Korean Journal of English Language and Linguistics, **Vol 26, May 2026, pp. 606-629**
DOI: 10.15738/kjell.26..202605.606

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

Classifying TED-Ed Texts by CEFR-Based Analytic Scale: Human Judgment, LLM Prompting, and Stability Analysis

Minji Kim (Northern Arizona University) · **Sumi Han** (Hallym University)

Received: February 19, 2026
Revised: March 21, 2026
Accepted: April 9, 2026

Kim, Minji (First author)
Graduate Student
Department of English
Northern Arizona University
Email: mk2729@nau.edu

Han, Sumi (Corresponding author)
Associate Professor, Digital Arts and Humanities/English Language & Learning Department
Hallym University
1 Hallymdaehak-gil, Chuncheon-si, Gangwon-do, Republic of Korea
Tel: +82-33-248-1532
Email: sumihan@hallym.ac.kr

ABSTRACT

Kim, Minji and Sumi Han. 2026. Classifying TED-Ed texts by CEFR-based analytic scale: Human judgment, LLM prompting, and stability analysis. *Korean Journal of English Language and Linguistics 26*, 606-629.

This study examines how large language model (LLM) prompting strategies align with human CEFR classifications of TED-Ed transcripts, educational video materials from the TED-Ed platform, for Korean learners of English as a foreign language (EFL). Two trained raters evaluated 321 texts using a CEFR-based analytic framework tailored to the Korean EFL context, incorporating vocabulary, syntax, discourse organization, cognitive demands, and cultural concepts. Most texts were rated at B2–C1, indicating upper-intermediate to advanced reading demands. In contrast, traditional readability indices, including Lexile and Oxford 3000/5000-based measures, placed most texts at A2–B1, reflecting their reliance on lexico-syntactic features rather than discourse, conceptual, and cultural demands. A stratified subset of 71 texts was then classified by GPT-5 under three prompting conditions: zero-shot (no exemplars), fixed few-shot (constant exemplars), and randomized few-shot (varying exemplars across runs). Agreement with human ratings was $\kappa = .45$ for zero-shot prompting and $\kappa = .48$ for fixed few-shot prompting, while randomized few-shot prompting yielded κ values between .60 and .64 with higher variability across runs. Human inter-rater reliability was $\kappa = .76$, indicating that LLM classifications did not reach the consistency of trained raters. The findings suggest that LLM-assisted CEFR classification can support, but not replace, human judgment, and that prompting design affects both agreement and stability. The difference between human and LLM evaluation procedures, including the simplified implementation of the human rating framework in the prompt, may partly explain this gap.

KEYWORDS

CEFR, TED-Ed, LLM prompting, text classification, automated assessment

1. Introduction

TED-Ed has emerged as one of the most widely used open educational resources for English as a foreign language (EFL) teaching and learning. The platform offers concise, animated videos across science, history, culture, and technology that present academic content to general audiences in a comprehensible form (Kim and Han 2025). Prior studies have demonstrated that TED-Ed videos can enhance learners' listening, speaking, and vocabulary skills (Bhurt et al. 2023, Lu and Chen 2020, Nguyen 2023, Rashtchi et al. 2021). However, prior studies have relied on learner perception measures, readability indices (e.g., Lexile), and vocabulary profiling (e.g., frequency-based word lists). As a result, it remains unclear whether TED-Ed materials are appropriate for learners at different proficiency levels, for instructional planning and materials selection in EFL contexts.

The Common European Framework of Reference for Languages (CEFR) provides a principled basis for evaluating text difficulty. Its reading descriptors specify what learners at different proficiency levels can understand and process, distinguishing systematic differences in text length, conceptual density, discourse structure, and inferencing demands (Council of Europe 2001, 2020). Yet applying CEFR to instructional texts requires more than surface-level readability measures, which capture only lexical frequency and sentence length. For TED-Ed texts that combine academic concepts, explanatory discourse, and cultural references, human judgment is needed to assess discourse, conceptual, and cultural demands that influence comprehension difficulty for EFL learners. This is particularly important in contexts such as Korea, where learners' exposure, cultural familiarity, and learning trajectories differ from those of CEFR's original European context (Negishi et al. 2013, Jeon 2022, Zhao et al. 2017).

However, CEFR-based evaluation by trained human raters is time-intensive and difficult to scale. Recent advances in large language models (LLMs) offer a potential solution. LLMs have demonstrated strong performance in text classification, proficiency estimation, and language assessment tasks (Mizumoto and Eguchi 2023, Uchida and Negishi 2025), raising the possibility that they could approximate human CEFR judgments. Yet prior work has primarily applied LLMs to learner-generated writing or isolated sentences rather than to pre-planned instructional texts such as TED-Ed texts (Benedetto et al. 2025, Cooper 2025). Moreover, LLM-based classification is sensitive to prompt design such as zero-shot (classification without examples) and few-shot prompting (classification with a small number of labeled examples), and the selection and ordering of examples can substantially affect classification outcomes (Brown et al. 2020, Chen et al. 2024, Yoshida 2024). Whether LLMs can reliably replicate human CEFR classifications of instructional texts, and under what prompting conditions, remains an open question.

To address this gap, the present study makes two contributions. First, it develops a CEFR-based analytic framework tailored to Korean learners of English as a foreign language. The framework extends existing CEFR-based text evaluation approaches by incorporating discourse organization, conceptual demands, and cultural–conceptual accessibility in addition to vocabulary and syntax. Second, the study evaluates how different prompting configurations affect agreement between LLM classifications and human CEFR ratings under controlled conditions. By examining not only mean agreement but also variability under different prompting conditions, the study clarifies the methodological constraints and practical potential of using LLMs to support CEFR-based text evaluation in EFL contexts.

2. Literature Review

2.1 TED-Ed as an Instructional Resource in EFL Contexts

TED-Ed (https://ed.ted.com/) is an open-access educational platform that provides brief animated instructional videos covering disciplines such as the humanities, natural sciences, social sciences, and technology. Created in collaboration with subject-area specialists, the videos combine spoken narration with animation, typographic cues, and visual design. This multimodal approach supports the presentation of academic content for online learners (Liu 2023a, 2023b, Song and Nah 2017, Xia 2023).

From a discourse perspective, TED-Ed draws on features of both spoken and written language. Analyses of the TED-Ed corpus indicate that the texts are informational and non-narrative in nature, with a moderate degree of explicitness, although linguistic characteristics vary across subject areas (Kim and Han 2025). While TED-Ed videos are delivered orally, their texts show relatively high lexical density, compact information structure, and cohesive syntactic organization, which are more typical of written academic texts. At the same time, the texts include features associated with spoken discourse, such as direct references to the audience, sequential clause linking, and discourse markers used to guide attention. These patterns reflect TED-Ed's instructional purpose of presenting complex academic content to non-specialist audiences in a concise and comprehensible manner (Yoon 2023).

In EFL contexts, TED-Ed has been widely used for vocabulary development, listening comprehension, content learning, and learner autonomy (Anggraeni and Indriani 2018a, 2018b, Bhurt et al. 2023, Hung 2015, Rashtchi et al. 2021). Empirical studies have examined its effectiveness across English for Academic Purposes and English for Specific Purposes instruction. Liu (2023b) reported that science and technology TED-Ed animations contain levels of specialized vocabulary comparable to those found in written scientific texts, suggesting their applicability to discipline-specific instruction. Anggraeni and Indriani (2018a) found that TED-Ed encourages active engagement through both guided classroom tasks and independent learning activities.

Despite this growing body of research, existing studies have primarily focused on learner perception measures, such as self-reported difficulty ratings and comprehension judgments, or on specific linguistic features, including vocabulary frequency, sentence length, and syntactic complexity (Anandha et al. 2024, Asghar et al. 2023, Kirana 2023, Liu 2023b, Utami et al. 2024). These approaches provide partial indicators of text difficulty but do not apply a consistent set of criteria across texts. As a result, it remains unclear whether TED-Ed materials are appropriate for learners at different proficiency levels. No study has evaluated TED-Ed texts using a proficiency framework with defined rating criteria, such as the CEFR.

2.2 The CEFR and Its Application in EFL Contexts

CEFR provides an internationally established framework for describing language proficiency across listening, speaking, reading, and writing (Council of Europe 2001). While originally designed to describe learner proficiency, the CEFR reading descriptors also offer criteria for evaluating the demands that texts place on readers. Across levels A1–C2, the descriptors distinguish systematic differences in text length, conceptual density, discourse structure, inferencing demands, and reliance on contextual support (Council of Europe 2020). From this perspective, CEFR can serve as a principled basis for estimating whether instructional texts match learners' reading proficiency.

At the same time, the direct application of CEFR descriptors in Asian EFL contexts often requires contextual adaptation. This concern has motivated a growing body of research that examines how CEFR descriptors can be recalibrated to reflect local learning conditions and instructional practices. In Japan, Negishi et al. (2013) reported that learners' self-assessments broadly followed the CEFR proficiency progression but lacked sufficient differentiation at lower levels, leading to the development of the CEFR-J, which subdivides levels A1 to B2 and introduces a Pre-A1 category. Similarly, Zhao et al. (2017) aligned the China Standards of English with the CEFR and identified substantial overlap and inconsistency among intermediate levels, particularly in lexical development, indicating the need to recalibrate descriptors to reflect local developmental trajectories. In the Korean context, Jeon (2022, 2024) analyzed textbook tasks using CEFR mediation descriptors and found that existing descriptors did not adequately capture the interactional and cognitive demands of classroom communication, thereby underscoring the need for refined, context-sensitive evaluation criteria.

