachers. It uses metrics at token, type and syllable level to identify a text readability difficulty. The second one was CVLA, which assesses the readability difficulty of a text based on the level of its vocabulary. It shows 65.7% accuracy in predicting CEFR levels on their evaluation dataset. Finally, we also compared to ChatGPT5. The model was prompted with Prompt 15, which showed good results in non-fine-tuned models.

5 Analysis of results

This section presents the results of the experiments carried out with different LLMs, prompts and fine-tuned models, as well as the comparison of our best-performing model with other state-of-the-art systems.

5.1 Prompting and fine-tuned LLMs

The main takeaway of our results is that, as expected, fine-tuned models perform better than non-fine-tuned ones, especially when both datasets are used. Figure 1 shows the detail of performance across language models and settings. Models smaller than 8B perform worse, and more irregularly, than models of size 8B or larger, as illustrated by the dispersion of the points representing each prompting approach. As regards models larger than 8B, there is no significant improvement: Llama 3.1. 8B, Llama 3.2. 11B, Gemma 2 9B and Mistral 12B, all reach comparable performance. As for fine-tuned models, the best performance was achieved when using both datasets to train, and when prompts specified the relation between the CEFR level and the class name used for fine-tuning (*numeric label*). This strategy produced improvements in performance, especially for the model fine-tuned with larger datasets (EDIA or combined), which performed better than that fine-tuned with the smallest dataset (CERD+A1).

Regarding differences in prompting approaches, Figure 2 shows that no approach is overwhelmingly superior. Fine-tuned models, represented by clearer dots, outperform all combinations of prompts and non-fine-tuned models, even in the "without prompt" approach, where a minimal instruction is provided. In-context learning, and most specifically, the use of examples, results in some slight improvement in the performance of non-fine-tuned models, especially when two examples are used, that is, prompts 15 and 20, respectively. However, examples do not have a positive impact on the performance of fine-tuned models. The most successful prompt, number 1 performs better than any prompt with in-context examples. This can be attributed to the fact that any possible information provided by examples has already been encoded in the model by the fine-tuning process.

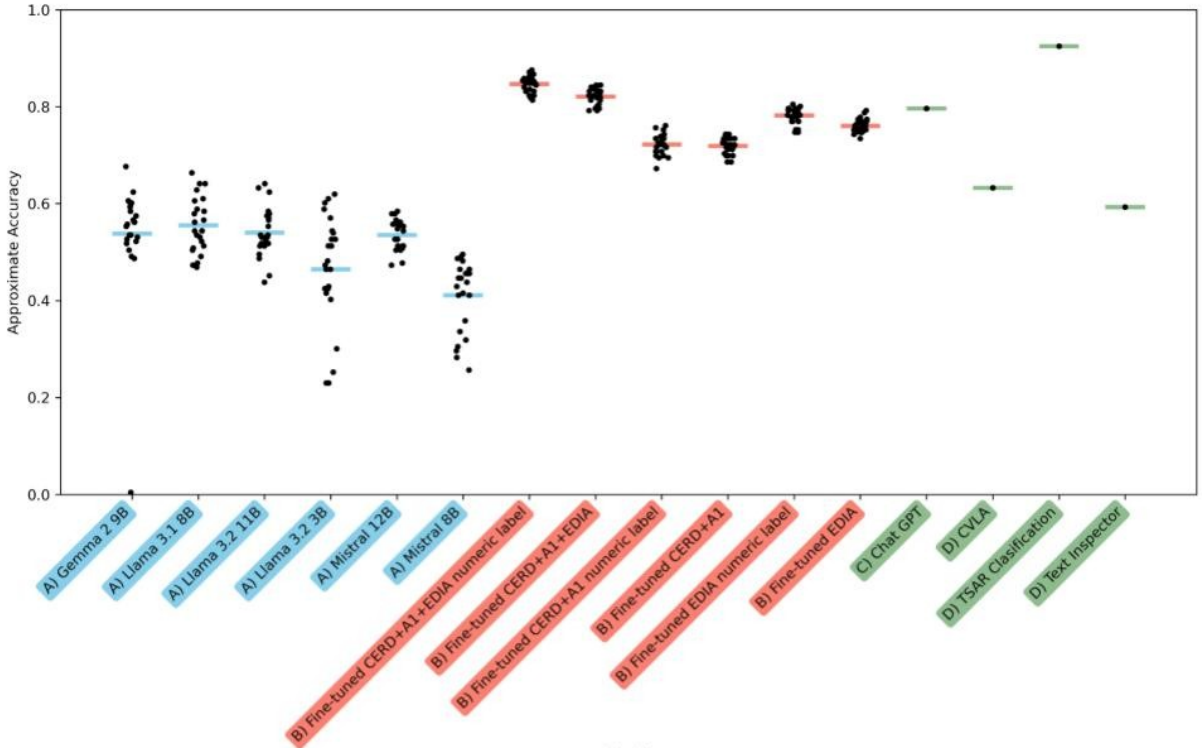

Figure 1: Approximate accuracy (y axis) obtained for different models to identify CEFR levels in texts, evaluated in the combined holdout corpus. Each point represents a different prompting approach, each language model or approach is represented in one of the columns in the x axis. From left to right: Non-fine-tuned models are in blue, fine-tuned models are in orange, and the state-of-the art classifiers are in green.

