# Using the CEFR for Guiding LLMs in Lexical Complexity Prediction

Maria Angelica Riera Machin[1] Adam Nohejl[1,2] Taro Watanabe[1]
[1]Nara Institute of Science and Technology [2]RIKEN
riera_machin.maria.rn9@naist.ac.jp adam.nohejl@riken.jp taro@is.naist.jp

## Abstract

Lexical complexity prediction is a key task for language teaching, as word difficulty varies across learners and proficiency levels. While prior work has largely focused on generating content aligned with the Common European Framework of Reference (CEFR), less attention has been paid on whether lexical complexity judgments themselves align with CEFR standards. This paper investigates whether large language models (LLMs) can reliably predict the CEFR level of individual words using zero-shot and few-shot prompting. We evaluate model outputs against three independent CEFR-annotated lexical datasets. Results show that prompt design strongly affects performance. Additionally, LLM predictions exhibit stronger alignment with certain lexical resources, suggesting that model judgments reflect biases present in their training data. These findings highlight both the potential and limitations of LLMs for CEFR-based lexical assessment.

## 1 Introduction

Research on text accessibility has increasingly focused on understanding and assessing word difficulty, driving progress in areas such as lexical simplification (LS) [1], lexical complexity prediction (LCP) [2], and automated language proficiency assessment [3]. These tasks aim to determine how challenging individual words are for different readers and to adapt text accordingly. With the rise of large language models (LLMs), both LS and LCP have benefited from stronger contextual understanding and generation capabilities, enabling more accurate complexity estimation and more effective simplification strategies.

Standardized frameworks such as the Common European Framework of Reference (CEFR) provide a structured way to describe language proficiency levels [4], and using CEFR levels as a proxy for lexical complexity helps align model predictions with language-learning needs while enabling controlled evaluation of simplification quality. However, despite growing interest in CEFR-controlled text generation [5] and CEFR-based classification of multi-word expressions [3, 6], relatively little work has examined whether LLMs can reliably predict the CEFR level of individual words. Consequently, it remains unclear how well LLMs perform on word-level CEFR classification, particularly across different prompting strategies and diverse lexical resources.

This work proposes an evaluation of LLMs on how well they can estimate the CEFR-aligned difficulty of a word, when supplied with a word list and given explicit CEFR-related information, including level descriptors and example vocabulary. We investigate the effect of different prompting strategies and compare multiple LLMs across three CEFR-annotated English lexical resources. The experimental results indicate that LLM performance is highly sensitive to prompt design. In particular, simpler prompts tend to yield better results, and the inclusion of CEFR level descriptions does not necessarily lead to performance improvements.

## 2 Related Work

Work on CEFR level classification has predominantly relied on masked language models, most notably BERT. Prior studies have framed CEFR prediction as a supervised classification task, leveraging contextualized BERT embeddings to capture semantic and syntactic cues relevant to lexical and sentence-level difficulty. [7] employs BERT-based contextual representations combined with traditional classifiers to predict CEFR levels for lexical items in context. [6] proposes a metric-based approach in which BERT sentence embeddings are compared against CEFR-level prototypes to address class imbalance. While these approaches achieve strong performance, they rely on carefully curated labeled datasets and task-specific training or model design.

Recently, LLMs have also been increasingly explored for simplifying and generating content aligned with learners' proficiency levels. One line of work investigates the ability of generative models to grade and produce vocabulary lists. LLMs have been examined in how they assign scores to words on a scale from 0 to 4 [8]. Although CEFR levels were not explicitly used for this study, the results demonstrated that modern LLMs can provide consistent difficulty judgments across multiple (English, Spanish, French, Swedish, and Dutch) languages in a zero shot setting.

A second line of research incorporates the CEFR directly into prompting. In [5], models such as GPT-4, Llama-2-7B, and Mistral-7B were evaluated on their ability to generate short stories that match a target CEFR levels. Models received a plot summary and a level (1–6) and were instructed to produce text aligned with both. Results showed that providing CEFR level descriptions significantly improved level control and reduced errors during generation.