These studies demonstrate that CEFR-based evaluation in Asian EFL settings requires contextual refinement to ensure interpretability and validity. They further suggest that effective use of the CEFR must account for local curricular expectations, learner characteristics, and the nature of instructional input, rather than relying on direct descriptor transfer alone.

More importantly, CEFR-based text evaluation requires attention to dimensions beyond surface-level readability. Traditional readability indices such as Flesch-Kincaid readability tests (Flesch 1948, Kincaid et al. 1975), Lexile scores, or vocabulary coverage measures capture lexical frequency and sentence length but do not reflect discourse organization, conceptual density, or cultural knowledge demands. For instructional texts like TED-Ed texts, which combine academic content with explanatory discourse and cultural references, CEFR-based human evaluation provides a more comprehensive assessment of text difficulty than readability formulas alone. The challenge, however, is that human-based evaluation is resource-intensive and difficult to apply at scale.

2.3 Large Language Models for Predicting CEFR Levels

Recent advances in large language models such as ChatGPT (OpenAI), Claude (Anthropic), and Gemini (Google) have created new possibilities for automating CEFR-aligned text classification. Unlike earlier natural language processing approaches that relied on surface-level indices (Vajjala and Lõo 2014, Vajjala and Meurers 2014, Xia et al. 2016), contemporary LLMs can integrate lexical, syntactic, and discourse-level information within a single analytical framework (Bermejo et al. 2025, Huang et al. 2025, Tang et al. 2025). This capacity makes LLMs potentially suitable for evaluating the multi-dimensional text characteristics specified in CEFR reading descriptors. In particular, LLMs may be better positioned than conventional readability metrics or vocabulary-based indices to capture discourse-level properties such as inferential density, conceptual abstraction, and genre-specific organization which are central to CEFR-based text evaluation but difficult to quantify through surface-level measures alone.

Empirical studies have begun to explore LLM-based CEFR classification. Mizumoto and Eguchi (2023), and Uchida and Negishi (2025) reported promising results in proficiency estimation tasks, while Wolfer and Lew (2025) demonstrated that LLMs can assess text complexity in ways that align with established readability constructs. Using a large-scale learner writing corpus, Mizumoto and Eguchi (2023) reported that GPT-3–based scoring achieved usable reliability and accuracy relative to benchmark levels and that incorporating linguistic features further improved performance. Extending CEFR adaptation to an EFL-localized scale, Uchida and Negishi (2025) introduced a CEFR-J–aligned writing level analyzer that combines lexical metrics with AI-based analytic scores and reported strong correspondence with human ratings, supported by distributional and expert-validation

evidence. At the lexical-resource level, Wolfer and Lew (2025) demonstrated that CEFR-graded vocabulary lists can be expanded by imputing CEFR levels for uncovered items using word-level predictors, while still recommending human review for pedagogical appropriateness. These studies suggest that CEFR-aligned automation is feasible when constructs are explicitly operationalized and validated, but they also underscore the continued role of human researchers and the sensitivity of automated outcomes to methodological choices (e.g., prompting design, feature selection, and local scale alignment).

Yet, prior work has focused primarily on learner-generated writing or isolated sentences (Benedetto et al. 2025, Cooper 2025, Schmalz and Brutti 2021, Vajjala and Lučić 2018). It has not been systematically examined whether LLMs can reliably classify pre-planned instructional texts that present different discourse and conceptual characteristics than learner production. As discussed earlier, TED-Ed texts are characterized by pre-planned, informational, and rhetorically compressed discourse that combines academic concepts with explanatory instruction and cultural references. These characteristics pose distinct challenges for automated classification, as CEFR-based evaluation requires attention not only to linguistic form but also to the cognitive and interpretive demands imposed on readers.

A critical consideration in LLM-based classification is prompt design. Research on in-context learning has shown that prompting strategy substantially affects model outputs (Brown et al. 2020, Min et al. 2022, Wei et al. 2022). Zero-shot prompting requires models to classify texts based solely on task instructions, whereas few-shot prompting provides labeled examples to guide interpretation. The selection, ordering, and balance of examples can introduce systematic biases, including recency effects and sensitivity to class distribution (Chen et al. 2024, Yoshida 2024). These findings suggest that prompting strategy is not merely an implementation detail but a central methodological factor that determines the reliability and consistency of LLM-based CEFR classification. From an assessment perspective, prompt design directly influences how models interpret proficiency criteria and apply them to previously unannotated texts, affecting both the consistency and construct representation of automated classifications.

Given the potential of LLMs to support scalable text evaluation, a key empirical question is whether LLM classifications can approximate human CEFR judgments and how different prompting strategies affect the degree of agreement. Addressing this question requires comparing LLM outputs against a human reference standard under systematically varied prompting conditions. Although LLMs have demonstrated promise in CEFR-related prediction tasks, no prior research has systematically examined LLM-based CEFR classification of instructional texts such as TED-Ed texts, nor investigated how prompting strategies influence agreement with human ratings in this context. The present study addresses this gap by first establishing human CEFR ratings of TED-Ed texts using a CEFR-based framework tailored to Korean learners and then comparing these ratings with LLM classifications under three prompting conditions—zero-shot, fixed few-shot, and randomized few-shot (the set of examples varies across runs). This design allows for empirical examination of both the accuracy and stability of LLM-based classification relative to human judgments. The study is guided by the following research questions:

1. What is the distribution of CEFR levels assigned to TED-Ed texts based on human ratings?
2. How do different prompting strategies influence the level of agreement between LLM classifications and human CEFR ratings?

3. Methodology

3.1 Data Collection

TED-Ed texts were collected and analyzed in this study. The dataset consists of the texts obtained from the TED-Ed's official channel on YouTube (https://www.youtube.com/@TEDEd) and was used for research purposes. As shown in Figure 1, the dataset was pre-processed to ensure consistency and suitability for text-level evaluation prior to analysis. Texts were cleaned to remove non-linguistic elements and formatting artifacts, such as metadata tags and extraneous symbols. Subject classification followed the framework proposed by Kim and Han (2025), which aligns TED-Ed's original metadata with the International Standard Classification of Education (ISCED) developed by UNESCO Institute for Statistics (2015). After filtering and cleaning, the final TED-Ed corpus comprised 2,336 English-language texts.

The full dataset includes 321 texts, each evaluated by two trained human raters using a CEFR-based analytic scale. These ratings were used to address the first research question by examining the distribution of CEFR levels across the dataset. For the second research question, a subset of 71 texts was selected for direct comparison with LLM classifications. The same texts were evaluated under three prompting conditions (zero-shot, fixed few-shot, and randomized few-shot), and each condition was run multiple times to examine variation across trials.

The selection was stratified by disciplinary domain, including Arts & Humanities, Social Sciences, Education, STEM, and Health & Welfare, to maintain topical diversity. Within each domain, texts were sampled to cover a wide range of lengths and topics while avoiding duplicate or near-duplicate content. A subset of 71 texts from the human-rated sample was subsequently used for LLM-based classification to compare model outputs with human ratings under different prompting conditions.

Figure 1. Data Collection and Sampling Procedure

3.2 Traditional Readability Measure Rating

As a surface-level baseline comparator, texts were analyzed using the Lexile Framework for Reading (Wei and Vande Moere 2021) and lexical profiling based on the Oxford 3000/5000 word lists (Oxford Learner's Dictionary 2019). These measures were selected because they are widely used in educational assessment and can be interpreted in relation to CEFR levels, allowing comparison with prior research on text readability and proficiency alignment (Jeon 2022, Negishi et al. 2013, Zhao et al. 2017). Table 1 summarizes the resulting CEFR-referenced distributions from the Lexile Framework and the Oxford 3000/5000 word lists.

For CEFR referencing, Lexile-based readability estimates were mapped to CEFR bands using published concordance ranges (Wei and Vande Moere 2021) (e.g., A1 < 535L, A2 = 540–800L, B1 = 805–1090L, B2 = 1095–1320L, C1 = 1325–1460L, C2 ≥ 1465L). Oxford profiling assigned CEFR levels based on the tagged level

of matched lemmas in the Oxford 3000/5000 lists. Each matched lemma was converted to an ordinal score (A1 = 1 to C2 = 6), and texts were classified to the nearest CEFR band based on the average score. These procedures provide CEFR-referenced baseline classifications rather than direct estimates of discourse-level proficiency demands.

Because these indices operationalize difficulty primarily through sentence-length and word-frequency information, they were treated as supplementary reference points rather than primary indicators of text level. Readability analyses were conducted on all 2,336 TED-Ed texts. Lexile estimates placed most texts in the A2–B2 range, with the highest concentration at B1 (N = 1,132). Oxford lexical profiling showed an even stronger concentration in lower-frequency bands (predominantly A2, N = 1,970), with negligible coverage beyond the Oxford 5000 list. Taken together, these baselines suggest that TED-Ed texts appear accessible at approximately the A2–B1 range when evaluated through surface-level indicators alone. However, such indices provide limited coverage of discourse organization, conceptual density, and culturally mediated background knowledge demands that can shape comprehension in Korean EFL contexts. Accordingly, the study developed the CEFR-based analytic scale tailored to the Korean EFL context and applied it in the human rating procedure described in Section 3.3.

Table 1. Results of Readability Measurement

Readability Measure	A1	A2	B1	B2	C1	C2	Unclassified
Lexile	21	220	1,132	706	100	114	43
Oxford 3,000/5,000	3	1,970	363	0	0	0	0

3.3 Human CEFR Rating

To establish the human reference standard for Research Question 1 and to provide the basis for comparison in Research Question 2, selected TED-Ed texts were evaluated by human raters using a CEFR-based analytic scale adapted for the Korean EFL context.

