A Benchmark Study of Multi-Granular CEFR Level Assessment

Eugene Matsumura[1,2] Yuki Arase[1]
[1]Institute of Science Tokyo, Japan [2]RWTH Aachen University, Germany
eugene.matsumura@rwth-aachen.de
arase@c.titech.ac.jp

Abstract

Assessing CEFR (Common European Framework of Reference for Languages) levels of sentences and documents is a crucial technique for language learning and education support. However, it remains unclear (1) whether increasing model size or adopting decoder-only architectures (i.e., large language models (LLMs)) leads to improvements over prevalent encoder-based approaches, (2) whether there is a gap in sentence and document level assessment performance, and (3) how text granularity affects and interacts in the assessment. To answer these questions, we present a systematic benchmark of CEFR level assessment. We experiment with a wide range of architectures and prompting methods on both document and sentence level CEFR assessment. Our results reveal that, for CEFR classification, decoder-only models do not outperform encoder-based baselines, and that performance remains largely stable regardless of model size. Our findings also suggest that document-level assessment achieves higher precision, while sentences and documents do not interact well to improve the assessment performance.

1 Introduction

Automatic assessment of text difficulty is an important problem in natural language processing, with applications in education, language learning, and controllable text simplification [1]. In particular, CEFR provides a widely adopted scale for describing language proficiency on a six-level scale ranging from A1 to C2, corresponding to Basic (A), Independent (B), and Proficient (C). Thus, automatic CEFR assessment enables scalable evaluation of learner texts and the adaptation of reading materials to specific proficiency levels [2], making it a key component in systems that model or manipulate linguistic complexity.

Text granularity is an important factor in CEFR prediction. CEFR-SP [3] introduced a professionally annotated English sentence-level dataset. README++ [4] expanded this line of work with a multilingual, multi-domain benchmark. To facilitate analysis across granularities, Universal-CEFR [5] aggregates multiple resources into a large-scale multilingual benchmark with CEFR annotations at several text lengths. This corpus enables systematic comparisons across granularities.

Automatic CEFR classification has traditionally been studied using feature-based methods [6] and encoder-based neural architectures [7]. More recent work has focused on neural supervised learning approaches applied to sentence-level CEFR-annotated corpora. Despite this progress, most prior work relies on encoder-based architectures explicitly trained to predict masked tokens. Decoder-only LLMs, which are primarily optimized for next-token prediction, remain underexplored for CEFR classification, since their hidden states are not explicitly designed for downstream classification. LLM2Vec [8] addresses this limitation by adapting decoder-only models to produce encoder-style text representations. However, its effectiveness for CEFR classification has not been evaluated, leaving open questions about how architectural choices (encoder vs. decoder-only), model scale, and inference strategies (e.g., prompting vs. non-prompting) affect CEFR prediction.

Another crucial aspect that has been underexplored is the effets of text granularity. While most prior work targets sentence-level classification, document-level prediction and document and sentence interactions remain unclear. Document-level proficiency labels do not necessarily correspond to the difficulty of individual sentences, making sentence-level prediction a distinct task rather than a straightforward extension of document-level readability [3]. This motivates us to investigate how text length interacts with model architecture in CEFR classification.

To answer these questions, we conduct a benchmark study using the UniversalCEFR dataset. We found that:

- Despite significantly smaller model sizes, encoder-based models remain competitive with decoder-only LLMs under comparable fine-tuning setups.
- Sentence-level CEFR is harder than the document-level assessment.
- Sentences and documents do not interact well, and multi-task learning does not lead to performance improvement.

2 Experiment Settings

We formulate CEFR level prediction as a multi-class classification task, adopting a unified setup across different model architectures to ensure fair comparison. Given a sentence, the goal is to predict its CEFR proficiency level among the six standard categories (A1–C2).

Filtering We use UniversalCEFR, a large-scale CEFR-annotated dataset covering multiple levels of textual granularity, including sentences, paragraphs, and full documents. We restrict our experiments to the English portion of the dataset and consider only reference texts with valid CEFR labels. Non-standard or ambiguous labels (e.g., A1+) are excluded, retaining only canonical CEFR levels from A1 to C2.

Experiment Data Construction From this data, we derive three dataset variants as subsets of the same source: (1) a sentence-level dataset, (2) a document-level dataset, and (3) a combined dataset containing both sentence- and document-level samples. To ensure balance between sentence- and document-level data in the combined setting, sentence-level samples are subsampled to match the number of document-level samples. The data is split into 80% training, 10% validation, and 10% test sets using stratified sampling by CEFR label to preserve label distributions across splits. Dataset sizes for each variant are reported in Table 1.