# 3 Methodology

## 3.1 Datasets

Several CEFR-graded word lists are available for English, including the Cambridge English Vocabulary Profile (EVP) [9], Pearson's Global Scale of English (GSE) [10], and the CEFRLex family of datasets [11]. CEFRLex is available in five languages (English, Spanish, French, Swedish, and Dutch), although only the English subset (EFFLex) is used in this study.

The **EVP** and **GSE** datasets were obtained by scraping their respective websites, with final outputs stored in JSON format. The **EFFLex** data set differs from EVP and GSE in that it contains only information on document-frequency for each word, rather than a predetermined CEFR label. Therefore, its CEFR levels had to be inferred from this information. Due to the small size of the corpus, words that have a low document frequency (approximately 70% of words appear in less than 5 documents) might get incorrectly identified, introducing noise. To mitigate this, we removed words with document frequency below 20% of the corpus-wide average. We then assign CEFR levels using a hybrid approach that combines first-occurrence heuristics with weighted frequency scoring across levels. Words showing clear dominance at beginner levels (A1–A2) are classified based on their earliest non-zero occurrence to prevent them from being pushed to higher levels. For the remaining words, a continuous CEFR score was computed by weighting normalized frequency distributions across levels and mapping this score to discrete CEFR categories. This approach was selected since it aligned more closely with expert-annotated resources such as EVP and GSE.

**Table 1** Distribution of CEFR levels after dataset alignment

|  | **Final data** |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Dataset | A1 | A2 | B1 | B2 | C1 | C2 |
| EVP | 499 | 679 | 1049 | 948 | 101 | 79 |
| GSE | 524 | 282 | 705 | 1482 | 452 | 0 |
| EFFLex | 500 | 504 | 561 | 912 | 968 | 0 |

All datasets were cleaned by removing proper nouns (e.g., personal names), as well as words containing digits (e.g., *12th*, *1930s*, etc). Multi-word expressions (e.g., *bullet hole*) were also excluded. After this process, only single-word nouns, verbs, adjectives, and adverbs present in all three datasets were retained. For polysemous words with different levels depending on their sense, we assigned the lowest level, typically corresponding to the most common sense. For instance, the word *run* can be classified as **A1** when used as the intransitive verb meaning 'to move on one's feet faster than walking,' but as **C1** when used metaphorically, as in 'a piece of equipment is running.' Since the former represents the most frequent and basic usage, *run* was assigned the **A1** level. The final word count is 3,445 and the level distributions are shown in Table 1.

## 3.2 Prompt setting

Six prompts were evaluated on the final datasets. They were organized into three groups based on the number of examples provided to the model (zero-shot, 30-shot, and 60-shot) and further distinguished by the inclusion of CEFR level descriptions (A1–C2), resulting in a total of six prompts for evaluation. The CEFR level descriptions were sourced from the official Council of Europe website [12]. All models used a temperature of 0.0.

**Zero shot:** No examples, no CEFR descriptions.

**Zero shot+CEFR:** No examples, CEFR descriptions.

**30-shot:** 5 word examples per CEFR level, totaling 30.

**30-shot+CEFR:** 5 word examples per CEFR level, totaling 30, and CEFR descriptions.

**60-shot:** 10 word examples per CEFR level, totaling 60.

**60-shot+CEFR:** 10 word examples per CEFR level, totaling 60, and CEFR descriptions.

The 5 or 10 examples per each of the CEFR levels that were used in the 30-shot and 60-shot prompts were drawn at random from a selected subset of words. Due to the nature of the CEFR itself, level boundaries can be ambiguous, which leads to disagreements across the three datasets. To mitigate this issue, a smaller subset containing only words for which all three datasets *agree with each other* was extracted to be used exclusively as examples in the prompts. The total number of words satisfying this condition is 507.

Since this subset does not contain any C2-level words due to the constraints of the GSE and EFFLex resources, the C2 examples were sourced from the final EVP dataset (which contains 79 C2 words). These examples were selected at random. An example of the simplest prompt that was used can be seen below.