3.3.1 CEFR-Based analytic scale tailored to the Korean EFL context

The analytic scale was developed through three sequential stages: (1) operationalizing CEFR descriptors into text-level rating criteria, (2) anchoring these criteria to Korean educational benchmarks, and (3) calibrating rater interpretations through norming sessions.

First, an eight-criterion analytic framework was established to capture multiple dimensions of reading difficulty relevant to EFL contexts (see Appendix A). The criteria encompassed vocabulary range, syntactic complexity, discourse organization, cultural and conceptual knowledge, cognitive load, rhetorical features, text length, and genre or register. Rather than treating these dimensions independently, raters evaluated them holistically when assigning a single overall CEFR level, allowing surface-level linguistic features to be weighed against deeper conceptual, rhetorical, and discourse-related demands. Second, each CEFR band was aligned with corresponding school stages in the Korean educational system (A1–A2 ≈ upper primary, B1–B2 ≈ lower–upper secondary, C1–C2 ≈ university). This alignment was informed by the 2022 national English curriculum, Ministry of Education (2022) approved textbook progressions, and the national college-entrance examination. Because no unified national curriculum exists at the tertiary level, university-level expectations were referenced using common institutional practices, such as placement standards and reading demands typical of academic coursework.

Finally, four norming sessions were conducted to establish shared interpretations of the analytic criteria. During these sessions, the draft scale was independently applied to sample TED-Ed texts, after which discrepancies were discussed and resolved. This process resulted in the formulation of an explicit hop-up adjustment rule (see Appendix B). When the conceptual, cultural, or rhetorical demands of a text clearly exceeded expectations based on its lexical and syntactic profile, raters raised the provisional CEFR level by half to one band. Typical triggers for this adjustment included dense references to culturally unfamiliar content, extended metaphor or satire, and sustained engagement with abstract or domain-specific concepts. This rule was incorporated into the finalized scale to ensure that level assignments reflected overall comprehension demands rather than surface linguistic features alone.

3.3.2 Rating procedure

The two researchers, both trained English-language teachers, served as independent human raters. Both held graduate-level qualifications in Applied Linguistics or English Education and had extensive teaching experience across primary, secondary, and university levels within the Korean EFL context. As former Korean EFL learners, they were familiar with national curricular sequencing, textbook content, and proficiency expectations across educational stages.

A subset of 321 texts selected from the TED-Ed corpus was evaluated using the adapted CEFR scale. Each rater independently assigned an overall proficiency level from A1 to C2 to each text. Prior to independent rating, the raters participated in three calibration sessions to review the finalized scale criteria and norming decisions established during scale development. In addition to assigning an overall CEFR level, raters recorded criterion-based notes aligned with the analytic scale (e.g., grammar, vocabulary range and precision, discourse organization, task demands). Rather than computing a mechanical average across criteria, raters made holistic level judgments informed by the integrated profile of linguistic, conceptual, and rhetorical features. When a text exhibited characteristics spanning adjacent bands, raters documented the rationale and considered whether the hop-up rule applied according to predefined scale guidelines. These structured notes supported calibration, discussion, and consistent adjudication decisions.

Following independent rating, cases of disagreement, most frequently occurring at adjacent CEFR boundaries (e.g., B2 vs. C1), were reviewed through discussion. For instance, texts explaining political systems (e.g., Athenian democracy) often combine relatively accessible sentence structures with dense informational load, discipline-specific terminology, and culture-dependent references (e.g., Greek civic institutions and transliterated terms), which can elevate discourse- and knowledge-based comprehension demands beyond what surface features alone suggest. In such cases, raters revisited the analytic criteria and, where applicable, applied the hop-up adjustment rule to resolve discrepancies, particularly when relatively accessible vocabulary and syntax co-occurred with high conceptual or rhetorical demands. Through this adjudication process, consensus ratings were established for all texts.

Inter-rater reliability was estimated using weighted Cohen’s kappa on the independent ratings prior to adjudication. The adjudicated consensus ratings then served as the human reference standard for subsequent comparisons with LLM-based classifications.

3.4 LLM-based Classification

GPT-5 (OpenAI, accessed September 2025) was used to classify a subset of 71 TED-Ed texts (sampled from the 321 human-rated texts) into CEFR levels under three prompting conditions. This flagship model was selected to provide a conservative test of LLM-assisted CEFR mapping by leveraging its improved reasoning consistency and instruction-following in complex classification tasks, thereby approximating the upper bound of model performance when benchmarking agreement against human CEFR judgments.

The LLM classifications were generated using GPT-5 through a local interface under the default decoding configuration. Parameters such as temperature and top-p were not modified by the user. In the randomized few-shot condition, a fixed random seed (42) was used to generate exemplar sequences. This setting controls the ordering of exemplars in the prompt but does not ensure identical model outputs across runs. To assess variability in model responses, each prompting condition was evaluated across multiple independent runs. The reported results therefore reflect both the effect of prompting design and variability introduced by the decoding process. GPT-5 was used to classify TED-Ed texts under three prompting configurations: zero-shot, fixed few-shot, and randomized few-shot. Agreement between GPT-5 classifications and human consensus ratings was then evaluated to examine how prompting strategy affects the extent to which automated CEFR labels align with expert judgment.

To evaluate stability across prompting conditions, the randomized few-shot configuration was executed multiple times using distinct exemplar sequences, and agreement statistics were calculated for each run. For the fixed few-shot condition, the same exemplar order was also repeated to examine run-to-run consistency under controlled input conditions. Mean κ values and corresponding standard deviations were reported to capture both agreement and variability. This procedure allows assessment of classification stability rather than relying on a single agreement estimate.

3.4.1 Prompting strategies

Across all prompting conditions, the model received an identical CEFR-based evaluation prompt so that any differences in classification outcomes could be attributed to prompting strategy rather than variation in task instructions themselves. Table 2 presents an excerpt from the base prompt used for GPT-based classification. Owing to space limitations, only the evaluation rubric portion of the full prompt is reproduced here. In the complete prompt, the model was instructed to act as a CEFR assessor and was provided with an explicit rubric aligned with the official CEFR can-do statements spanning levels A1 to C2.

The human rating procedure included a hop-up rule, which allowed raters to raise the assigned CEFR level when conceptual, cultural, or rhetorical demands exceeded the lexical and syntactic profile of the text. This rule was not directly implemented in the LLM prompt in its original form. Because the rule requires multi-dimensional judgment across discourse, conceptual load, and contextual knowledge, it was operationalized in a simplified manner through the nine analytic criteria provided in the prompt. Each criterion was phrased in a more explicit and localized form to ensure interpretability by the model. As a result, the LLM applied a reduced representation of the human rating framework rather than the full decision procedure used by human raters. The comparison between human and LLM classifications therefore reflects agreement between aligned but not identical evaluation procedures, and this difference should be considered when interpreting discrepancies in reliability. The evaluation rubric described proficiency characteristics at each level in terms of grammatical control, vocabulary range and precision, coherence and cohesion, and overall task fulfillment. Quantitative guidance (e.g., approximate error rates, indicative vocabulary ranges, and cohesion markers) was included to narrow the model's interpretive space and support consistent application of the criteria.

The analytic scale was grounded in CEFR reading descriptors. To operationalize text difficulty for LLM-based classification, the rubric also incorporated production-oriented descriptors (labeled *writing*), as these provide explicit, text-observable cues of linguistic complexity. These features were included not to assess writing performance, but to capture how text complexity is realized through lexical range, syntactic structure, and cohesion. In TED-Ed texts, these features increase systematically across proficiency levels. This integration reflects the need to evaluate not only the functional demands described in the CEFR, but also the linguistic forms through which those demands are instantiated in actual texts. Accordingly, the LLM was instructed to evaluate these observable features to estimate the CEFR level of the text for readers, while human ratings remained anchored in the CEFR-based analytic scale tailored to the Korean EFL context. We treat this as a pragmatic operationalization rather than a claim that the model assessed writing performance.

Table 2. Excerpt from the Base Prompt for GPT-based CEFR Classification

Type	Prompt
Persona	You are a CEFR assessor. You will identify the following aspects and evaluate the given text.
Evaluation Rubric	This rubric is designed to assess writing proficiency from CEFR level A1 to C2. Each level's "can-do" statements are based on the official CEFR Can-Do Statements (Council of Europe 2020), and the grammar and vocabulary descriptors are adapted from Macmillan Education resources. At the A1 level, the text shows very limited sentence structures with frequent spelling and grammar errors, but can still convey messages at the word level…At the C2 level, the text demonstrates complete mastery of grammar and syntax, with only negligible errors …
Specific Evaluation Criteria	Some of the above criteria are defined more specifically below. However, the following definitions are not absolute. At the A1 level, grammar allows for up to three errors per fifty words, vocabulary covers no more than fifteen types per fifty words, cohesion shows zero markers, the text is very short, and no contextual background is required. … At the C1–C2 level, grammar is error-free, vocabulary exceeds sixty types per fifty words, cohesion includes varied rhetorical devices, the text is long or more, and contextual competence includes free understanding of interdisciplinary, historical, and cultural contexts.
Instructions	Use the rubric above to evaluate the text at one of the CEFR levels (A1–C2). For each evaluation category (grammar and accuracy, vocabulary range and accuracy, coherence and cohesion, and task achievement), provide a score from 1 to 6 along with justification. Finally, determine the overall CEFR level.

