Automatic CEFR Level Classification for Spanish Learner Texts: A Comparative Study of a Pretrained LLM and Prompt Engineering

Antonio Tamayo-Herrera[1,*,†], Juan Felipe Zuluaga-Molina[2,†]

[1]Universidad de Antioquia, 050010, Medellín, Colombia
[2]Institución universitaria ITM, 050034, Medellín, Colombia

Abstract

This paper, developed by the In2Lab-ITM research team, presents a comparative study between an open-access pretrained large language model for text classification and several few-shot prompt engineering approaches based on a commercial LLM that has reported strong performance across multiple natural language processing tasks. The experiments were conducted within the framework of the NivELE@IberLEF2026 shared task, which focuses on the classification of texts written in Spanish by language learners into five CEFR proficiency categories (A1, A2, B1, B2, and C1). The best-performing system achieved a Macro-averaged F1 of 0.57251, placing third in the competition, above six participating teams and showing competitive performance relative to the first- and second-ranked systems, which obtained Macro-averaged F1 of 0.60279 and 0.57606, respectively

Keywords

Automatic Spanish Proficiency Classification, Pretrained LLMs, Prompt Engineering, CEFR Spanish Level

1. Introduction

Over the last five years, growing interest in the automation of processes related to the assessment, teaching, and promotion of additional languages has been strongly shaped by the emergence of generative artificial intelligence [1]. Alongside the cautious findings regarding the incorporation of AI into educational environments, multiple technological and computational developments have outlined new educommunicative horizons and forms of digital educational culture, including the integration of augmented reality [2] and the increasing digitization and agency of human–machine interactions in language teaching and learning [3]. Nevertheless, in languages other than English, the pedagogical development of these technologies remains incipient, and there is still limited evidence concerning the necessity, viability, and implications of their integration into additional language and culture classrooms, particularly at a time when other educational contexts increasingly advocate for their restriction or prohibition [4].

One of the most rapidly developing technological domains concerns academic-administrative tasks such as assessment, student placement through proficiency testing [5, 6], and the automation of standardized examinations [7]. In many cases, these activities demand considerable time investment despite being highly automatable due to their recurrence, repetitiveness, and relative procedural regularity. Regarding the automatic classification of learner texts through proficiency-based assessments aimed at assigning a language level to learners' interlanguage production, a substantial body of research across multiple languages has focused on aspects such as lexical sophistication [8], grammatical and morphosyntactic features [9], syntactic complexity [10, 11, 12], and discourse-related dimensions including coherence and cohesion [13, 14, 15]. These approaches have contributed to the consolidation of a research area situated at the intersection of Natural Language Processing (NLP) and automated writing evaluation in second language contexts.

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
†These authors contributed equally.
antonio.tamayo@udea.edu.co (A. Tamayo-Herrera); juanzuluagam@itm.edu.co (J. F. Zuluaga-Molina)
0000-0002-5984-7463 (A. Tamayo-Herrera); 0000-0001-8751-4992 (J. F. Zuluaga-Molina)

In Spanish as an Additional Language (SAL) [16], research on these issues has taken diverse directions. Broadly speaking, advances have been made in the measurement of written proficiency using generative AI [17, 18, 19] and language models [20, 21], even in longitudinal studies [22]. Likewise, efforts have been devoted to automatic error identification [23] and the development of standardized assessments, including the automation of their evaluation procedures [7]. Collectively, these studies provide preliminary evidence regarding the interface between interlanguage analysis, automated error detection, proficiency-level classification, and language modeling within NLP-oriented frameworks.

Despite these recent advances, significant questions remain regarding the robustness, transferability, and generalization capacity of automatic proficiency classification models, particularly when general-purpose generative models are employed in highly specialized tasks such as the leveling of texts produced by SAL learners. Although some systems have reported outstanding performance on specific learner corpora, there is still limited evidence concerning the actual behavior of prompting-based models in multiclass classification tasks sensitive to lexical, syntactic, and discourse complexity phenomena. In this context, Large Language Models (LLMs) have increasingly been employed for automatic classification tasks due to their apparent capacity for contextual learning and generalization through few-shot prompting and in-context learning strategies. However, studies critically examining their performance in specialized linguistic assessment tasks in SAL remain scarce, especially when compared to models specifically trained for Common European Framework of Reference for Languages (CEFR) level classification.