**Zero-shot prompt:**

> You are an expert in language learning and CEFR levels. The CEFR stands for the Common European Framework of Reference for Languages. It's an international standard for describing language ability and it categorizes language proficiency into six levels (A1, A2, B1, B2, C1, C2).
>
> Evaluate the CEFR level of the following list of English words: `[word_list]`.
>
> If a word has multiple senses or grammatical roles, classify it according to its most frequent everyday meaning. Provide the output in a valid JSON format that only contains 'word' and 'level' keys. Do not repeat any word, each word should appear exactly once.

## 4 Results

The experiments were conducted using GPT4o-mini [13], Gemini-2.5-flash [14], Mistral-7B-Instruct [15] and LlaMA3-8B-Instruct [16]. The models were evaluated using the six prompts discussed earlier. Performance was assessed using F1 score and quadratic weighted kappa (QWK) on each of the three datasets.

The results in Table 2 show a clear performance gap between closed-source and open-source models. The closed-source models, GPT4o-mini and Gemini-2.5-flash, consistently achieve higher agreement scores than the open-source models Mistral-7B-Instruct and LLaMA-3-8B-Instruct across all prompts. The open-source models, however, outperform a simple frequency-based baseline (see Appendix C), which achieves QWK of **0.3712**. GPT4o-mini yields the strongest overall performance, with its best configuration (60-shot without CEFR descriptions) reaching a QWK of **0.6669**. Gemini-2.5-flash follows closely, achieving a comparable peak QWK of **0.6546** in the zero-shot setting. In contrast, both open-source models perform substantially worse, with Mistral-7B-Instruct having the weakest overall performance. This trend is also reflected in the F1 scores, where closed-source models consistently outperform open-source ones across all prompting strategies. In F1 too GPT4o-mini achieves the highest overall performance, while open-source models remain notably behind, indicating weaker per-class prediction accuracy in addition to lower ordinal agreement. Similar results for the GSE and EFFLex datasets are reported in Appendix A.

**Table 2** Scores across prompts against the EVP dataset

| | **GPT4o-mini** | | **Gemini-2.5-flash** | |
|---|---|---|---|---|
| Prompt | F1 | QWK | F1 | QWK |
| Zero-shot | 0.4835 | 0.6616 | 0.4303 | **0.6546** |
| Zero-shot+CEFR | 0.4687 | 0.6314 | 0.3746 | 0.5803 |
| 30-shot | 0.4767 | 0.6610 | 0.4122 | 0.5978 |
| 30-shot+CEFR | 0.4651 | 0.6365 | 0.4129 | 0.6006 |
| 60-shot | 0.4766 | **0.6669** | 0.4063 | 0.6120 |
| 60-shot+CEFR | 0.4694 | 0.6431 | 0.4179 | 0.6024 |
| | **Mistral-7B** | | **LLaMA-3-8B** | |
| Prompt | F1 | QWK | F1 | QWK |
| Zero-shot | 0.3142 | 0.3678 | 0.3276 | 0.4135 |
| Zero-shot+CEFR | 0.2867 | 0.3140 | 0.3452 | 0.3807 |
| 30-shot | 0.3335 | 0.3942 | 0.3368 | 0.4356 |
| 30-shot+CEFR | 0.3291 | 0.3712 | 0.3578 | 0.4453 |
| 60-shot | 0.3479 | **0.4256** | 0.3644 | **0.4550** |
| 60-shot+CEFR | 0.3432 | 0.3766 | 0.3655 | 0.4530 |

Beyond quantitative differences, the LLM's behavior also varies across models. Mistral-7B-Instruct frequently struggles to follow prompt instructions, repeating entries, or hallucinating labels. These issues become more pronounced as the number of words evaluated per prompt increases; reducing batch size from the original 30 words per request to 10–15 mitigates this behavior. Additionally, it occasionally misspells target words (e.g., “sharpely” instead of “sharply”), further complicating evaluation. LLaMA-3-8B-Instruct presents a different set of chal-

**Table 3** QWK scores between dataset pairs

| Datasets | F1 score | QWK |
|---|---|---|
| EVP/GSE | 0.4412 | **0.7257** |
| EVP/EFFLex | 0.2466 | 0.5386 |
| GSE/EFFLex | 0.3518 | 0.5219 |
| Mean | 0.5954 | 0.3049 |

lenges. Without explicit output instructions, it frequently deviates from the expected JSON structure or includes explanatory reasoning alongside predictions. It also shows a tendency to split words, such as separating “interestingly” into “interesting” and “ly”. In contrast, GPT4o-mini and Gemini-2.5-flash exhibit stable behavior and consistently adhere to the required output format.