In the zero-shot condition, the model was provided with only the evaluation prompt and the target text, with no labeled examples. By contrast, in the few-shot and randomized few-shot conditions, the prompt additionally included labeled exemplar texts for each CEFR level, as presented in Table 3. More specifically, each example text was accompanied by a CEFR level label assigned based on human ratings (e.g., CEFR level: A1). The example order remained constant across all trials (A1-A2-B2-B1-C1-C2). Lastly, in the randomized few-shot condition, the same exemplar texts were provided, but the order was shuffled across trials. This manipulation was implemented to examine the model's sensitivity to prompt sequencing, as reported in previous work on LLM-based evaluation (Chen et al. 2024, Yoshida 2024). Randomization was also used to counterbalance potential recency effects, whereby models may assign disproportionate weight to recently presented examples and to mitigate the impact of class imbalance, given that lower-level texts (A1–B1) were underrepresented in the dataset. In Table 3, example texts for each CEFR level were partially extracted from the original prompt due to space constraints. The original prompt configurations used in Tables 3 and 4 are provided in Appendix C.

Table 3. CEFR A1 and C2 Exemplars Used in Few-Shot and Randomized Few-Shot Prompting

CEFR Level	Exemplars
A1	I just forgot by mercer mare. Sometimes i remember and sometimes i just forget. This morning i remembered to brush my teeth, but i forgot to make my bed. I put my dishes in the sink after breakfast, but i forgot to put the milk away. …Did i forget to turn off the tub too? But there's one thing i never forget: i always remembered to have mom read me a bedtime story, and i always remember to kiss her good night you.
C2	Take an adjective such as "implacable," or a verb like "proliferate," or even another noun, "crony," and add a suffix, such as "-ity," or "-tion," or "-ism." You've created a new noun. "Implacability," "proliferation," "cronyism." Sounds impressive, right? Wrong! … A paragraph heavily populated by nominalizations will send your readers straight to sleep. Rescue them from the zombie apocalypse with vigorous verb-driven sentences that are concrete and clearly structured. You want your sentences to live, not to join the living dead.

3.4.2 Experimental design

From the full corpus of 321 TED-Ed texts, a subset of 71 texts was selected for LLM-based classification with three prompting strategies. This subset was not intended to be statistically representative of the entire TED-Ed repository but was purposefully stratified to cover all six CEFR levels (A1–C2) and include a range of disciplinary domains commonly used in EFL instruction. The primary goal of this sampling strategy was procedural validation, enabling controlled comparison across prompting conditions rather than population-level generalization. Each prompting strategy was executed across multiple independent trials to evaluate both the accuracy and stability of LLM classifications relative to human ratings. For initial analysis, three trials were conducted. Additional trials (up to five) were included in extended analysis to examine variability. Table 4 presents the initial prompting configurations used across three experimental trials.

In the zero-shot condition, no labeled examples were provided, so example order is not applicable (indicated by "—"). In the fixed few-shot condition, the same six CEFR-labeled exemplar texts were presented in an identical order across all trials (A1-A2-B2-B1-C1-C2), allowing assessment of baseline few-shot performance without ordering effects. In the randomized few-shot condition, the same exemplars were presented in a different shuffled order for each trial to examine sensitivity to example sequencing. The fixed order (A1-A2-B2-B1-C1-C2) was intentionally non-sequential, with lower levels (A1, A2) presented first, followed by intermediate and advanced levels in a mixed arrangement. This design placed underrepresented lower-level exemplars earlier in the prompt to counteract potential recency bias, whereby models may weight later examples more heavily.

The randomized orders varied this arrangement across trials. Trial 1 reversed the sequence (C2 first, A1 last), Trial 2 interspersed levels (A1-C2-B1-A2-B2-C1), and Trial 3 began with intermediate levels (B1-B2). By comparing κ values and their variability across these configurations, the study assessed whether example ordering systematically influenced classification agreement with human ratings. To assess the stability of model outputs under the randomized few-shot condition, we conducted four additional runs using the Trial 1 exemplar order (C2–C1–B1–B2–A2–A1) with identical settings. These runs were used to estimate within-configuration variability.

Table 4. Prompting Example Order for LLM Classification

Trial	Zero-shot	Fixed few-shot	Randomized few-shot
1	—	A1-A2-B2-B1-C1-C2	C2-C1-B1-B2-A2-A1
2	—	A1-A2-B2-B1-C1-C2	A1-C2-B1-A2-B2-C1
3	—	A1-A2-B2-B1-C1-C2	B1-B2-C2-C1-A2-A1

3.5 Data Analysis

The primary analysis compared LLM-generated CEFR classifications with human ratings to assess how prompting strategy affects agreement. Agreement was evaluated using weighted Cohen's kappa (Cohen 1968), which quantifies agreement between categorical ratings while correcting for chance. Following the interpretation, κ values of .21–.39 were interpreted as minimal agreement, .40–.59 as weak agreement, .60–.79 as moderate agreement, .80–.90 as strong agreement, and values above .90 as almost perfect agreement. Weighted κ is appropriate for ordinal proficiency scales such as the CEFR, as it assigns smaller penalties to disagreements between adjacent levels (e.g., B2 vs. C1) and larger penalties to more distant mismatches (e.g., A2 vs. C2). Quadratic weighting was applied to reflect the graded severity of disagreement across CEFR bands, consistent with practices in large-scale language assessment and rater-reliability research (Eckes 2019, North 2014).

For Research Question 1, Cohen's kappa (κ) was computed between the two human raters to establish inter-rater reliability and to validate the human ratings as a stable reference standard. For Research Question 2, the kappa was computed separately for each trial and prompting configuration to assess agreement between LLM classifications and human consensus ratings. The standard deviation of Cohen's kappa across repeated trials was calculated to assess the stability of model performance under different prompting conditions. Lower standard deviation values indicate greater consistency across trials, whereas higher variability reflects sensitivity to prompt configuration. Together, mean κ and its variability provide complementary indicators of agreement strength and classification stability, allowing systematic comparison of prompting strategies in terms of both accuracy and reliability.

4. Results and Discussion

This study examined how CEFR-based analytic scale tailored to the Korean EFL context and large language model–based prompting strategies can be applied to classify TED-Ed texts by proficiency level. Two main findings emerged. First, prompting strategy substantially influenced both the agreement and stability of LLM-based CEFR classification. Second, even under the most favorable prompting conditions, model performance remained below human inter-rater reliability, indicating that LLMs can assist but not replace expert judgment in text-level CEFR evaluation.

4.1 Human CEFR Ratings of TED-Ed Texts

A total of 321 texts from the TED-Ed corpus were independently evaluated by two trained raters using the CEFR-based analytic scale developed in this study. Inter-rater reliability for the initial independent ratings indicated moderate agreement (κ = .76). Disagreements occurred most frequently at adjacent CEFR boundaries (e.g., B2 vs. C1), reflecting the interpretive demands of CEFR-aligned text classification in which discourse complexity, conceptual density, and culture-dependent references can affect proficiency judgments. Disagreements were subsequently resolved through discussion and adjudication, and the resulting consensus ratings were used as the human reference standard for all subsequent analyses.

Table 5 presents the distribution of final consensus ratings across the CEFR levels. The distribution was heavily skewed toward upper-intermediate and advanced levels: 4 texts at A1, 6 at A2, 14 at B1, 100 at B2, 140 at C1, and

57 at C2. The concentration of texts at B2–C1 (75% of the sample) indicates that TED-Ed materials generally require upper-secondary to early-university level English proficiency for Korean EFL learners.

As described in Section 3.2, Traditional Readability Measure Rating, readability indices applied to the full corpus (N = 2,336) produced markedly lower estimates. Lexile scores placed most texts at A2–B2, while Oxford 3000/5000 coverage classified 85% of texts as A2. This divergence reflects the limitation of surface-level measures, which do not capture the discourse complexity, conceptual density, and cultural knowledge demands identified by human raters.

Table 5. CEFR-Level Distribution of TED-Ed Texts by Human Raters (N = 321)

CEFR Level	A1	A2	B1	B2	C1	C2
n (%)	4 (1)	6 (2)	14 (4)	100 (31)	140 (44)	57 (18)

The discrepancy between readability-based and CEFR-based human classifications represents a significant finding regarding text difficulty assessment. Readability indices based on the Lexile measurement and Oxford 3000/5000 word lists placed most TED-Ed texts in the A2–B2 range, reflecting short sentence structures and a high proportion of high-frequency vocabulary. At this surface level, the texts appear suitable for upper-beginner or lower-intermediate readers. Still, readability formulas capture difficulty primarily through lexical frequency and sentence length, and prior research has shown that such indices are calibrated mainly on native or near-native reading corpora (Crossley et al. 2011, Graesser et al. 2014). In the present study, these limitations were evident in texts, which combine syntactically simple sentences with abstract logical reasoning that imposes advanced inferential demands not reflected in readability scores.

Human ratings using the CEFR-based analytic scale tailored to the Korean EFL context yielded a markedly different profile. For the subset of 321 texts, most texts were classified in the B2–C1 range, with relatively few below B1. This distribution indicates that when discourse organization, conceptual load, and cultural references are considered, TED-Ed materials more closely resemble texts used at upper-secondary and early-university levels in the Korean curriculum. Reliance on surface-level readability measures alone therefore risks systematic underestimation of difficulty for Korean EFL learners.

These findings demonstrate the necessity of adapting CEFR descriptors for text-level evaluation in EFL contexts. The analytic scale and hop-up procedure used in this study integrated linguistic, cognitive, and cultural dimensions, allowing texts with intermediate lexical profiles but advanced conceptual or rhetorical demands to be more accurately classified. This pattern parallels findings from CEFR adaptation studies in Japan, China, and Korea (Jeon 2022, Negishi et al. 2013, Zhao et al. 2017), extending prior work from learner performance to instructional input.