The length of the prompt seems to have the biggest negative impact on the performance of the model, which is the reason why some prompting approaches with examples perform so poorly. This may also explain why the use of linguistic information, that is, the use of CEFR descriptors and CEFR-classified wordlists, does not seem to have a positive effect either. Prompt 23, which includes a list of 8000 words organised by CEFR levels, shows the worst performance of all prompts with linguistic information. However, Prompt 22, which includes all CEFR levels overall reading descriptors, achieves much better results.

5.2 Comparison with state-of-the-art systems

The comparison between our models and state-of-the art systems shows that the fine-tuned Llama 3.1 8B model, trained on the combination of datasets (CERD+A1 and EDIA) and prompted with Prompt 1 performs closer to the TSAR model, as shown in Table 2. It should be noted, though, that as discussed above, the TSAR classifier was trained using the holdout portion of our dataset, which likely contributes to its excellent results.

System	Acc	Approx Acc
Fine-tuned Llama	77%	87.6%
TSAR	**85%**	**92.5%**
CVLA	43.4%	63.3%
Text Inspector	37.2%	59.3%
ChatGPT	61.1%	79.6%

Table 2: Accuracy and Approximate Accuracy for state-of-the-art systems for CEFR classification.

In contrast, the rest of systems, namely Text Inspector, CVLA, and ChatGPT, perform substantially worse, with only ChatGPT reaching above 50% accuracy. These results are particularly striking in the case of Text Inspector, whose performance in terms of strict accuracy is notably low (37.2%). When considering approximate accuracy, results improve remarkably across all systems, with most

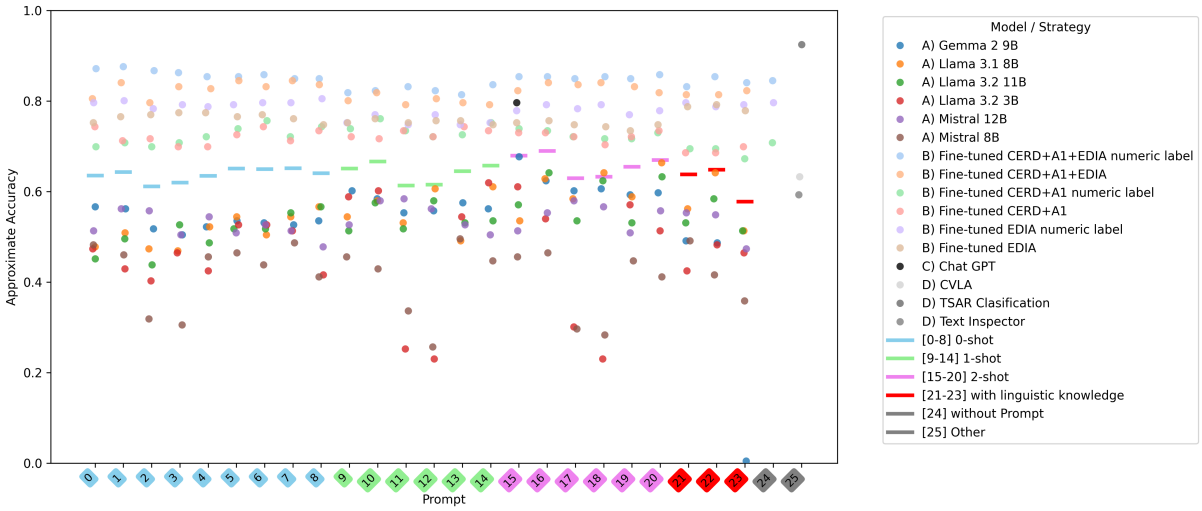

Figure 2: Approximate accuracy (y axis) for different prompts (x axis) with different language models and systems. Prompts in blue (0 to 8) are zero-shot prompts. Prompts in green (9 to 14) are one-shot, prompts in pink (15 to 20) are two-shot, prompts in red (21 to 23) incorporate extensive descriptions of CEFR level descriptors and vocabulary lists, and the no-prompt approach for fine-tuned models (named *Without Prompt*) and the state-of-the-art classifiers (TSAR classifier, Text Inspector, CVLA and ChatGPT) are in grey, to the right. Clearer dots correspond to fine-tuned language models. The average performance for each prompting strategy across language models is marked with a line.