With the exception of Gemini-2.5-flash, the overall best-performing configuration for each model corresponds to the 60-shot prompt without CEFR descriptions, indicating that providing concrete, labeled examples is more beneficial than descriptive guidance. Conversely, adding CEFR level descriptions rarely improves performance and often degrades it: in nearly all prompt pairs, the version including CEFR descriptions performs worse than its simpler counterpart. This pattern suggests that LLMs benefit more from practical examples than from abstract descriptions. Furthermore, longer, more complex prompts may also introduce ambiguity that hinders effective CEFR-level prediction, skewing prediction towards higher levels.

## 4.1 Agreement between datasets

QWK and F1 score were used as a metric to measure the datasets’ agreement level between each other, thus they were calculated for each pair (EVP/GSE, EVP/EFFLex, GSE/EFFLex). The mean QWK and F1 scores across all three datasets were also calculated and are shown in 3.

The strong agreement between GSE and EVP indicates that expert-curated CEFR resources share a largely consistent interpretation of lexical difficulty, despite differences in scale design and annotation procedures. In contrast, the lower agreement that GSE and EVP achieve with EFFLex suggests a systematic mismatch. This divergence points to a fundamental difference in what is being measured: while GSE and EVP reflect expert-judgement of word difficulty, EFFLex instead captures usage frequency across educational texts. As a result, frequency-based CEFR approximations do not reliably align with expert judgments.

Compared to the inter-dataset agreement scores, the best-performing LLM configuration (GPT with a 60-shot prompt without CEFR descriptions), achieves substantial but still lower agreement, QWK of 0.6669 on EVP and 0.6410 on GSE, than the agreement observed between these two datasets with QWK of 0.7257. This indicates that, although the evaluated LLMs can approximate expert judgments of lexical difficulty under optimized prompting, they do not fully match the consistency achieved by human-annotated CEFR resources. Notably, LLM agreement is considerably higher than the agreement involving EFFLex, reinforcing the view that LLM predictions align more closely with expert-annotated data than with frequency-based approximations. Together, these results suggest that expert-annotated datasets remain a stronger reference for CEFR evaluation, while LLMs offer a competitive but limited approximation of expert consensus.

# 5 Conclusions

This work evaluated four LLMs (GPT4o-mini, Gemini-2.5-flash, Mistral-7B-Instruct, and LLaMA-3-8B-Instruct) covering both closed and open-source systems. The task focused on CEFR-guided lexical classification, assessing how effectively these models predict the difficulty level of English words when provided with different types of prompt information. Across nearly all prompt configurations and datasets, GPT4o-mini outperformed the other models, while Mistral-7B-Instruct exhibited the weakest overall performance. Notably, prompts that included explicit CEFR level descriptions typically underperformed compared to their simpler counterparts, suggesting that additional descriptive information may introduce ambiguity. In contrast, providing labeled word examples generally improved performance for all models except Gemini-2.5-flash, indicating that exemplar-based prompting is more effective than descriptive guidance for this task.

These results highlight the strong sensitivity of CEFR-based lexical classification to both model choice and prompt design. While more detailed prompts might seem beneficial, our findings suggest that concise, example-driven prompts better align with how current LLMs internalize and apply notions of lexical difficulty. This reinforces the role of expert-annotated data as a benchmark when deploying LLMs for educational and proficiency-oriented language tasks.

## References

[1] Kyle North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. Deep learning approaches to lexical simplification: A survey. *Journal of Intelligent Information Systems*, Vol. 63, pp. 111–134, February 2025.