Within this context, the shared task NivELE@IberLEF2026 [24, 17] emerges as a benchmark focused on the automatic classification of texts written by learners of Spanish as a foreign language according to CEFR proficiency levels (A1, A2, B1, B2, and C1). The task provides a particularly relevant scenario for evaluating the ability of different computational approaches to automatically identify linguistic complexity features associated with interlanguage development. Furthermore, it opens a space for discussing the actual suitability of general-purpose generative models in specialized linguistic assessment tasks, especially within multiclass settings and heterogeneous learner corpora.

Accordingly, this contribution, framed within the field of Natural Language Processing applied to language teaching and learning, analyzes the performance of different computational approaches for the automatic classification of texts produced by learners of Spanish as an Additional Language in the NivELE@IberLEF2026 shared task. More specifically, the study explores the behavior of specialized pretrained models and few-shot prompting strategies using GPT-4o-mini to discuss their scope, limitations, and generalization capacity in CEFR-based proficiency classification tasks. Preliminary results reveal significant differences between the approaches under examination. While certain specialized models achieve moderate performance, prompting-based strategies exhibit important limitations in capturing fine-grained linguistic complexity features in written interlanguage production, even when advanced generative models are employed.

2. State-of-the-Art

2.1. Automatic Language Proficiency Classification and Educational NLP

One of the most established areas within NLP applied to educational settings is automated writing evaluation and language proficiency classification. Automated Essay Scoring (AES) systems have traditionally been developed following two major approaches: models based on manually engineered linguistic features and neural models capable of learning complex representations automatically [23]. The former typically focus on lexical, syntactic, discourse-related, and accuracy-oriented indicators [25, 26], whereas the latter include deep learning architectures such as convolutional neural networks, recurrent neural networks, and, more recently, transformers and large language models [27, 28, 29].

Within this field, the automatic classification of proficiency levels according to the Common European Framework of Reference for Languages (CEFR) [6] has gained increasing relevance due to its applications in language teaching, certification, and student placement. Numerous studies have addressed the automatic prediction of A1–C2 proficiency levels through supervised learning over annotated learner

corpora, combining measures of linguistic complexity with statistical or neural models [30, 31, 32]. These studies suggest that lexical, syntactic, and discourse complexity can effectively model differences across proficiency levels, yielding relatively robust results across several languages, including Spanish [33].

In recent years, transformer-based approaches have increasingly displaced models relying exclusively on manually designed linguistic features [30, 34]. Models such as BERT [35], RoBERTa [35], and GPT [36] have achieved competitive performance in automatic proficiency classification tasks, particularly when large annotated corpora are available. Nevertheless, several studies have pointed out that the effectiveness of these models remains highly dependent on factors such as domain specificity, corpus size and homogeneity, and the particular characteristics of the assessment task itself.

With regard to Spanish as an Additional Language, although significant advances have been made in automated proficiency classification, learner corpus analysis, and automated assessment, the available evidence remains considerably more limited than in English and other European languages [17, 7, 20, 23]. Questions also remain concerning the actual generalization capacity of pretrained models and the transferability of systems trained on specific learner corpora to new CEFR classification scenarios.

2.2. Linguistic Complexity and Interlanguage in CEFR Classification

Research on automatic proficiency classification has consistently shown that CEFR levels can be modeled through multiple dimensions of linguistic complexity associated with the progressive development of learner interlanguage [37]. In this context, automated assessment systems typically operationalize observable linguistic phenomena through quantifiable indices related to syntactic complexity, lexical richness, discourse cohesion, verbal variation, and error patterns [10, 11, 12].

Syntactic complexity remains one of the most extensively studied dimensions. Previous studies have shown that measures such as mean sentence length, subordination, phrasal density, and nominal complexity can distinguish CEFR levels with relative consistency, particularly across lower and intermediate proficiency bands [10, 11, 12]. In parallel, lexical sophistication and vocabulary diversity have repeatedly been identified as strong predictors of language proficiency, especially through metrics related to lexical richness, lexical density, and low-frequency vocabulary use [14, 38, 39].

At the discourse level, more recent work has incorporated measures of cohesion and coherence based on the use of connectives, anaphoric mechanisms, lexical repetition, and thematic organization [13, 14, 15]. These dimensions make it possible to model not only the structural complexity of learner texts, but also the extent to which learners can establish increasingly sophisticated discourse relations as their interlanguage develops.