models reaching 60% accuracy or higher. ChatGPT gets up to almost 80%, whereas, CVLA and Text Inspector, both of which rely heavily on traditional readability metrics and vocabulary-based classification, show the worst performance, achieving 63.3% and 59.3% approximate accuracy, respectively. Overall, these findings suggest that while lexical features might be a strong indicator of a text level, other factors like morphosyntactic features, types of clauses or connective devices must also be factored in for more accurate CEFR level classification.

Next, the confusion matrices of the 5 systems are analysed in detail (see Figure 3 in Appendix). The top-right matrix in Figure 3 shows the error patterns for the fine-tuned Llama model with Prompt 1. The lowest class (A1) exhibits the highest proportion of error (33%), which may be attributed to the limited availability of A1 examples in CEFR-labelled corpora like CERD, as well as in general corpora. This may result in LLMs failing to adequately model this level. As for the remaining levels, most misclassification occur with adjacent levels, with only a few exceptions, namely two B2 texts classified as C2 and one C1 text classified as B1. Section 5.3 provides a detailed analysis of these cases.

The top-left matrix in Figure 3 corresponds to the confusion matrix for the TSAR CEFR Evaluation Model. Most texts are correctly classified, and misclassifications are typically limited to adjacent levels. It is worth noting that the model never predicts level A1, possibly due to the lack of representation of this level in the fine-tuning corpus. However, the model excels at classifying higher proficiency levels.

In contrast to the TSAR model, CVLA (middle-right matrix in Figure 3) correctly classifies all A1 texts. However, misclassifications tend to span more than one level, unlike both TSAR and our fine-tuned model. That is the case of a C1 text classified as A1, which coincides with the text that our system classified as B1. This finding reinforces the claim that lexical features alone are insufficient for accurate CEFR classification. Finally, CVLA shows a tendency to classify A2 texts as A1.

Similar to TSAR, Text Inspector (bottom matrix in Figure 3) fails to correctly classify A1 texts. It should also be noted that the number of A1 texts evaluated is smaller, as the system cannot process texts shorter than 100 words. As for A2 texts, they are generally classified correctly, and misclassifications are

typically limited to adjacent levels. A similar pattern is observed across the rest of levels, although some deviations occur, such as one C1 text classified as A2, and two C2 texts classified as A2. It is also worth highlighting that the system appears to struggle with B1 texts particularly, 44% of which are misclassified as A2.

Finally, in line with our fine-tuned model, ChatGPT's misclassifications (middle-left matrix in Figure 3) are generally limited to adjacent levels, with the only exception of one B2 text classified as C2. The highest proportion of errors is observed in the B1 (48%) and A2 (42%) levels, whereas the model performs comparatively well at classifying B2 texts. This pattern may be explained by the greater availability and representativeness of B2 texts in both online resources and training datasets, which could facilitate robust modelling at this level.

5.3 Error analysis

A detailed analysis of the three worst classification errors of our system was carried out, as well as additional human evaluation to verify the original levels. The evaluators were two experienced EFL teachers familiar with CEFR standards.

The first B2 text (CERD+A1 corpus) was consistently rated as B2 by both human evaluators, thus confirming the original classification. Although relatively long (596 words) and containing some advanced vocabulary and idioms (e.g. *bits of*, *felt like a bit of*, *seemed so unlikely*), as well as some phrasal verbs (e.g. *took off*, *turned out*, *put forward an argument*), the text remains accessible to B2 learners due to contextual support and a clear chronological structure. It does not seem to be cognitively demanding beyond B2. Therefore, we attribute its misclassification as C2 to its length and relatively advanced lexical items.

The second B2 text (EDIA corpus) is about the Magna Carta. The text is syntactically simple, relying mainly on present and past tense forms, thus it does not seem particularly challenging from a syntactic perspective. However, it includes a vast amount of highly specialised vocabulary (e.g. *oppressed serfs*, *Anglo-Norman barons*, *habeas corpus*), as well as many abstract and historical concepts. Notably, the evaluators themselves judged this text as C1, differing from the original classification. This suggests that the presence of domain-specific vocabulary may have significantly influenced the system's higher level classification.