[2] Kai North, Marcos Zampieri, and Matthew Shardlow. Lexical complexity prediction: An overview. *ACM Comput. Surv.*, Vol. 55, No. 9, January 2023.

[3] Joseph Marvin Imperial, Barayan, and Othersxf. UniversalCEFR: Enabling open multilingual research on language proficiency assessment. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9714–9766, Suzhou, China, November 2025. Association for Computational Linguistics.

[4] Council of Europe. The CEFR Levels – Common European Framework of Reference for Languages. https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions, 2025. Accessed: 2025.

[5] Ali Malik, Stephen Mayhew, Christopher Piech, and Klinton Bicknell. From tarzan to Tolkien: Controlling the language proficiency level of LLMs for content generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15670–15693, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[6] Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. CEFR-based sentence difficulty annotation and assessment. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6206–6219, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[7] Desislava Aleksandrova and Vincent Pouliot. CEFR-based contextual lexical complexity classifier in English and French. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch, editors, *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pp. 518–527, Toronto, Canada, July 2023. Association for Computational Linguistics.

[8] David Alfter. Out-of-the-box graded vocabulary lists with generative language models: Fact or fiction? In Thomas Gaillat, Cyriel Mallart, Fabienne Moreau, Jen-Yu Li, Griselda Drouet, David Alfter, Elena Volodina, and Arne Jönsson, editors, *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pp. 1–19, Rennes, France, October 2024. LiU Electronic Press.

[9] Annette Capel. The english vocabulary profile. In Julia Harrison and Fiona Barker, editors, *English Profile in Practice*, chapter 2, pp. 9–27. Cambridge University Press, 2015.

[10] Pearson. Gse teacher toolkit. https://www.english.com/gse/teacher-toolkit/user/vocabulary, 2017. Accessed 2025.

[11] Luise Dürlich and Thomas François. EFLLex: A graded lexical resource for learners of English as a foreign language. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[12] Council of Europe. Common European Framework of Reference for Languages: Global Scale. https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale, 2025. Accessed: 2025-01-01.

[13] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report, 2024.

[14] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

[15] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[16] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models, 2024.

[17] Adam Nohejl, Frederikus Hudi, Eunike Andriani Kardinata, Shintaro Ozaki, Maria Angelica Riera Machin, Hongyu Sun, Justin Vasselli, and Taro Watanabe. Beyond film subtitles: Is YouTube the best approximation of spoken vocabulary? In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9566–9585, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

# A Results in GSE and EFFLex data

The results on the GSE and EFFLex datasets follow trends similar to those observed for EVP. Across all three datasets, the best-performing prompt configurations are largely consistent, with the 60-shot prompt yielding the highest performance for all models except Gemini-2.5-flash. As in the EVP results, adding CEFR level descriptions generally degrades performance across most models. In addition, and consistent with earlier observations, all LLMs exhibit substantially weaker agreement with the frequency-based EFFLex dataset compared to the expert-annotated resources.

**Table 4** Scores across prompts against the GSE dataset

| | **GPT4o-mini** | | **Gemini-2.5-flash** | |
|---|---|---|---|---|
| Prompt | F1 | QWK | F1 | QWK |
| Zero-shot | 0.4181 | 0.6108 | 0.2672 | **0.5410** |
| Zero-shot+CEFR | 0.4133 | 0.5828 | 0.2313 | 0.4620 |
| 30-shot | 0.4499 | 0.6288 | 0.2536 | 0.4881 |
| 30-shot+CEFR | 0.4441 | 0.6130 | 0.2549 | 0.4836 |
| 60-shot | 0.4611 | **0.6410** | 0.2545 | 0.4954 |
| 60-shot+CEFR | 0.4489 | 0.6178 | 0.2574 | 0.4926 |
| | **Mistral-7B** | | **LLaMA-3-8B** | |
| Prompt | F1 | QWK | F1 | QWK |
| Zero-shot | 0.2559 | 0.3168 | 0.3259 | 0.4302 |
| Zero-shot+CEFR | 0.2084 | 0.2481 | 0.2612 | 0.3404 |
| 30-shot | 0.2785 | 0.3371 | 0.3829 | 0.4593 |
| 30-shot+CEFR | 0.2610 | 0.3074 | 0.3727 | 0.4476 |
| 60-shot | 0.2815 | **0.3699** | 0.3908 | **0.4731** |
| 60-shot+CEFR | 0.2754 | 0.3055 | 0.3767 | 0.4562 |