Linguistic errors have likewise been examined as relevant indicators of interlanguage development. Several studies suggest that certain error types — including agreement, tense usage, punctuation, and word order — tend to cluster around specific CEFR levels, supporting the interpretation of proficiency as a developmental continuum rather than as a set of fully discrete categories [39, 40].

Taken together, this body of research frames CEFR classification as a multidimensional problem involving lexical, syntactic, discourse-related, and pragmatic phenomena. At the same time, it provides a useful theoretical basis for operationalizing explicit linguistic features within automatic classification systems and prompting strategies guided by linguistic complexity criteria.

2.3. Large Language Models and Few-Shot Prompting in Educational NLP

The emergence of large language models (LLMs) has significantly reshaped the landscape of educational NLP. Unlike traditional approaches that relied on task-specific supervised training, LLMs can address a wide range of problems through in-context learning strategies such as zero-shot and few-shot prompting, partially reducing the dependence on large annotated corpora and extensive fine-tuning procedures [41, 42, 43].

In educational settings, these models have increasingly been applied to tasks involving automated assessment, feedback generation, open-ended response classification, learner profiling, and tutoring

support. Within writing assessment in particular, several studies have reported that models such as GPT can achieve reasonably high agreement with human raters in automatic scoring and proficiency classification tasks [44, 45, 46].

Recent literature, however, also emphasizes that LLM performance is highly sensitive to instruction design and to the quality of the examples included in prompts [43]. In this regard, strategies such as few-shot prompting, chain-of-thought prompting, and rubric-based prompting have shown notable improvements over simpler prompting configurations [42]. Some studies further suggest that incorporating explicit domain knowledge into prompts may improve model performance in specialized educational tasks [47].

Despite these advances, research specifically addressing CEFR classification for Spanish as an Additional Language through prompting-based approaches remains scarce. Likewise, there is still limited evidence regarding the extent to which LLMs can capture fine-grained phenomena related to linguistic complexity and interlanguage in multiclass proficiency classification tasks.

2.4. Linguistically Informed Prompting

Recent developments in prompt engineering have shifted prompting practices from relatively intuitive interactions toward more structured approaches guided by task-specific information. Broadly speaking, prompting can be understood as the systematic design of natural language instructions intended to shape the inferential behavior of large language models without modifying their internal parameters [48].

Within this line of research, several studies have examined how explicit linguistic properties of prompts — including discourse organization, lexical selection, syntactic structure, and task descriptors — affect model performance [49, 50]. These studies suggest that even small morphological, syntactic, or pragmatic variations in instruction design may lead to noticeable differences in model outputs.

In educational and language assessment tasks, some recent approaches have begun incorporating proficiency descriptors, analytic rubrics, and explicit linguistic criteria directly into prompts used with LLMs. The goal of these approaches is to direct model attention toward phenomena relevant to language assessment, such as syntactic complexity, lexical richness, discourse cohesion, or interlanguage-related error patterns [51].

Within CEFR classification contexts, such strategies appear particularly relevant because proficiency levels are not determined solely by thematic content, but also by multiple formal and discourse-related dimensions. Linguistically informed prompting therefore emerges as a potential way of integrating linguistic theory, proficiency descriptors, and assessment criteria into LLM-based systems [52].

At the same time, empirical evidence regarding the actual effectiveness of this type of prompting for automatic proficiency classification in Spanish as an Additional Language remains limited, especially when compared with specialized models trained on learner corpora.

3. Experiments

This section describes the four experiments conducted in the present study. The NivELE@IberLEF2026 shared task did not provide official training or validation datasets; instead, participants were given only a testing set consisting of 316 texts written in Spanish by language learners. The task consisted of assigning each text to one of the five proficiency levels of the Common European Framework of Reference for Languages (CEFR): A1, A2, B1, B2, or C1 [6].

3.1. Run 1

The first experiment consisted of performing the classification task using a pretrained model (https://huggingface.co/pymlex/roberta-spanish-cefr) specialized in CEFR level classification for Spanish available on Hugging Face (pymlex/roberta-spanish-cefr). This system is based on BERTIN [53], a

foundational Spanish language model built on the RoBERTa architecture (https://huggingface.co/project/bertin-roberta-base-spanish).