Dataset	Train	Validation	Test
Sentence-level	10,260	1,283	1,283
Document-level	743	93	93
Combined	1,486	186	186

Table 1: Dataset sizes for each variant after stratified splitting by CEFR level.

Evaluation Metric The CEFR level distribution is naturally biased: the lower and higher ends (A1 and C1) are scarce [3]. Thus, we employ Macro-F1 as the evaluation metric to prioritise models with balanced performance across levels while penalising models that ignore minor classes.

3 Methods

To enable a comprehensive evaluation, we include models spanning a range of sizes and architectures.

3.1 Models with different architectures

We evaluate all model architectures at both the sentence and document levels using the corresponding dataset variants. Encoder-based models are BERT-style bidirectional Transformers. To examine the effect of model capacity within this family, we include both base- and large-sized variants: BERT-base/large [9], RoBERTa-base/large [10], DistilBERT-base [11], and ALBERT-base-v2 [12].

Decoder-only models are autoregressive generative LLMs. To study scaling effects in this class, we consider small models (1-2B parameters) and mid-sized models (7-8B parameters). Our small LLMs are TinyLlama-1.1B-Chat [13], Gemma-3-1B-it [14], and OLMo-2-0425-1B [15], while our mid-sized LLMs are LLaMA-3.1-8B-Instruct [16], Qwen2.5-7B-Instruct [17], and Mistral-7B-Instruct-v0.2 [18].

In addition to standard encoder- and decoder-only models, we evaluate LLM2Vec models, which adapt decoder-only LLMs to produce encoder-style text representations; we use the author-released LLM2Vec model based on LLaMA-3-8B-Instruct [8].

By comparing established encoder-based models, decoder-only LLMs, and LLM2Vec under a unified classification framework, we assess how different model architectures perform on the CEFR prediction task.

3.2 Prompting

For decoder-only models, we evaluate both prompted and non-prompted variants. The prompted setting uses an instruction-style prompt, while the non-prompted setting receives only the raw input sentence.

Model Type		Macro-F1
Encoder (Base)	BERT-base	0.56
	RoBERTa-base	0.55
	DistilBERT-base	0.54
	ALBERT-base-v2	0.57
Encoder (Large)	BERT-large	0.56
	RoBERTa-large	0.57
Decoder (1–2B)	TinyLlama-1.1B-Chat	0.54
	Gemma-3-1B-it	0.56
	OLMo-2-0425-1B	0.49
Decoder (7–8B)	LLaMA-3.1-8B-Instruct	0.59
	Qwen2.5-7B-Instruct	0.58
	Mistral-7B-Instruct-v0.2	0.56
LLM2Vec-based	LLM2Vec-8B	0.50

(a) Sentence-level CEFR prediction performance (Macro-F1) on the UniversalCEFR test set.

Model Type		Macro-F1
Encoder (Base)	BERT-base	0.86
	RoBERTa-base	0.90
	DistilBERT-base	0.77
	ALBERT-base-v2	0.76
Encoder (Large)	BERT-large	0.77
	RoBERTa-large	0.87
Decoder (1–2B)	TinyLlama-1.1B-Chat	0.86
	Gemma-3-1B-it	0.75
	OLMo-2-0425-1B	0.84
Decoder (7–8B)	LLaMA-3.1-8B-Instruct	0.77
	Mistral-7B-Instruct-v0.2	0.76
	Qwen2.5-7B-Instruct	0.79
LLM2Vec-based	LLM2Vec-8B	0.85

(b) Document-level CEFR prediction performance (Macro-F1) on the UniversalCEFR test set

Table 2: Sentence and document level CEFR prediction results (LLM2Vec-8B denotes the LLM2Vec model based on LLaMA-3-8B-Instruct)

Prompting is implemented in a zero-shot setting using an instruction-based prompt adapted from the ReadMe++ template, with all in-context examples removed. The full prompt is provided in Appendix A. All prompting experiments are conducted on the combined dataset.

3.3 Classification Head

To obtain a fixed-length representation from token-level hidden states, we apply pooling over the final hidden layer to form an embedding, which is passed to a classification head to predict the CEFR proficiency level. For all models, we apply standard mean pooling over token representations from the final hidden layer.