The third case involves a C1 text (EDIA corpus) classified as B1. The text is relatively short (286 words) for a C1 text and takes the form of a dialogue with direct speech, very short sentences and mostly simple verb forms (e.g. Present Simple and *will* or *going to* for future reference). Most of the vocabulary is highly frequent and specific, with a few exceptions (e.g. *"greedy losers*, *contest the will*, *lawsuit* or *hope for a settlement*), which may fall within the B2-C1 range. In contrast to the original classification, the evaluators rated this text as B2, noting that while the linguistic surface is simple, some degree of inference and pragmatic interpretation is required. The system, however, seems unable to capture such nuances. Consequently, features such as text length, sentence simplicity and high-frequency vocabulary may have contributed to the lower classification.

All in all, the errors produced by our best performing model are relatively minor, as most involve adjacent-level misclassifications. Errors are concentrated in the B1 and B2 levels, which likely reflects the overlapping of linguistic features between neighbouring CEFR bands at these levels.

6 Discussion

Our experiments show that generative LLMs can reliably identify the CEFR level of a text, especially when fine-tuned. Fine-tuned models consistently outperform all prompting strategies, even those including CEFR descriptors or CEFR-level vocabulary lists, which have historically been the basis of traditional commercial systems for CEFR classification. While in-context examples slighlty improve the performance of some models, these gains never exceed those obtained through fine-tuning with a relatively small amount of labelled data. This confirms the importance of task-specific adaptation, even for models that exhibit broad linguistic knowledge.

The quantity and quality of labelled data strongly influences performance. Fine-tuning on the larger and more diverse EDIA dataset yields higher accuracy than training on CERD+A1 alone, and combining

both datasets leads to the best overall results. This underscores the value of heterogeneous, manually-labelled corpora for capturing subtle differences among CEFR levels. Nonetheless, errors remain frequent at intermediate levels (especially B1 and B2), a pattern also observed in the TSAR Evaluator Model and consistent with the ambiguity of these adjacent bands in the CEFR scale.

The persistent difficulty with A1 texts merits additional discussion. Both our fine-tuned models and the TSAR Evaluator Model struggle to correctly identify A1, often misidentifying these texts as A2 or B1. This likely reflects two intertwined factors: the scarcity of simple texts in publicly available corpora, and the possibility that pretraining data contains far fewer prototypical A1 inputs than texts at intermediate or advanced levels. As simplicity increases, models may also over-rely on lexical or structural features that are not distinctive enough when difficulty is low.

Another important observation is that models smaller than 8B parameters behave inconsistently and are more sensitive to prompt wording, whereas 8B models or bigger converge towards similar performance.

Finally, our comparison with off-the-shelf tools like Text Inspector or CVLA, and proprietary LLMs like ChatGPT, shows that proprietary LLMs perform remarkably well in this task, but comparable, if not slightly lower, to fine-tuned open-source LLMs. State-of-the-art tools like CVLA or Text Inspector show slightly better performance than prompting strategies, but perform significantly worse than fine-tuned models. Finally, a fine-tuned classifier (encoder only) model shows very good performance, but this must in part be attributed to data leakage: the model was possibly trained on the data we held out for testing.

These findings suggest that dedicated, open-weight LLMs offer a robust solution for CEFR level aware adaptation pipelines, with performance comparable to commercial models and beyond non-LLM approaches.

7 Conclusions and Future Work

Our study shows that open-source generative LLMs can reliably identify the CEFR level of texts, provided that they are appropriately fine-tuned. While zero-shot and in-context prompting lead to moderate and sometimes unstable performance, fine-tuned models achieve accuracy levels comparable to state-of-the-art CEFR classifiers and commercial models, despite being trained on substantially smaller datasets. Model size proves relevant only up to a certain threshold, with models of 8B parameters or larger showing similar performance profiles. Additionally, the availability of diverse well-curated labelled data, particularly at the lower and intermediate CEFR bands, plays a key role in overall accuracy. Persistent misclassifications around A1 and at the B1–B2 boundary may reflect inherent ambiguities in the CEFR scale, as well as data scarcity, and highlight the continued need for richer and more balanced corpora.

Future work will address several limitations identified in this study. First, we plan to expand the training datasets, particularly for underrepresented CEFR levels such as A1, in order to improve robustness and reduce class imbalance effects. In doing so, we will systematically record and calculate inter-annotator agreement.

Then, we want to better assess the differences in performance of encoder-only models, like BERT, and decoder-only models, like the Llama we have been using, both in terms of accuracy and performance.

In addition, because of hardware constraints, the current fine-tuning setup was relatively restricted, as it relies on a linear classification layer with the rest of the model frozen. We will explore more expressive and contemporary approaches, including instruction tuning and parameter-efficient fine-tuning methods, which may better capture the subtle linguistic features underlying CEFR distinctions.