The model was originally trained on UniversalCEFR (https://huggingface.co/datasets/UniversalCEFR/caes_es), a learner corpus annotated according to CEFR levels containing approximately 31,100 text samples written by learners of Spanish. Since the model had already been optimized for language proficiency classification tasks in Spanish, it was selected as the main baseline for the present study.

3.2. Run 2

The second experiment implemented a few-shot prompt engineering strategy using GPT-4o-mini[1] through the API. The aim of this run was to explore whether a general-purpose LLM could compete with the specialized model used in run 1 through example-based contextual learning.

The experimental setup used the parameter $temperature = 0$ in order to minimize stochastic variation in the model outputs, together with $max\_tokens = 2$, restricting the response exclusively to the CEFR label. For the prompting strategy, 15 examples extracted from the UniversalCEFR corpus used to train the run 1 model were selected, with three examples per proficiency level. The complete prompt and a representative sample of the examples used are presented in Appendix A.1.

3.3. Run 3

The third experiment maintained the few-shot prompting strategy with GPT-4o-mini, as well as the same parameters used in run 2. However, in this case the examples extracted from UniversalCEFR were replaced with the five official examples provided on the competition website, one for each CEFR level.

The purpose of this experiment was to evaluate whether examples more closely aligned with the specific domain of the task could improve the performance of the generative model. The examples used are presented in Appendix A.2.

3.4. Run 4

The fourth experiment preserved the few-shot prompting strategy with GPT-4o-mini and the same parameters used in the previous runs but incorporated a linguistically enriched prompt written entirely in Spanish.

Unlike the previous experiments, this run included explicit instructions related to linguistic features associated with proficiency development in Spanish as a foreign or additional language and with CEFR proficiency levels. These instructions incorporated information related to syntactic complexity, verbal variation, subordination, discourse cohesion, lexical richness, lexical density, and error patterns, drawing on previous research on automatic proficiency classification and interlanguage analysis.

The purpose of this approach was to explore whether the explicit incorporation of linguistic criteria could guide the model's inferential behavior toward phenomena relevant to language proficiency classification. Table 1 summarizes the main linguistic features included in the prompt. A complete description of the instructions used is provided in Appendix A.3.

The source code corresponding to all experiments is available in a GitHub repository at: https://github.com/ajtamayoh/In2Lab-ITM-at-IberLEF-NivELE-2026.

4. Results & Analysis

Table 2 presents the official results obtained by each of the runs implemented in the present study within the NivELE@IberLEF2026 leaderboard [17]. Run 1, based on a pretrained model specialized in CEFR level classification, achieved a Macro-averaged F1 of 0.57251, placing the submitted system

[1]This language model was selected not only because of its widely recognized performance across different natural language processing tasks, but also due to its relatively low cost and accessibility compared to more recent GPT versions and other more expensive language models such as Gemini.

Table 1
Instructions used in the prompt for run 4.

Linguistic Feature	CEFR Indicator in Spanish as a Foreign/Additional Language	What to Observe in the Text	Suggested Computational Feature
Verb tenses	A1–A2: present / B1+: past tenses / B2+: wider variety	Use of past tense, subjunctive, conditional	Frequency by verb tense (count and proportion)
Verbal variation	A1: limited / B2+: broader repertoire	Number of distinct verb forms	Number of unique verb forms
Subjunctive	B1–B2 onward	Presence and correct usage	Percentage of verbs in the subjunctive
Sentence length	A1: shorter / B2+: longer	Number of words per sentence	Average sentence length
Subordination	B1+: subordinate clauses / B2+: more complex structures	Use of "que," "aunque," "cuando"	Number of subordinate clauses per sentence
Syntactic depth	B2–C1: embedded structures	Sentences with multiple levels of embedding	Syntactic tree depth
Basic connectors	A1–A2: y, pero, porque	Limited use	Frequency of basic connectors
Advanced connectors	B2+: sin embargo, por lo tanto	Discourse variation	Frequency and diversity of connectors
Lexical diversity	A1: low / B2+: high	Repetition vs. variation	Type-Token Ratio (TTR)
Sophisticated vocabulary	B2–C1	Use of abstract or low-frequency words	Frequency of rare words (based on corpus frequency)
Grammatical errors	A1–B1: frequent	Agreement, gender, verb-related errors	Errors per 100 words
Error types	Vary by level	Predominant error type	Error classification (gender, tense, etc.)
Cohesion	B1+: referential mechanisms	Pronouns, substitutions	Frequency of pronouns and anaphoric references
Lexical repetition	A1–A2: high	Word repetition	Lexical repetition ratio
Discourse structure	B2+: clearer organization	Introduction, development, conclusion	Structure detection (heuristic or model-based)
Text length	Generally increases with proficiency level	Total number of words	Total token count
Lexical density	B2+: higher density	More content words than function words	Ratio of lexical words to total words
Overall complexity	Increases across levels	Combination of multiple factors	Combined indices (e.g., syntactic + lexical complexity)