**Table 5** Scores across prompts against the EFFLex dataset

| | **GPT4o-mini** | | **Gemini-2.5-flash** | |
|---|---|---|---|---|
| Prompt | F1 | QWK | F1 | QWK |
| Zero-shot | 0.2391 | 0.4540 | 0.1680 | **0.4045** |
| Zero-shot+CEFR | 0.2345 | 0.4421 | 0.1455 | 0.3588 |
| 30-shot | 0.2591 | 0.4671 | 0.1657 | 0.3764 |
| 30-shot+CEFR | 0.2533 | 0.4677 | 0.1588 | 0.3687 |
| 60-shot | 0.2720 | **0.4824** | 0.1788 | 0.3765 |
| 60-shot+CEFR | 0.2562 | 0.4732 | 0.1680 | 0.3700 |
| | **Mistral-7B** | | **LLaMA-3-8B** | |
| Prompt | F1 | QWK | F1 | QWK |
| Zero-shot | 0.1814 | 0.2380 | 0.2207 | 0.3243 |
| Zero-shot+CEFR | 0.1554 | 0.2290 | 0.1635 | 0.2691 |
| 30-shot | 0.1846 | 0.2827 | 0.2511 | 0.3482 |
| 30-shot+CEFR | 0.1841 | 0.2651 | 0.2274 | 0.3318 |
| 60-shot | 0.1924 | **0.2854** | 0.2519 | **0.3515** |
| 60-shot+CEFR | 0.1801 | 0.2595 | 0.2310 | 0.3413 |

# B Other Prompt Templates

The same prompt templates were used across all three models, with one exception: for LLaMA-3-8B-Instruct, it was not possible to obtain a valid JSON output unless the output structure was explicitly specified. Consequently, the instruction “*Provide the output in a valid JSON format that only contains* `'word'` *and* `'level'` *keys. Do not repeat any word, each word should appear exactly once*” was replaced with “*Provide the output in a valid JSON format in the form of words: [word: <word>, level: <level>, ...] that contains both ‘word’ and ‘level’ keys for each word. Do not give any reasoning.*”

For prompts that included word examples, the word list was always provided in a simple comma-separated format, i.e., `example1, example2, etc.`

**30-shot prompt:**

> You are an expert in language learning and CEFR levels. The CEFR stands for the Common European Framework of Reference for Languages. It's an international standard for describing language ability and it categorizes language proficiency into six levels (A1, A2, B1, B2, C1, C2). Below are some word examples that belong to each CEFR level:
>
> Examples of 'A1' level words: `[word_examples]`.
> Examples of 'A2' level words: `[word_examples]`.
> Examples of 'B1' level words: `[word_examples]`.
> Examples of 'B2' level words: `[word_examples]`.
> Examples of 'C1' level words: `[word_examples]`.
> Examples of 'C2' level words: `[word_examples]`.
>
> Based on these examples and your own knowledge on the CEFR, evaluate the CEFR level of the following list of English words: `[word_list]`.
>
> If a word has multiple senses or grammatical roles, classify it according to its most frequent everyday meaning. Provide the output in a valid JSON format that only contains `'word'` and `'level'` keys. Do not repeat any word, each word should appear exactly once.

# C Frequency Baseline

The linear regression baseline was trained using log-frequency in TUBELEX [17] on 60 examples. Predictions were rounded to the nearest CEFR level.

**Table 6** Linear regression using log-frequency in TUBELEX.

| Dataset | F1 | QWK |
|---|---|---|
| EVP | 0.3026 | 0.3712 |
| GSE | 0.4171 | 0.5134 |
| EFFLex | 0.2275 | 0.2601 |