Furthermore, although this study focuses on CEFR classification as an intermediate task, it does not directly validate its impact on downstream text adaptation, which remains the ultimate objective of the project. Future research will investigate how improvements in classification translate into the quality and pedagogical adequacy of automatically generated text adaptations.

Finally, the systematization of the methodology to assess LLMs for the task of CEFR adaptation makes it more feasible to apply to low-resource languages. As a proof-of-concept, we are planning to apply these to Russian as part of our project.

Acknowledgements

This research has been funded and made possible by Grant PID2023-149648NB-I00, funded by the Ministerio de Ciencia, Innovación y Universidades of the Spanish government, and the National Research Agency (AEI). This work used computational resources from UNC Supercómputo (CCAD) – Universidad Nacional de Córdoba, which are part of SNCAD, República Argentina. We would also like to thank the EFL teachers that carried out the manual evaluation, as well as the reviewers for their insightful comments and suggestions, which have contributed to improve this work.

References

Alva-Manchego, F., R. Stodden, J. M. Imperial, A. Barayan, K. North, and H. Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China, November. Association for Computational Linguistics.

Bax, S. 2025. Text inspector. `https://textinspector.com/`. Computer program. Accessed 18 November 2025.

Benedetto, L., G. Gaudeau, A. Caines, and P. Buttery. 2025. Assessing how accurately large language models encode and apply the common european framework of reference for languages. *Computers and Education: Artificial Intelligence*, 8:100353.

Breuker, M. 2023. Cefr labelling and assessment services. In G. Rehm, editor, *European Language Grid*, Cognitive Technologies. Springer, Cham.

Castellví Vives, J., R. Gilabert Guerrero, and E. Comelles Pujadas. 2024. Automating task design: Bridging the gap between second language research and l2 instruction. *TEISEL. Tecnologías para la investigación en segundas lenguas*, 3:1–21.

Chen, X. and D. Meurers. 2018. Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, 41(3):486–510.

Coleman, M. and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing, Strasbourg.

Deutsch, T., M. Jasbi, and S. Shieber. 2020. Linguistic features for readability assessment. In J. Burstein, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, H. Yannakoudakis, and T. Zesch, editors, *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online, July. Association for Computational Linguistics.

Díez-Bedmar, M. B. and G. Luque-Agulló. 2023. Analysing the cefr/cv in university language centres in spain: The rater perspective. In *Global Perspectives on Effective Assessment in English Language Teaching*. IGI Global, pages 1–33.

Farajidizaji, A., V. Raina, and M. Gales. 2024. Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9325–9339, Torino, Italia.

Hancke, J., S. Vajjala, and D. Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In M. Kay and C. Boitet, editors, *Proceedings of COLING 2012*, pages 1063–1080, Mumbai, India, December. The COLING 2012 Organizing Committee.

Heilman, M., K. Collins-Thompson, J. Callan, and M. Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In C. Sidner, T. Schultz, M. Stone, and C. Zhai, editors, *Human Language Technologies 2007: The Conference*

of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pages 460–467, Rochester, New York, April. Association for Computational Linguistics.

Huang, C. Y., J. Wei, and T. H. K. Huang. 2024. Generating educational materials with different levels of readability using llms. In *Proceedings of the 3rd Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2024), co-located with the ACM CHI Conference on Human Factors in Computing Systems (CHI 2024)*, ACM International Conference Proceeding Series, pages 16–22. Association for Computing Machinery.

Imperial, J. M., A. Barayan, R. Stodden, R. Wilkens, R. Muñoz Sánchez, L. Gao, M. Torgbi, D. Knight, G. Forey, R. R. Jablonkai, E. Kochmar, R. J. Reynolds, E. Ribeiro, H. Saggion, E. Volodina, S. Vajjala, T. François, F. Alva-Manchego, and H. Tayyar Madabushi. 2025. UniversalCEFR: Enabling open multilingual research on language proficiency assessment. In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9714–9766, Suzhou, China, November. Association for Computational Linguistics.

Kincaid, P., R. P. Fishburne, R. L. Rogers, and B. S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Lee, B. W., Y. S. Jang, and J. Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Lee, J. and S. Vajjala. 2022. A neural pairwise ranking model for readability assessment. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3802–3813, Dublin, Ireland, May. Association for Computational Linguistics.

Malik, A., S. Mayhew, C. Piech, and K. Bicknell. 2024. From tarzan to Tolkien: Controlling the language proficiency level of LLMs for content generation. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15670–15693, Bangkok, Thailand, August. Association for Computational Linguistics.

McLaughlin, G. H. 1969. Smog grading–a new readability formula. *The Journal of Reading*, 12(8):639–646.