third among the nine participating teams. This performance remained relatively close to that of the second-ranked system (0.57606) and the competition winner (0.60279).

Runs 2, 3, and 4, which relied on few-shot prompting strategies with GPT-4o-mini, produced considerably lower scores. It should be noted that these results were calculated on approximately 50% of the testing dataset, according to the official competition settings. Nevertheless, the gap observed with respect to the specialized model is sufficiently large to reveal relevant tendencies regarding the behavior of the evaluated approaches.

Overall, the results suggest that the pretrained model used in run 1 was better able to capture patterns associated with linguistic complexity and interlanguage development in texts written by learners of

Table 2
Results obtained across the four experiments.

Run	Macro-averaged F1
Run 1	0.57
Run 2	0.31
Run 3	0.31
Run 4	0.24

Spanish. This finding is consistent with previous literature indicating that CEFR classification tasks depend heavily on syntactic, lexical, and discourse-related phenomena that are highly specific to both the domain and the training corpus. In this sense, the BERTIN-based model appears to benefit from having been previously fine-tuned on a large learner corpus annotated according to CEFR levels, which likely increased its sensitivity to linguistic patterns associated with each proficiency level. By contrast, the few-shot prompting strategies implemented with GPT-4o-mini showed important limitations in addressing the proposed classification task. Although the model used in runs 2, 3, and 4 corresponds to a contemporary LLM with competitive results across multiple NLP tasks, the experiments suggest that its generalization capacity through prompting alone was insufficient to adequately model gradual differences in language proficiency within written interlanguage texts.

The behavior of run 4 is particularly noteworthy. Unlike runs 2 and 3, this experiment incorporated explicit instructions related to linguistic features associated with CEFR levels, including syntactic complexity, verbal variation, subordination, discourse cohesion, lexical richness, and error patterns. These features were selected based on previous literature on automatic proficiency classification and interlanguage analysis. However, the explicit inclusion of this linguistic information did not improve the model's performance; on the contrary, run 4 obtained the lowest score among all experiments. A plausible explanation for this behavior is that the linguistically enriched prompt required GPT-4o-mini to simultaneously attend to a large number of heterogeneous linguistic indicators, many of which correspond to computational metrics rather than directly observable textual features. Since the model was not provided with these measurements but rather with textual descriptions of them, it may have attempted to approximate such indicators intuitively from the learner texts, potentially introducing noise into the classification process. In addition, the considerable length and complexity of the prompt may have diluted the relative importance of the few-shot examples and the primary classification objective, creating competing signals that hindered effective decision-making.

This result suggests that explicitly incorporating linguistic criteria into prompts does not necessarily guarantee that the model will operationalize such phenomena effectively in complex classification tasks. One possible explanation is that CEFR classification involves multidimensional interactions among linguistic, discourse-related, and pragmatic features that are difficult to model through relatively superficial instructions or limited examples. At the same time, the results seem to indicate that general-purpose LLMs, at least under relatively simple few-shot configurations, still struggle to capture fine-grained regularities associated with the progressive development of learner interlanguage.

Another relevant aspect concerns the difference between the performance originally reported for the pretrained model used in run 1 and the results obtained in the competition. Although the model achieved metrics close to 99% on the UniversalCEFR corpus, its score decreased considerably in the NivELE@IberLEF2026 setting [17, 24]. This behavior reinforces recent discussions surrounding transferability and robustness in automatic language proficiency classification models, particularly in scenarios involving differences across corpora, learner profiles, or annotation conditions. Consequently, the results obtained in this study suggest that even highly specialized models may be affected by domain shift phenomena when confronted with new datasets.