4.2 LLM Classification and Agreement with Human Ratings

The remainder of this study focuses on evaluating whether LLM-based classification can approximate the human CEFR ratings. A stratified subset of 71 texts was classified by GPT-5 under the three prompting conditions, zero-shot, fixed few-shot, and randomized few-shot prompting. Agreement with human ratings was also assessed using weighted Cohen's kappa.

Table 6 presents the distribution of CEFR levels for the 71 TED-Ed texts included in the LLM analysis. As shown in the table, the distribution was highly skewed toward higher proficiency levels, with most texts falling

within B2 to C2, while lower levels (A1 to B1) were minimally represented. This class imbalance provides important context for interpreting agreement patterns across prompting strategies.

Table 6. Distribution of CEFR Levels in the LLM Sample (*n* = 71)

CEFR Level	A1	A2	B1	B2	C1	C2
n (%)	1 (1.4)	1 (1.4)	1 (1.4)	15 (21.1)	34 (47.9)	19 (26.8)

4.2.1 Agreement by prompting strategy

Table 7 summarizes agreement statistics across prompting conditions, and Figure 2 visualizes mean κ values and variability. Zero-shot prompting yielded κ = .45 (*SD* = .03), indicating weak agreement with human CEFR ratings. Fixed few-shot prompting showed only marginal improvement (κ = .48, *SD* = .05). Randomized few-shot prompting produced the highest mean agreement (κ = .64) but also substantially greater variability (*SD* = .20), indicating reduced stability across trials relative to the fixed conditions. The high variability observed in the randomized few-shot condition (*SD* = .20) should be interpreted with caution. In addition to sensitivity to exemplar selection and ordering, this variability may be influenced by the distribution of CEFR levels in the sample. The dataset includes a very limited number of lower-level texts (A1–B1), which constrains the range of possible classifications and increases the influence of individual exemplars in few-shot prompting. As a result, the observed variability cannot be attributed solely to the prompting strategy itself. It likely reflects a combined effect of prompt sensitivity and sample imbalance.

Trial-level estimates further illustrate this instability in the randomized condition. Across the three randomized exemplar orders, the first randomized order (C2–C1–B1–B2–A2–A1) yielded κ = .87, whereas the other randomized orders produced markedly lower agreement (κ = .46–.56). This pattern suggests that the elevated agreement observed in Trial 1 was contingent on a favorable exemplar configuration, particularly example ordering, rather than reflecting a consistent advantage of randomization. To evaluate within-configuration consistency, the Trial 1 exemplar order was repeated four additional times. Under this fixed order, agreement decreased and exhibited lower run-to-run variability, yielding a five-run mean κ = .60 (*SD* = .16). Collectively, these results indicate a trade-off: randomization can yield higher agreement in some runs, but it does so with reduced reproducibility, whereas using a constant exemplar set and order yields more consistent performance. However, this variability should be interpreted with caution. Given the highly imbalanced distribution of CEFR levels in the dataset, particularly the limited representation of A1–B1 texts, the observed instability is likely influenced by sampling effects.

Human inter-rater reliability (κ = .76) exceeded all LLM conditions, indicating that even the best-performing prompting configuration did not reach the level of agreement achieved by trained human raters. Thus, while LLM prompting strategies can approximate human CEFR ratings to some extent, agreement estimates are sensitive to prompt configuration, particularly under randomized few-shot prompting.

The observed sensitivity aligns with prior research showing that exemplar ordering and class balance can influence LLM classification outputs (Chen et al. 2024, Yoshida 2024). In the present study, the LLM subsample was highly imbalanced, with minimal representation at A1–B1 and a concentration at B2–C2. Under such conditions, variation in exemplar placement may disproportionately affect borderline decisions and, consequently, agreement estimates across runs. From an applied perspective, these findings raise concerns about the suitability of randomized few-shot prompting for use cases that require consistent CEFR classification outcomes (e.g.,

instructional placement or content sequencing). Fixed prompt templates and empirically validated exemplar sets may yield lower peak agreement than the best randomized run, but they offer greater reproducibility and clearer interpretability for practical deployment.

Table 7. Agreement Between GPT-5 and Human Ratings by Prompting Strategy

Prompting Strategy	Mean κ	*SD*	Interpretation
Zero-shot (3 trials)	.45	.03	Weak
Fixed few-shot (3 trials)	.48	.05	Weak
Randomized few-shot (3 trials)	.64	.20	Moderate
Randomized few-shot (5 trials)*	.60	.16	Moderate
Human inter-rater	.76	—	Moderate

** = Subsequent randomized few-shot with Trial 1 example order after the initial trial*

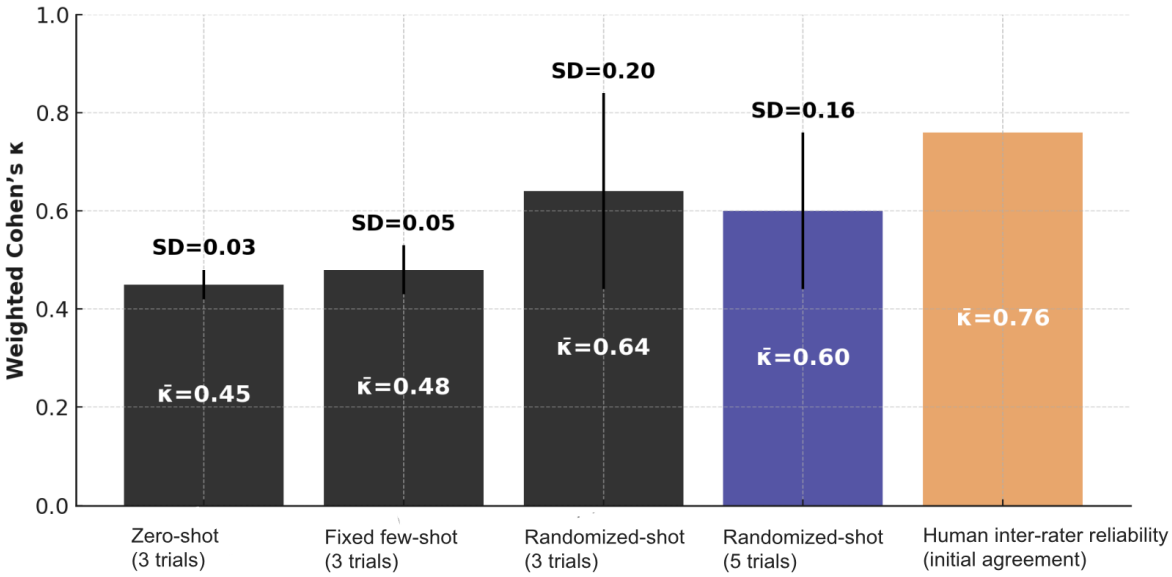


Figure 2. Mean Agreement Between Human Ratings and Prompting Strategies

4.2.2 Misclassification patterns from LLM classification

Misclassification patterns from GPT-5 classification showed systematic clustering rather than random dispersion. Most disagreements occurred between adjacent CEFR levels, particularly at the B2–C1 and C1–C2 boundaries, where distinctions rely on subtle differences in discourse coherence, conceptual abstraction, and cultural familiarity. In these boundary cases, GPT-5 frequently produced hedged classifications in its narrative justifications (e.g., C1, bordering on C2), indicating uncertainty when texts approached upper-level thresholds. Importantly, these misclassifications were concentrated near adjacent levels rather than spanning distant CEFR bands, suggesting that model outputs were sensitive to fine-grained proficiency distinctions even when exact level alignment with human ratings was not achieved.

Across all prompting conditions, human inter-rater reliability (κ = .76) exceeded model–human agreement, including the best-performing randomized configuration. This gap of approximately .12 in κ represents a

meaningful difference in classification consistency. Analysis of misclassification patterns showed that most disagreements occurred at adjacent CEFR levels, particularly at the B2–C1 and C1–C2 boundaries. These boundary regions depend on dimensions such as discourse coherence, conceptual abstraction, and cultural familiarity, all of which human raters could evaluate by drawing on curricular knowledge and contextual judgment. The frequent occurrence of adjacent-level or hedged classifications (e.g., B2 approaching C1) indicates that the model's predictions were not arbitrary but sensitive to fine-grained proficiency distinctions.

For example, Text 2218 (*Why meat is the best worst thing in the world*; Health & Welfare) was rated very highly across categories (Grammar 5/6, Vocabulary 6/6, Coherence 6/6, Task 6/6) and was described as displaying broad, precise, idiomatic vocabulary and a clear macro-structure supported by varied cohesive devices. At the same time, the justification noted minor surface issues (e.g., a sentence fragment and occasional punctuation irregularities) and suggested that register calibration would be needed for academic publication. This combination of near-maximal rhetorical control with minor accuracy/register caveats resulted in an overall classification of C1 (high), bordering on C2. A similar pattern appeared for Text 1244 (*The benefits of a bilingual brain*; Education), which was evaluated as generally well controlled but not fully C1-like in precision and consistency (Grammar 4/6, Vocabulary 5/6, Coherence 5/6, Task 5/6). The rationale emphasized strong organization and appropriate domain terminology, while also identifying small lexical/orthographic slips and mild register shifts that kept the text from full mastery, recommending added data/citations and more consistently academic cohesion devices to reach the next band. Accordingly, GPT-5 assigned B2 (upper range, approaching C1). Together, these cases show how GPT-5 used hedged, adjacent-level labels when the textual evidence supported a high level overall but retained localized features that, in the model's rationale, distinguished the upper boundary of one band from the lower boundary of the next.