Meng, C., M. Chen, J. Mao, and J. Neville, 2020. *ReadNet: A Hierarchical Transformer Framework for Web Article Readability Analysis*, page 33–49. Springer International Publishing.

Meta. 2024. meta-llama/llama-3.1-8b-instruct. Accessed: 2025-02-21.

Milton, J. and T. Alexiou. 2009. Vocabulary size and the common european framework of reference for languages. In B. Richards, M. H. Daller, D. D. Malvern, P. Meara, J. Milton, and J. Treffers-Daller, editors, *Vocabulary Studies in First and Second Language Acquisition*. Palgrave Macmillan, London, pages 194–211.

Milton, J. and T. Alexiou. 2020. Vocabulary size assessment: Assessing the vocabulary needs of learners in relation to their CEFR goals. In M. Dodigovic and M. P. Agustín-Llach, editors, *Vocabulary in Curriculum Planning*. Palgrave Macmillan, Cham.

Montgomerie, A. 2021. CEFR Levelled English Texts. Kaggle Dataset. Accessed: 2025-11-20.

Oviedo, J. C., E. Comelles Pujadas, L. Alonso Alemany, and J. Atserias Batalla. 2025. taskGen at TSAR 2025 shared task exploring prompt strategies with linguistic knowledge. In M. Shardlow, F. Alva-Manchego, K. North, R. Stodden, H. Saggion, N. Khallaf, and A. Hayakawa, editors,

Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025), pages 160–172, Suzhou, China, November. Association for Computational Linguistics.

Oxford University Press. n.d. The oxford 5000: The list of the 5000 most important words to learn in English, from A1 to B2 level. `https://www.oxfordlearnersdictionaries.com/external/pdf/wordlists/oxford-3000-5000/The_Oxford_5000.pdf`. Accessed: 2025-09-21.

Ribeiro, L. F. R., M. Bansal, and M. Dreyer. 2023. Generating summaries with controllable readability levels. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11669–11687, Singapore. Association for Computational Linguistics.

Simons, M. and J. Colpaert. 2015. Judgmental evaluation of the cefr by stakeholders in language testing. *Revista de Lingüística y Lenguas Aplicadas*, 10(1):66–77.

Tran, H., Z. Yao, L. Li, and H. Yu. 2025. Readctrl: Personalizing text generation with readability-controlled instruction learning. In *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)*, pages 19–36, Albuquerque, New Mexico, US. Association for Computational Linguistics.

Uchida, S. and M. Negishi. 2025. Estimating the cefr-j level of english reading passages: Development and accuracy of cvla3. *English Corpus Studies*, 32:165–174.

Vajjala, S. 2022. Trends, limitations and open challenges in automatic readability assessment research. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France, June. European Language Resources Association.

Vajjala, S. and D. Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In J. Tetreault, J. Burstein, and C. Leacock, editors, *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montréal, Canada, June. Association for Computational Linguistics.

Vajjala, S. B. 2015. *Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications*. Phd thesis, Universität Tübingen, Philosophische Fakultät, Allgemeine und Vergleichende Sprachwissenschaft, Tübingen, Germany, August. Date of defense: 2015-08-03.

Warner, B., A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, G. T. Adams, J. Howard, and I. Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria, July. Association for Computational Linguistics.

Xia, M., E. Kochmar, and T. Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, San Diego, California. Association for Computational Linguistics.

A Appendix: Prompts

A.1 Baseline: 0-shot prompts

Prompt 0: you are an English teacher and I want you to classify the following text according to the CEFR classes. I want you to respond only with the class (A1—B1—C1—A2—B2—C2). Without introduction.

Prompt 1: you are an English teacher and I want you to classify the following text for reading comprehension according to the CEFR classes. I want you to respond only with the class (A1—B1—C1—A2—B2—C2), without introduction.