Finally, although the competition constraints —particularly the absence of gold labels for the testing set and the lack of official training and development splits— prevented detailed error analysis, the experiments nevertheless make it possible to identify relevant tendencies regarding current limitations of prompting in specialized language assessment tasks. Overall, the findings reinforce the idea that

automatic proficiency classification in Spanish as a foreign or additional language still depends heavily on specialized models trained on learner corpora and on representations sensitive to complex interlanguage and linguistic complexity phenomena. In addition, the study raises questions about the reliance on exclusively linguistic data without incorporating sociodemographic variables that could also inform classification processes, an aspect that may play an important role in real-world student placement decisions.

5. Conclusions

This study presented a comparative analysis between a specialized pretrained model and several few-shot prompt engineering strategies for the automatic classification of texts written in Spanish by language learners into five CEFR proficiency levels (A1, A2, B1, B2, and C1) within the framework of the NivELE@IberLEF2026 shared task [17, 24]. The system based on the pretrained model used in run 1 achieved a Macro-averaged F1 of 0.57251 and ranked third among the nine participating teams. Beyond demonstrating the competitiveness of the proposed system relative to the first- and second-ranked submissions, the experiments also made it possible to compare the behavior of specialized models and prompting-based strategies implemented with a contemporary general-purpose LLM.

The results show that the BERTIN-based specialized model substantially outperformed the few-shot prompting strategies implemented with GPT-4o-mini. This finding is particularly relevant given that the prompting strategies used in runs 2, 3, and 4 were designed on the basis of previous literature on automatic proficiency classification, linguistic complexity, and interlanguage analysis. Even the explicit incorporation of linguistic features associated with CEFR levels into the prompt failed to improve the performance of the generative model.

Taken together, the findings suggest that automatic proficiency classification tasks in Spanish as a foreign or additional language still require models capable of capturing fine-grained phenomena related to syntactic complexity, lexical richness, discourse cohesion, and interlanguage development. At the same time, the study highlights current limitations of few-shot prompting for highly specialized language assessment tasks when using general-purpose LLMs.

Finally, the results also contribute to ongoing discussions surrounding transferability and robustness. Although the model used in run 1 reported metrics close to 99% on the UniversalCEFR corpus, its performance decreased considerably within the competition setting, suggesting the presence of domain shift phenomena associated with differences across corpora and learner profiles.

6. Future Work

If the labels for the testing set are eventually released by the competition organizers, one of the main directions for future work will involve fine-tuning the pretrained model used in run 1 in order to evaluate whether domain-specific adjustment to the NivELE@IberLEF2026 dataset can improve the obtained metrics.

Future research could also explore the use of more recent and higher-capacity generative models, such as GPT-5 or advanced versions of Gemini, in order to examine whether more robust architectures are better able to capture complex interlanguage and CEFR-related phenomena through prompting strategies.

Another promising line of research involves further developing linguistically informed prompting strategies, particularly through structured prompts, chain-of-thought reasoning, and hybrid approaches combining explicit computational linguistic features with the inferential capacities of LLMs.

As stated in the Discussion session for run 4, future work should investigate the extent to which prompt complexity affects proficiency classification performance. In particular, future experiments could systematically compare short prompts, linguistically enriched prompts, and prompts augmented with automatically computed linguistic features.

Finally, future experiments should consider variables such as learners' first language, dialectal variation in Spanish, and differences across training corpora, as these factors could contribute to a better understanding of the transferability and generalization phenomena observed in automatic proficiency classification systems.

Declaration on Generative AI

During the preparation of this work, the authors used generative AI-assisted tools, including Consensus and Elicit, to support literature exploration and organization during the development of the state-of-the-art section (Generate literature review and Content enhancement). The authors also used ChatGPT (GPT-4) to assist with text translation from Spanish into English (Text Translation), language revision (Grammar and spelling check), paraphrasing and stylistic refinement (Paraphrase and reword; Improve writing style), and the revision of academic phrasing in selected sections of the manuscript. The use of these tools was limited to writing support, language assistance, and literature exploration. All methodological decisions, experimental designs, coding procedures, analyses, interpretations, and conclusions were developed, verified, reviewed, and approved exclusively by the authors. The authors take full responsibility for the final content of the manuscript.