At the same time, these results highlight persistent difficulty in operationalizing such dimensions through prompt-based, text-only evaluation. Methodologically, the findings support the use of LLMs as complementary tools rather than substitutes for expert judgment. LLMs can efficiently process large text repositories and perform initial proficiency screening, but human raters remain essential for adjudicating borderline cases and texts with high conceptual or cultural demands.

4.3 Implications for CEFR-Based Text Evaluation and LLM-Supported Mapping

The findings have two main implications for CEFR-based text evaluation. First, they demonstrate the necessity of adapting CEFR descriptors for text-level evaluation in EFL contexts. The analytic scale and hop-up procedure used in this study integrated linguistic, cognitive, and cultural dimensions, allowing texts with intermediate lexical profiles but advanced conceptual or rhetorical demands to be more accurately classified. This pattern parallels findings from CEFR adaptation studies in Japan, China, and Korea (Jeon 2022, Negishi et al. 2013, Zhao et al. 2017), extending prior work from learner performance to instructional input.

Second, the results show that prompting design is central to the reliability of LLM-based CEFR mapping. Example-based prompting can bring model predictions closer to human judgments, but such gains depend on careful control of example selection, ordering, and class balance. In addition, the human and LLM procedures were not fully equivalent. The human raters applied a hop-up rule that incorporated contextual and conceptual judgments, whereas the LLM prompt used a simplified set of criteria derived from this rule. This difference in implementation likely contributed to the observed discrepancy between human and model agreement. For applied use, LLM-based systems should rely on fixed, empirically validated prompts and report both mean agreement and variability across runs. Since LLM outputs are not strictly deterministic even under fixed decoding settings,

reporting a single agreement statistic may overstate classification stability. By implementing repeated runs and documenting variability across prompting configurations, the present study introduces a reproducibility-oriented evaluation protocol for LLM-assisted CEFR mapping. Such reporting practices are necessary if LLM-based classification is to be used in assessment-sensitive educational contexts.

It should be also noted that the LLM evaluation was conducted on a subset of 71 texts that was heavily concentrated in the upper proficiency bands (B2–C2). Such distributional imbalance may affect κ statistics, which are sensitive to class prevalence and boundary density. In this context, variability observed across randomized prompting conditions should be interpreted cautiously, as fluctuations may reflect the limited representation of lower-level texts in addition to prompt-order effects. Replication with larger and more balanced CEFR samples is therefore necessary to determine the stability of randomized prompting under more evenly distributed conditions.

Taken together, the combination of the CEFR-based analytic scale tailored to the Korean EFL context and controlled prompting strategies provides a basis for integrating LLMs into the evaluation of academically oriented digital materials. In hybrid workflows, LLMs can perform large-scale preliminary classification, while human experts focus on validation and refinement, particularly for texts near proficiency boundaries or with substantial conceptual and cultural load.

5. Conclusion

This study investigated how the CEFR-based analytic scale tailored to the Korean EFL context and large language model–based prompting strategies can be used to classify TED-Ed texts by proficiency level for Korean learners of English. While readability-based indices (Lexile and Oxford 3000/5000) placed most texts in the A2–B1 range, human ratings using the CEFR-based analytic scale tailored to the Korean EFL context located the majority of texts at B2–C1. This divergence suggests that lexical- and sentence-level indices may underestimate the discourse, conceptual, and cultural demands of TED-Ed materials for EFL readers. The findings further demonstrated that prompting design substantially shaped the extent to which GPT-5 approximated human classifications. Randomized few-shot prompting yielded the highest mean agreement with human ratings (mean $\kappa \approx .60$–$.64$ across conditions) but also greater variability across trials, whereas zero-shot and fixed few-shot prompts produced lower but more stable agreement. Overall, LLM-based CEFR classification approached, but did not reach, agreement levels observed between trained human raters (initial independent ratings $\kappa \approx .76$).

These results suggest several considerations for mapping texts to proficiency levels with LLMs. First, text classification should account not only for surface linguistic features but also for topic familiarity, background knowledge, and cultural accessibility, which shape EFL learners' comprehension. The CEFR-based analytic scale tailored to the Korean EFL context used in this study integrated vocabulary, syntax, discourse organization, cognitive load, and cultural–conceptual demands, providing a more realistic account of TED-Ed text difficulty than readability formulas alone. Second, prompt design and example sequencing are central components of the measurement procedure, as they influence the consistency with which an LLM applies CEFR-related criteria to new texts. By implementing repeated runs and reporting variability across prompting configurations, the present study underscores the importance of evaluating not only mean agreement but also classification stability in LLM-assisted text evaluation. Such reproducibility-oriented reporting is essential if automated CEFR mapping is to be applied in assessment-sensitive educational contexts.

For language pedagogy and assessment, the findings suggest that TED-Ed texts can function as level-appropriate resources for upper-secondary and early-university learners when classification relies on frameworks that capture

discourse and conceptual demands rather than readability scores alone. The study also offers a procedural template for using LLMs as a first-pass tool to screen large collections of educational texts, with human raters responsible for calibrating the framework, validating classifications, and adjudicating borderline cases. In this sense, LLM-supported CEFR mapping is most appropriately positioned as a support mechanism for text selection and curriculum design rather than as a stand-alone replacement for human judgment.

Several limitations should be considered when interpreting these findings. First, the dataset contained relatively few texts at lower CEFR levels (A1–B1), resulting in substantial class imbalance. Under this condition, few-shot prompting is highly sensitive to exemplar composition and ordering. Therefore, the high variability observed in the randomized few-shot condition (SD = .20) may reflect sampling bias rather than an inherent limitation of the prompting strategy itself. Second, the LLM analysis was conducted on a subset of 71 texts, whereas human ratings were available for the full dataset (N = 321). The reported agreement results are therefore restricted to this subset. Third, the LLM prompt implemented a simplified version of the human rating procedure. In particular, the full hop-up rule and Korean contextualization guidelines used by human raters were not directly incorporated into the prompt, as these procedures involve complex conditional judgments that are difficult to operationalize within a single prompt structure. Although the prompt reflects the core analytic criteria, this difference represents a methodological gap between human and model-based evaluation and likely contributes to discrepancies in reliability. Fourth, the analytic framework used in the LLM prompt combines CEFR reading descriptors with observable linguistic and discourse features (e.g., sentence complexity, lexical range, and text length) to better capture variation across TED-Ed texts. While this hybrid approach reflects systematic textual differences, it does not directly model reader comprehension processes. As a result, the construct being measured may only partially align with CEFR-defined reading proficiency. Fifth, decoding parameters such as temperature, top-p, and seed control were not explicitly managed and were determined by the local interface. Although multiple runs were conducted to account for variability, the lack of controlled decoding settings limits replicability and makes it difficult to isolate the effects of prompting design. Sixth, the few-shot exemplars included a mixture of genres (e.g., children's literature, advice texts, and informational writing) that do not fully align with the informational-expository register of TED-Ed texts. This mismatch may affect model generalization and contribute to instability across prompting conditions. Finally, the human reference standard was based on two trained raters, and the LLM evaluation was limited to a single model (GPT-5) under specific prompting conditions. In addition, the CEFR-adapted framework was developed within the Korean EFL context and applied only to TED-Ed materials. These constraints limit the generalizability of the findings.

Future research should address these limitations by using more balanced CEFR samples, controlling decoding settings, and examining the effects of exemplar selection under domain-matched conditions. Replication with the full dataset would allow more robust evaluation of LLM–human agreement. Comparative analyses across multiple LLM architectures and controlled prompting designs are also needed to determine the stability of classification outcomes. Further validation is required to examine whether the CEFR-based framework aligns with learner proficiency progression. This includes cross-validation with Korean EFL curricular materials and standardized textbook levels, as well as empirical linkage to learner comprehension data and placement benchmarks.

References

Anandha, A., D. Anggraheni and A. Yogatama. 2024. Students' perspective on the use of TED-Ed video on enhancing English comprehension. In *Proceedings of the English Language & Literature International Conference (ELLiC)*, 402-411.

Anggraeni, C. W. and L. Indriani. 2018a. TED-ED for autonomous listener. In *Proceedings of the 5th Asia Pacific Education Conference (AECON 2018)*, 18-22.

Anggraeni, C. W. and L. Indriani. 2018b. Teachers' perceptions toward TED-ED in listening class insight the era of disruptive technology. *Metathesis: Journal of English Language Literature and Teaching* 2(2), 222-235.

Asghar, R., A. Khan and M. Farooq. 2023. The role of TED-ED animations in enhancing the speaking fluency of undergraduate ESL learners in a Pakistani setting. *University of Chitral Journal of Linguistics and Literature* 7(1), 288-297.

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.

Bermejo, V. J., A. Gago, R. H. Gálvez and N. Harari. 2025. LLMs outperform outsourced human coders on complex textual analysis. *Scientific Reports* 15(1), 40122.

Bhurt, S. M., S. A. Memon and B. Bhurt. 2023. The impact of using TED-Ed as learning instrument on enhancing undergraduate ESL learners listening skill. *Orient Research Journal of Social Sciences* 8(2), 63-77.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter and D. Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 1877-1901.

Chen, X., R. Chi, X. Wang and D. Zhou. 2024. Premise order matters in reasoning with large language models. In *Proceedings of the 41st International Conference on Machine Learning*, 6596-6620.

Cohen, J. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 70(4), 213-220.

Cooper, C. 2025. Predicting the CEFR level of English listening texts with machine learning methods. *Research Methods in Applied Linguistics* 4(3), 1-16.