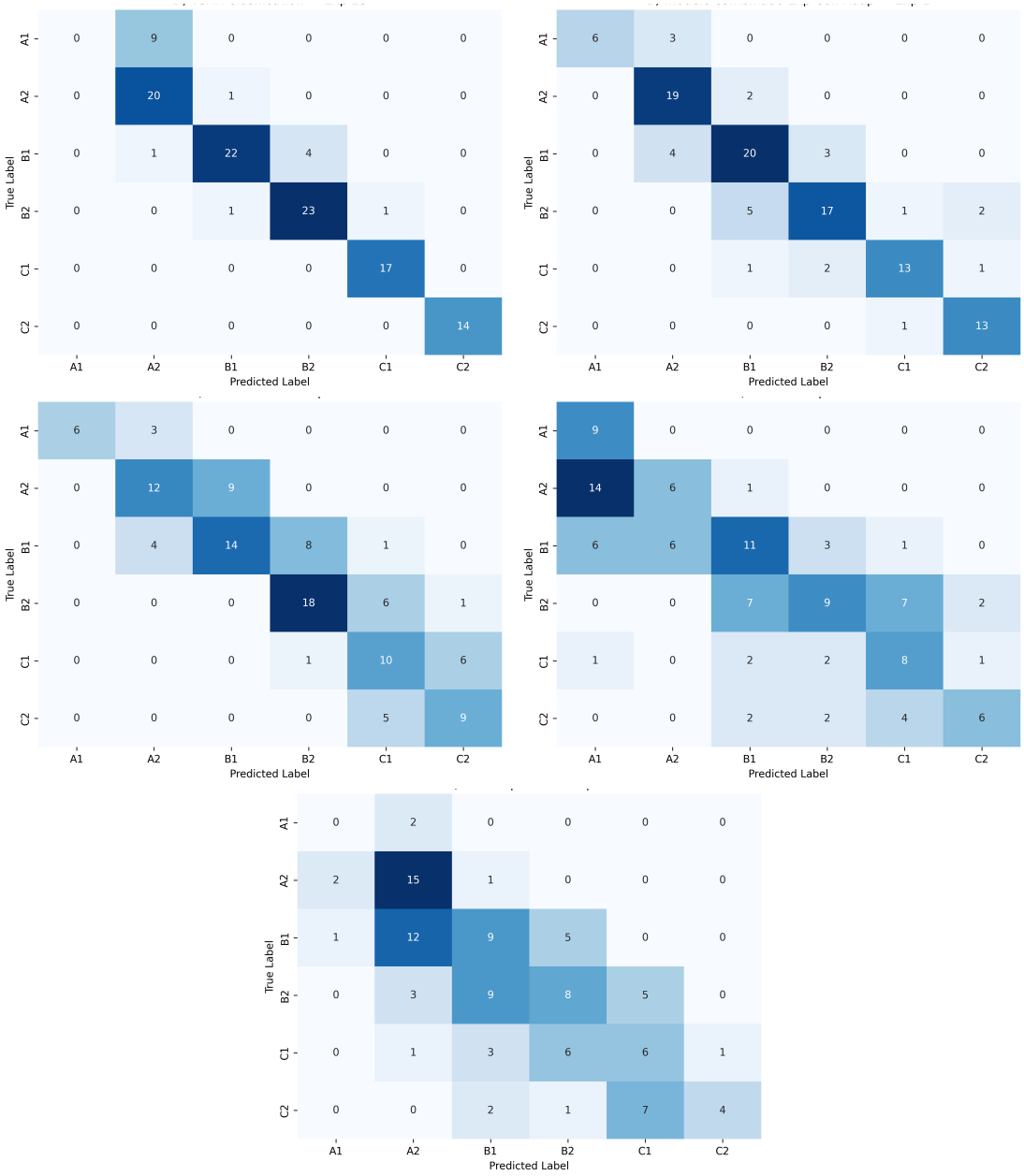

Figure 3: From left to right, top to bottom, confusion matrices for: the TSAR Model, the best performing model, the fine-tuned Llama with prompt 1, ChatGPT with prompt 15, CVLA and Text Inspector.

Prompt 2: you are an English teacher and I want you to classify the following text according to the CEFR classes. I want you to responds only with the class (A1—A2—B1—B2—C1—C2). Without introduction. I detect that you have a bias to classify everything as B2, correct it.

Prompt 3: you are an English teacher and I want you to classify the following text for reading comprehension according to the CEFR classes. I want you to responds only with the class (A1—A2—B1—B2—C1—C2). Without introduction. I detect that you have a bias to classify everything as B2, correct it.

Prompt 4: Can you classify the following text according to its Common European Framework of Reference (CEFR) readability level? I want you to respond only with the class (A1—A2—B1—B2—C1—C2). Without introduction.

Prompt 5: I'm a teacher of English who is preparing a reading task. Can you classify the following text regarding its readability based on the CEFR? I

want you to respond only with the class (A1—A2—B1—B2—C1—C2). Without introduction.

Prompt 6: I'm a teacher of English as a foreign language. I'm preparing a reading task for my students and I would like to use the following text. Can you tell me which CEFR level is the following text suitable for? I want you to respond only with the class (A1—A2—B1—B2—C1—C2). Without introduction.

Prompt 7: I'm a teacher of English as a foreign language. I'm preparing a reading task for my students and I would like to use the following text. Can you classify the following text according to its CEFR readability level? I want you to respond only with the class (A1—A2—B1—B2—C1—C2). Without introduction.