All models are fine-tuned with a lightweight classification head applied on top of the pooled representations, and the classification head is trained jointly with the base model parameters. The classification head adopts a multilayer perceptron design inspired by prior work on CEFR assessment [7]. The classifier is implemented as a two-layer multilayer perceptron with ReLU activations and dropout regularization, projecting from the model embedding dimension to a 128-dimensional hidden layer, followed by a linear projection to the CEFR label space. A dropout rate of 0.2 is applied between layers.

3.4 Training Settings

For decoder-only models, parameter-efficient fine-tuning is performed using Low-Rank Adaptation (LoRA) [19] via the Hugging Face PEFT library. We apply LoRA with rank $r = 16$, scaling factor $\alpha = 32$, and a dropout rate of 0.05. LoRA adapters are injected into both attention and feed-forward layers.

All models are optimized using the AdamW optimizer with a learning rate of 2×10^{-5}, which is used consistently across all experiments.

4 Results

We evaluate all models using macro-averaged F1 score (Macro-F1), which treats all CEFR classes equally.

4.1 Effects of Model Architecture

Finding 1: Model architectural and size differences have limited effects on CEFR classification performance across sentence and document granularity. At the sentence level (Table 2a), performance differences across architectures are modest. Encoder base models and decoder models achieve broadly similar Macro-F1 scores, and improvements from larger model sizes are inconsistent. At the document level (Table 2b), too, no consistent advantage of decoder-only models is observed in this setting, encoder models remain highly competitive.

Model	w/o prompt	w/ prompt
TinyLlama-1.1B-Chat	0.66	0.48
Gemma-3-1B-it	0.64	0.42
OLMo-2-0425-1B	0.65	0.31
LLaMA-3.1-8B-Instruct	0.66	0.64
Mistral-7B-Instruct-v0.2	0.58	0.59
Qwen2.5-7B-Instruct	0.68	0.51

Table 3: Comparison of prompting for decoder-only models on the UniversalCEFR combined test set (Macro-F1).

The LLM2Vec-based model also achieves strong performance compared to the corresponding standard decoder-only model. Overall, we observe no systematic performance advantage for decoder-only or LLM2Vec models over encoder-based models, suggesting that architectural differences among encoders, decoders, and LLM2Vec-based models have a limited effect in our setting. Similarly, model size alone is not a reliable predictor of CEFR classification performance.

Table 3 compares decoder-only models trained with and without an explicit CEFR-specific prompt. For larger models such as LLaMA-3.1-8B-Instruct and Mistral-7B-Instruct-v0.2, performance differences between the prompted and unprompted settings are small. In contrast, smaller models, including OLMo-2-0425-1B, Gemma-3-1B-it, and TinyLlama-1.1B-Chat, show substantially lower performance when prompting is applied. Qwen2.5-7B-Instruct also exhibits a decrease under prompting. Overall, these results indicate that CEFR-specific prompting often leads to lower prediction performance across the models.

4.2 Effects of Text Lengths

Finding 2: Document-level CEFR assessment performance is consistently and substantially higher than that of the sentence level. As shown in (Table 2), this is the case on all models, across architectures. It is expected because documents provide more clues to assess CEFR levels. However, the best macro-F1 is limited as 0.90 (RoBERTa-base); improvements are desired to use it for educational purposes.

4.3 Effects of Multi-Task Learning

Finding 3: Sentence and document level CEFR does not interact well. Table 4 reports CEFR prediction performance for models trained jointly on sentence- and document-level samples, evaluated separately on the sentence- and document-level subsets of the dataset. For sentence-level prediction, performance is substantially lower than in the single-granularity setting. Also for document-level prediction, multi-task training degraded the macro-F1 scores for most cases (only LLaMA3-8B and Qwen2.5-7B gained improvements) and does not show consistent improvements over document-only training. Overall, the results suggest that joint sentence–document training offers limited advantages over single-granularity settings. This is consistent with [3] discussed that document and sentence CEFR holds different challenges.

Model Type		Sentence	Document
Encoder (Base)	BERT-base	0.39	0.76
	RoBERTa-base	0.36	0.67
	DistilBERT-base	0.40	0.69
	ALBERT-base-v2	0.37	0.71
Encoder (Large)	BERT-large	0.34	0.66
	RoBERTa-large	0.42	0.77
Decoder (1–2B)	TinyLlama-1.1B	0.37	0.77
	Gemma-3-1B	0.36	0.77
	OLMo-2-1B	0.33	0.77
Decoder (7–8B)	LLaMA3-8B	0.25	0.84
	Mistral-7B	0.38	0.74
	Qwen2.5-7B	0.39	0.82
LLM2Vec-based	LLM2Vec-8B	0.34	0.64

Table 4: CEFR prediction performance (Macro-F1) trained under the multi-task learning using the combined dataset and tested on the document and sentence portion of the combined test set.