Council of Europe. 2001. *Common European Framework of Reference for Languages*. Available online at https://rm.coe.int/1680459f97

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Available online at https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

Crossley, S., D. Allen and D. McNamara. 2011. Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language* 23(1), 86-101.

Eckes, T. 2019. Many-facet Rasch measurement: Implications for rater-mediated language assessment. In V. Aryadoust and M. Raquel, eds., *Quantitative Data Analysis for Language Assessment*, 153-175. London: Routledge.

Flesch, R. 1948. A new readability yardstick. *Journal of Applied Psychology* 32(3), 221-233.

Graesser, A. C., H. Li and C. Forsyth. 2014. Learning by communicating in natural language with conversational agents. *Current Directions in Psychological Science* 23(5), 374-380.

Huang, Y., D. Li and A. Cheung. 2025. Evaluating the linguistic complexity of machine translation and LLMs for EFL/ESL applications: An entropy weight method. *Research Methods in Applied Linguistics* 4(3), 100229.

Hung, H. T. 2015. Flipping the classroom for English language learners to foster active learning. *Computer Assisted Language Learning* 28(1), 81-96.

Jeon, J-H. 2022. A systematic review of CEFR-related research of English education in South Korea. *Journal of Curriculum and Teaching* 11(8), 363-375.

Jeon, J-H. 2024. A textbook task analysis based on the CEFR mediation scale of basic user. *Primary English Education* 30(1), 91-115.

Kim, M. and S. Han. 2025. Register variation in TED-Ed videos: A multidimensional analysis across academic disciplines. *Korean Journal of English Language and Linguistics* 25, 1026-1047.

Kincaid, J. 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* (Res. Rep. No. 8-57). Orlando, FL: University of Central Florida, Institute for Simulation and Training.

Kirana, L. 2023. TED-Ed animated videos' impact on vocabulary gain of Indonesian EFL middle schoolers. *Research on English Language Teaching in Indonesia* 11(3), 39-45.

Liu, C. Y. 2023a. Suitability of TED-Ed animations for academic listening. *English for Specific Purposes* 72, 4-15.

Liu, C. Y. 2023b. Specialized vocabulary in TED talks and TED-Ed animations: Implications for learning English for science and technology. *Journal of English for Academic Purposes* 65, 101293.

Lu, K. and Q. Chen. 2020. A study on the learning effects of a blended listening and speaking course: A case study of medicine-related EFL learners. In *2020 Conference on Education, Language and Inter-cultural Communication (ELIC 2020)*, 97-102.

Min, S., X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi and L. Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11048-11064.

Ministry of Education. 2022, December. *2022 Revised National Curriculum for Primary, Secondary, and Special Schools Announced*. Available online at https://english.moe.go.kr/boardCnts/viewRenewal.do?boardID=265&boardSeq=93810&lev=0&statusYN=W&s=english&m=0201&opType=N

Mizumoto, A. and M. Eguchi. 2023. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics* 2(2), 1-13.

Negishi, M., T. Takada and Y. Tono. 2013. A progress report on the development of the CEFR-J. In E. D. Galaczi and C. J. Weir, eds., *Exploring Language Frameworks: Proceedings of the ALTE Kraków Conference*, 135-163. Cambridge, UK: Cambridge University Press.

Nguyen, C-D. 2023. TED Ed for incidental L2 academic vocabulary learning: A corpus-driven study. In B. L. Reynolds, ed., *Vocabulary Learning in the Wild*, 241-261. Singapore: Springer.

North, B. 2014. *English Profile Studies: The CEFR in Practice*. Cambridge, UK: Cambridge University Press.

Oxford Learner's Dictionary. 2019. *Oxford 3000 and 5000*. Oxford: Oxford University Press.

Rashtchi, M., B. Khoshnevisan and M. Shirvani. 2021. Integration of audiovisual input via TED-ED videos and language skills to enhance vocabulary learning. *MEXTESOL* 45(1), 1-18.

Schmalz, V. and A. Brutti. 2021. Automatic assessment of English CEFR levels using BERT embeddings. In *Proceedings of the Eighth Italian Conference on Computational Linguistics*, 293-299.

Song, J. Y. K. and K. Nah. 2017. Audio-visual elements of motion graphics in online educational video: Focused on Ted-Ed originals. *Journal of the Korean Society of Design Culture* 23(3), 452-461.

Tang, B., H. Liang, K. Jiang and X. Dong. 2025. On the importance of task complexity in evaluating LLM-based multi-agent systems. In *Proceedings of the NeurIPS 2025 Scaling Environments for Agents (SEA) Workshop*.

Uchida, S. and M. Negishi. 2025. Assigning CEFR-J levels to English learners' writing: An approach using lexical metrics and generative AI. *Research Methods in Applied Linguistics* 4(2), 1-14.

UNESCO Institute for Statistics. 2015. *International Standard Classification of Education: Fields of Education and Training 2013 (ISCED-F 2013)*. Montreal, Canada: UNESCO Institute for Statistics.

Utami, S., S. Noerjanah and N. Ibnus. 2024. The effectiveness of TED-ED videos on students' speaking skills in ninth grade of junior high school. *FLIP: Foreign Language Instruction Probe* 3(1), 1-10.

Vajjala, S. and K. Lõo. 2014. Automatic CEFR level prediction for Estonian learner text. In *Proceedings of The Third Workshop on NLP for Computer-assisted Language Learning*, 113-127.

Vajjala, S. and I. Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 297-304.

Vajjala, S. and D. Meurers. 2014. Readability assessment for text simplification: From analyzing documents to identifying sentential simplifications. *International Journal of Applied Linguistics* 165(2), 194-222.

Wei, J. and A. Vande Moere. 2021. *Aligning the Lexile® Framework for Reading to the Common European Framework of Reference*. Technical report. MetaMetrics. Available online at https://metametricsinc.com/wp-content/uploads/2018/07/Aligning-the-Lexile-Framework-to-theCEFR.pdf

Wei, J., X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, F. Xia, Q. Le and D. Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 24824-24837.

Wolfer, S. and R. Lew. 2025. Supplementing CEFR-graded vocabulary lists for language learners by leveraging information on dictionary views, corpus frequency, part-of-speech, and polysemy. *Humanities and Social Sciences Communications* 12(1), 1151.

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*, 12-22.

Xia, S. 2023. Explaining science to the non-specialist online audience: A multimodal genre analysis of TED talk videos. *English for Specific Purposes* 70, 70-85.

Yoon, S. 2023. Multimodality in online animated lectures: A case study of TED-Ed from a cognitive linguistic approach. *The Journal of Linguistic Science* 107, 545-570.

Yoshida, L. 2024. The impact of example selection in few-shot prompting on automated essay scoring using GPT models. In *Proceedings of the 25th International Conference on Artificial Intelligence in Education*, 61-73.

Zhao, W., B. Wang, D. Coniam and B. Xie. 2017. Calibrating the CEFR against the China standards of English for college English vocabulary education in China. *Language Testing in Asia* 7, 1-18.

Examples in: English
Applicable Languages: English
Applicable Level: Secondary/Tertiary

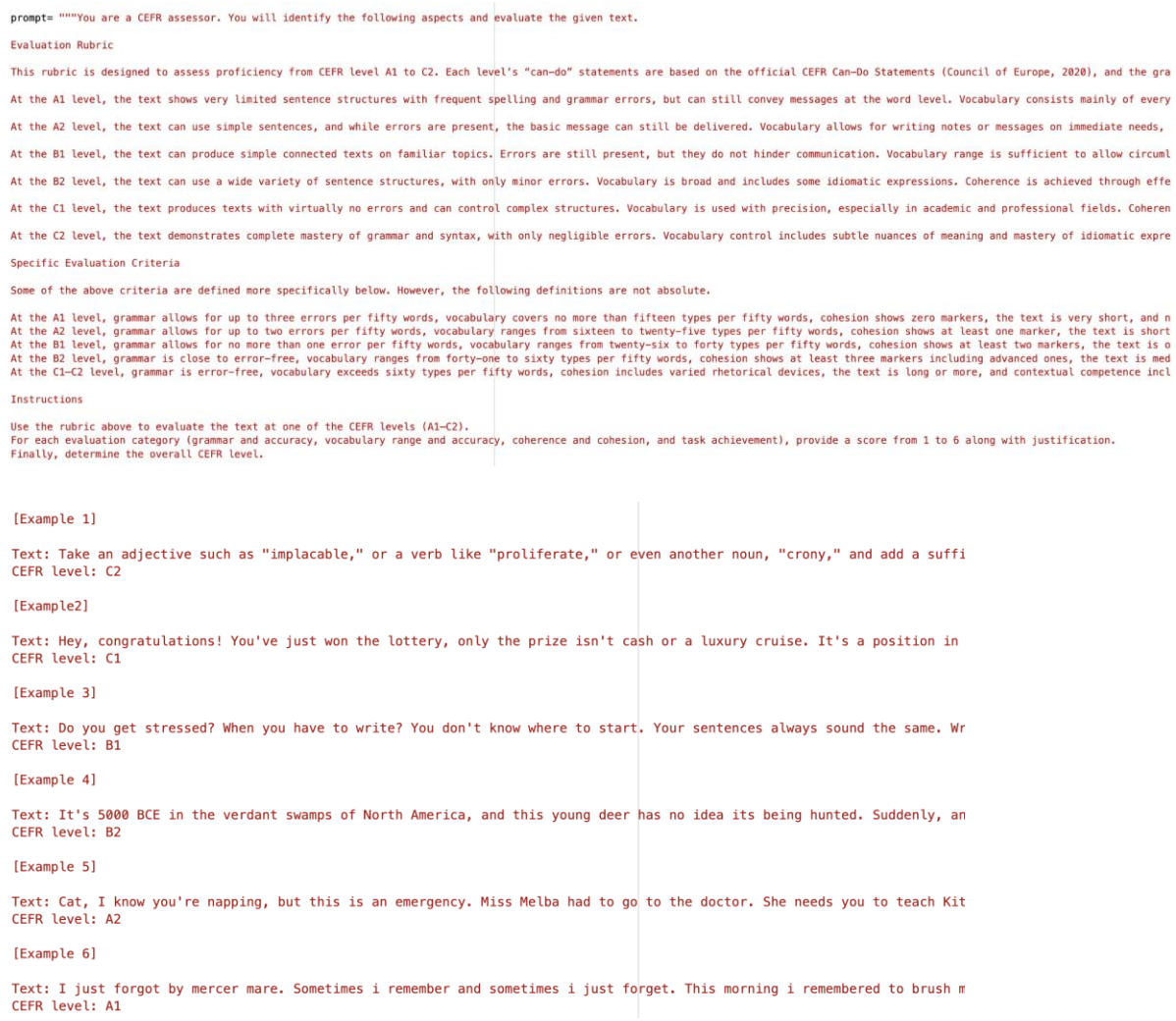
Appendix A. CEFR Bands for Korean EFL Learners (Evaluation Framework for TED-Ed Text)