Prompt 8: I'm a teacher of English as a foreign language. I'm preparing a reading task for my students and I would like to use the following text. Can you annotate the following text according to its CEFR readability level? I want you to respond only with the class (A1—A2—B1—B2—C1—C2). Without introduction.

A.2 1-shot and 2-shot prompts

Prompt 9 (1 example for each level + Prompt 0): this is an example of a text for level A1 <EXAMPLE>
this is an example of a text for level B1 <EXAMPLE>
this is an example of a text for level B2 <EXAMPLE>
this is an example of a text for level C1 <EXAMPLE>
this is an example of a text for level C2 <EXAMPLE>
you are an English teacher and I want you to classify the following text according to the CEFR classes. I want you to responds only with the class (A1—A2—B1—B2—C1—C2). Without introduction.

Prompt 10 (1 example for each level + Prompt 1), as in Prompt 9.

Prompt 11 (1 example for each level + Prompt 2), as in Prompt 9.

Prompt 12 (1 example for each level + Prompt 3), as in Prompt 9.

Prompt 13 (1 example for each level except level B2 + Prompt 0), as in Prompt 9.

Prompt 14 (1 example for each level except level B2 + Prompt 1), as in Prompt 9. 2-shot prompts

Prompt 15 (2 examples for each level + Prompt 0), as in Prompt 9.

Prompt 16 (2 examples for each level + Prompt 1), as in Prompt 9.

Prompt 17 (2 examples for each level + Prompt 2), as in Prompt 9.

Prompt 18 (2 examples for each level + Prompt 2), as in Prompt 9.

Prompt 19 (2 examples for each level except level B2 + Prompt 0), as in Prompt 9.

Prompt 20 (2 examples for each level except level B2 + Prompt 1), as in Prompt 9.

A.3 Linguistic knowledge prompts

Prompt 21: this is a list of CEFR level descriptors for level A1: <DESCRIPTORS A1>
this is a list of CEFR level descriptors for level A2: <DESCRIPTORS A2>
this is a list of CEFR level descriptors for level B1: <DESCRIPTORS B1>
this is a list of CEFR level descriptors for level B2: <DESCRIPTORS B2>
this is a list of CEFR level descriptors for level C1: <DESCRIPTORS C1>
this is a list of CEFR level descriptors for level C2: <DESCRIPTORS C2>
Using the above descriptors I want you to classify the following text according to its Common European Framework of Reference (CEFR) level. I want you to responds only with the class (A1—A2—B1—B2—C1—C2). Without introduction.

Prompt 22: same as Prompt 21, without descriptors for level B2

Prompt 23: This is a list of words sorted by levels A1, A2, B1, B2, C1, C2. In each level you will find words that a text of that level should have. In addition, the words of each higher level include the words of the lower level.
These are the words of level A1 <VOCABULARY A1>
These are the words of level A2 <VOCABULARY A2>
These are the words of level B1 <VOCABULARY B1>
These are the words of level B2 <VOCABULARY B2>
These are the words of level C1 and C2 <VOCABULARY C1 and C2>

Using the above words, I want you to classify the following text according to its Common European Framework of Reference (CEFR) level. I want you to responds only with the class (A1—A2—B1—B2—C1—C2). Without introduction.

B Appendix: Hyperparameters for fine-tuning

.

We fine-tuned a Llama 3.1 8B (Meta, 2024) with the training and tuning portions of the manually labelled corpus, combining both labelled datasets.

We added a 6-neuron linear layer to the model, with each neuron corresponding to a CEFR class. Model parameters were frozen, except for the classification layer.

While parameter-efficient fine-tuning (PEFT) alternatives such as LoRA or QLoRA are widely used, our experimental methodology was conditioned by technical infrastructure and hardware compatibility constraints.

More specifically, our computational environment relied on an AMD GPU architecture within the ROCm ecosystem. Due to system-level permission limitations that prevented the installation of specialized low-level dependencies and custom quantization kernels—which are required to support 4-bit or 8-bit quantization libraries in this hardware ecosystem—implementing a stable LoRA or QLoRA pipeline was technically unfeasible.

Consequently, we opted for a robust linear probing approach, freezing all base model parameters and training exclusively the 6-neuron classification layer. This strategy effectively circumvented hardware-software stack conflicts while minimizing computational overhead and preventing overfitting.

After performing a grid search resorting to the validation portion of our dataset, the best hyperparameters for fine-tuning were:

- Optimizer: AdamW
- Learning rate: 1×10^{-3}
- Batch size: 8 × accumulated gradient at 4 steps
- Epochs: 30
- Loss function: cross entropy
- Regularization
  - no additional dropout
  - weight decay = 0
  - label smoothing = 0
- Main selection metric: Accuracy