5 Conclusion

We compared encoder-based, decoder-only, and LLM2Vec-based models for CEFR level prediction on sentence- and document-level inputs. Our experiment results revealed that, across both granularities, decoder-only and LLM2Vec-based models achieve performance comparable to base encoder models, with no consistent architectural advantage observed. Sentence-level CEFR assessment is more challenging than document-level assessment, while document-level assessment still requires improvement for classroom use. And our results further show that joint sentence–document training does not consistently outperform single-granularity baselines.

Future work will explore regression and ranking methods to better capture the ordinal nature of CEFR levels.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP25K03233. This study was carried out using the TSUBAME4.0 supercomputer at Institute of Science Tokyo.

References

[1] Guanlin Li, Yuki Arase, and Noel Crespi. Aligning sentence simplification with ESL learner's proficiency for language acquisition. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, **Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 492–507, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[2] Alison Chi, Li-Kuang Chen, Yi-Chen Chang, Shu-Hui Lee, and Jason S. Chang. Learning to paraphrase sentences to different complexity levels. **Transactions of the Association for Computational Linguistics**, Vol. 11, pp. 1332–1354, 2023.

[3] 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.

[4] Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 12230–12266, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[5] Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Joshua Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. Universal-CEFR: 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. 9703–9755, Suzhou, China, November 2025. Association for Computational Linguistics.

[6] Elma Kerz, Daniel Wiechmann, Yu Qiao, Emma Tseng, and Marcus Ströbel. Automated classification of written proficiency levels on the CEFR-scale through complexity contours and RNNs. In Jill Burstein, Andrea Horbach, Ekaterina Kochmar, Ronja Laarmann-Quante, Claudia Leacock, Nitin Madnani, Ildikó Pilán, Helen Yannakoudakis, and Torsten Zesch, editors, **Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 199–209, Online, April 2021. Association for Computational Linguistics.

[7] Veronica Juliana Schmalz and Alessio Brutti. Automatic assessment of English CEFR levels using BERT embeddings. In Elisabetta Fersini, Marco Passarotti, and Viviana Patti, editors, **Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021)**, pp. 295–301, Milan, Italy, June 2021. CEUR Workshop Proceedings.

[8] Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. LLM2Vec: Large language models are secretly powerful text encoders. In **First Conference on Language Modeling**, 2024.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[11] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[12] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In **8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020**. OpenReview.net, 2020.

[13] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model, 2024.

[14] Gemma Team. Gemma 3. 2025.

[15] Team OLMo. 2 OLMo 2 furious (COLM's version). In **Second Conference on Language Modeling**, 2025.

[16] AI @ Meta Llama Team. The llama 3 herd of models, 2024.

[17] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[18] 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.

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In **International Conference on Learning Representations**, 2022.

A Prompt for CEFR Prediction

Figure 1 shows the prompt used for CEFR prediction with decoder-only models. The same prompt template was used across all models to ensure a fair comparison.

```
[SYSTEM]
Rate the following sentence on its
    readability level. The readability is
    defined as
the cognitive load required to understand
    the meaning of the sentence. Rate the
readability on a scale from very easy to
    very hard. Base your scores on the CEFR
scale for L2 learners. You should use the
    following key:

1 = Can understand very short, simple
    texts a single phrase at a time,
    picking up
    familiar names, words and basic
        phrases and rereading as required.
2 = Can understand short, simple texts on
    familiar matters of a concrete type.
3 = Can read straightforward factual texts
     on subjects related to his/her field
    and interest with a satisfactory level
         of comprehension.
4 = Can read with a large degree of
    independence, adapting style and speed
    of
    reading to different texts and purpose
        .
5 = Can understand in detail lengthy,
    complex texts, whether or not they
    relate to
    his/her own area of speciality,
        provided he/she can reread
        difficult sections.
6 = Can understand and interpret
    critically virtually all forms of the
    written
    language including abstract,
        structurally complex, or highly
        colloquial
    literary and non-literary writings.

[USER]
<sentence>

[ASSISTANT]
Given the above key, the readability of
    the sentence is (scale=1-6):
```

Figure 1: Prompt used for sentence-level CEFR prediction.