Criterion	A1	A2	B1	B2	C1	C2
School Stage	Upper Primary (G4-5)	Lower or middle-Middle (G7)	High-proficiency MS / early HS or HS Year 1	HS Year 2 (exam track) / HS Year 3 (university entrance)	University	Senior undergraduate / Graduate
Vocabulary	Basic everyday/school words (e.g., food, school, family, weather)	Broader everyday lexicon, frequent emotion/ location/ frequency words	Concrete explanation/ process terms, topic word families, some academic labels	Academic/ domain vocabulary, some abstract terms, meaning disambiguation needed	High-register lexis, collocations/ Idioms, figurative use	Specialist conceptual terms, highly abstract/ technical vocabulary
Syntax	Simple/repeated simple sentences, basic tenses	Some conjunctions; simple past/future; negatives	Includes subordination varied tenses, if/because clauses	Diverse complex structures, coordination & subordination relative clauses, longer sentences	Long, dense sentences, insertions/ Contrasts, comparative structures	Non-linear/fragmented analytic structure, multi-layered cohesion
Discourse	Simple narration, sequential explanation	Time/ reason/ comparison links; simple cause–effect	Opinion–reason pattern; comparisons; examples	Topic development with examples & transitions; synthesis; some counter-moves	Logical architecture (claim–reason–evidence, counterargument, framing)	Advanced discourse: refutation, paradox/antithesis, intertextual weaving, multi-voiced argument
Cultural / Conceptual Knowledge	None required, 100% everyday	Basic cultural/ social knowledge (e.g., school, family)	General social knowledge helpful	Core cultural/ historical or academic background points (1–2 background jumps)	Specific disciplinary/ social-theory background, humanities/ culture knowledge	Extensive theoretical frames (Western & non-Western), without background, interpretation is difficult
Cognitive Load	Labeling/ stating/recall level	Feelings/experience linked to simple reasons	Opinion with basic justification; limited critique/ compare	Multiple viewpoints, analysis; compare/contrast / argument	Hypothesis & evaluation, abstraction, public reasoning	Philosophical/ metadiscourse, counter-factuals, paradox chains, irony/ synthesis
Rhetorical Features	None—plain statements	Repetition/ Emphasis, simple examples/comparisons	Examples/ comparison, simple rhetorical linking	Metaphor, emphasis, rhetorical questions; layered strategies	Allusion, stance-shifts, academic argumentative devices	Rich allusion/ irony/ metaphor, meta-commentary, dense citation & polyphony
Length	~100 words	120–180	200–350	400–600	600–900	900+
Genre / Style	Daily-life pieces, dialogue, short introductions / textbook snippets	Diary/ journal, short reports, brief informational texts	Article summary, column/ opinion, short expository prose	News/ reportage, review, feature; policy/ guideline style possible	Academic essay/ critique, complex argumentation/ discussion	Scholarly article, critical review/ synthesis across sources (e.g., news + analysis)

Appendix B. Dimensions and Hop-Up Adjustment Rules Used in the CEFR-Based Analytic Scale Tailored to the Korean EFL Context

Dimension (→ scored for every text)	Rationale for East-Asian Learners	Hop-Up Rule
● Linguistic Load ● Syntax depth ● Low-frequency lexis	Mirrors upper-secondary English-exam passages	Baseline CEFR band
● Cultural Familiarity ● Western religion, Greco-Roman myths, pop-culture unfamiliar in Korea	Cultural decoding requires extra schema building	+1 sub-level if ≥ 2 dense references +1 full band if references are essential for main idea
● Genre & Rhetoric ● Extended metaphor, satire, parody, court-style argument	Korean textbooks contain few such patterns → higher inference burden	+1 sub-level if one high-density device +1 full band if multiple intertwined devices
● Domain-specific Cognitive Load ● Abstract math, paradoxes, symbolic logic, advanced STEM	Even technically strong students learn these concepts in L1, rarely in L2	+1 sub-level if concept density comparable to senior-HS STEM text +1 full band if specialized graduate-level content
● Discourse Management ● Non-linear structure, nested frames, time jumps	Text-organization signals differ across languages; weak discourse cues hinder comprehension	+½ band when global cohesion markers are sparse
● Strategic Competence Needed ● Skimming for gist, schematic mapping, note-taking	Skills explicitly taught only from Grade 11; weak below that	No hop if skills aligned with Grade; +½ band if skills exceed average HS training
● Emotional / Philosophical Abstraction ● Existential ethics, aesthetic criticism, neuroscience of emotion	High affective-cognitive demand adds overload	+½ band if primary thrust rests on abstract reasoning

Appendix C. CEFR-Based Evaluation Prompt for LLM Classification (Screenshot of Local Prompt Interface)



```
prompt= """You are a CEFR assessor. You will identify the following aspects and evaluate the given text.

Evaluation Rubric

This rubric is designed to assess proficiency from CEFR level A1 to C2. Each level's "can-do" statements are based on the official CEFR Can-Do Statements (Council of Europe, 2020), and the gra

At the A1 level, the text shows very limited sentence structures with frequent spelling and grammar errors, but can still convey messages at the word level. Vocabulary consists mainly of every

At the A2 level, the text can use simple sentences, and while errors are present, the basic message can still be delivered. Vocabulary allows for writing notes or messages on immediate needs,

At the B1 level, the text can produce simple connected texts on familiar topics. Errors are still present, but they do not hinder communication. Vocabulary range is sufficient to allow circuml

At the B2 level, the text can use a wide variety of sentence structures, with only minor errors. Vocabulary is broad and includes some idiomatic expressions. Coherence is achieved through effe

At the C1 level, the text produces texts with virtually no errors and can control complex structures. Vocabulary is used with precision, especially in academic and professional fields. Coheren

At the C2 level, the text demonstrates complete mastery of grammar and syntax, with only negligible errors. Vocabulary control includes subtle nuances of meaning and mastery of idiomatic expre

Specific Evaluation Criteria

Some of the above criteria are defined more specifically below. However, the following definitions are not absolute.

At the A1 level, grammar allows for up to three errors per fifty words, vocabulary covers no more than fifteen types per fifty words, cohesion shows zero markers, the text is very short, and n
At the A2 level, grammar allows for up to two errors per fifty words, vocabulary ranges from sixteen to twenty-five types per fifty words, cohesion shows at least one marker, the text is short
At the B1 level, grammar allows for no more than one error per fifty words, vocabulary ranges from twenty-six to forty types per fifty words, cohesion shows at least two markers, the text is o
At the B2 level, grammar is close to error-free, vocabulary ranges from forty-one to sixty types per fifty words, cohesion shows at least three markers including advanced ones, the text is med
At the C1-C2 level, grammar is error-free, vocabulary exceeds sixty types per fifty words, cohesion includes varied rhetorical devices, the text is long or more, and contextual competence incl

Instructions

Use the rubric above to evaluate the text at one of the CEFR levels (A1-C2).
For each evaluation category (grammar and accuracy, vocabulary range and accuracy, coherence and cohesion, and task achievement), provide a score from 1 to 6 along with justification.
Finally, determine the overall CEFR level.

[Example 1]

Text: Take an adjective such as "implacable," or a verb like "proliferate," or even another noun, "crony," and add a suffi
CEFR level: C2

[Example2]

Text: Hey, congratulations! You've just won the lottery, only the prize isn't cash or a luxury cruise. It's a position in
CEFR level: C1

[Example 3]

Text: Do you get stressed? When you have to write? You don't know where to start. Your sentences always sound the same. Wr
CEFR level: B1

[Example 4]

Text: It's 5000 BCE in the verdant swamps of North America, and this young deer has no idea its being hunted. Suddenly, an
CEFR level: B2

[Example 5]

Text: Cat, I know you're napping, but this is an emergency. Miss Melba had to go to the doctor. She needs you to teach Kit
CEFR level: A2

[Example 6]

Text: I just forgot by mercer mare. Sometimes i remember and sometimes i just forget. This morning i remembered to brush m
CEFR level: A1
```

Note. The prompt is presented as a screenshot from the local interface used in this study. The visible area reflects screen settings, but the full prompt content is included without omission.