References

[1] R. Baskara, M. Mukarto, Exploring the implications of chatgpt for language learning in higher education, Indonesian Journal of English Language Teaching and Applied Linguistics 7 (2023) 343–358.

[2] N. Punar Özçelik, G. Yangın Ekşi, M. H. Baturay, Augmented reality (ar) in language learning: A principled review of 2017–2021, Participatory Educational Research 9 (2022) 131–152. doi:10.17275/per.22.83.9.4.

[3] R. Godwin-Jones, Distributed agency in language learning and teaching through generative ai, Language Learning & Technology 28 (2024) 4–31. URL: https://hdl.handle.net/10125/73570.

[4] N. Selwyn, J. Aagaard, Banning mobile phones from classrooms—an opportunity to advance understandings of technology addiction, distraction and cyberbullying, British Journal of Educational Technology 52 (2021) 8–19. doi:10.1111/bjet.12943.

[5] Centro Virtual Cervantes, Diccionario de términos clave de ele, 1997–2026. URL: https://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/.

[6] Council of Europe, Marco común europeo de referencia para las lenguas: Aprendizaje, enseñanza, evaluación, Servicio de Publicaciones del Consejo de Europa, Strasbourg, 2001. URL: https://www.coe.int/lang-cefr.

[7] J. García Laborda, C. Navarro Laboulais, Automatizar los exámenes del diploma de español como lengua extranjera (dele): ¿utopía o realidad?, Revista de Lingüística Teórica y Aplicada 46 (2008) 81–93. doi:10.4067/S0718-48832008000200005.

[8] S. A. Crossley, T. Salsbury, D. S. McNamara, Predicting the proficiency level of language learners using lexical indices, Language Testing 29 (2012) 243–263. doi:10.1177/0265532211419161.

[9] J. Hancke, Automatic Prediction of CEFR Proficiency Levels Based on Linguistic Features of Learner Language, Master's thesis, Universität Tübingen, Tübingen, Germany, 2013. International Studies in Computational Linguistics, Seminar für Sprachwissenschaft.

[10] G. A. Khushik, A. Huhta, Investigating syntactic complexity in efl learners' writing across common european framework of reference levels a1, a2, and b1, Applied Linguistics 41 (2020) 506–532. doi:10.1093/applin/amy064.

[11] L. Ortega, Syntactic complexity measures and their relationship to l2 proficiency: A research synthesis of college-level l2 writing, Applied Linguistics 24 (2003) 492–518. doi:10.1093/applin/24.4.492.

[12] A. Alzahrani, Utility of kolmogorov complexity measures: Analysis of l2 groups and l1 backgrounds, PLOS ONE 19 (2024) e0301806. doi:10.1371/journal.pone.0301806.
[13] Y.-L. Lin, Y.-C. Tsai, C.-Y. C. Hsieh, Discourse competence across band scores: An analysis of speaking performance in the general english proficiency test, International Review of Applied Linguistics in Language Teaching 62 (2024) 1719–1746. doi:10.1515/iral-2022-0154.
[14] N. Iwashita, L. May, P. Moore, Features of Discourse and Lexical Richness at Different Performance Levels in the Aptis Speaking Test, Technical Report AR-G/2017/002, British Council, London, 2017. URL: https://www.britishcouncil.org/sites/default/files/iwashita_et_al_layout_1_revised.pdf.
[15] C.-N. Hsieh, Y. Wang, Speaking proficiency of young language students: A discourse-analytic study, Language Testing 36 (2019) 27–50. doi:10.1177/0265532217734240.
[16] J. F. Zuluaga, J. Gómez, Suplemento: Enseñanza, promoción y aprendizaje del español como lengua adicional en colombia, Lenguaje 52 (2024) e10014515. doi:10.25100/lenguaje.v52i2S.14515.
[17] M. V. Cantero Romero, J. Collado Montañez, I. Cabrera de Castro, J. Fruns Giménez, A. Arjonilla Sampedro, J. Cruz, A. Montejo-Ráez, S. M. Jiménez-Zafra, Resumen de nivele en iberlef 2026: Nivelación automática de textos de estudiantes de ele (español como lengua extranjera), Procesamiento del Lenguaje Natural 77 (2026). Forthcoming.
[18] M. V. Cantero-Romero, M. E. Vallecillo-Rodríguez, A. M. Ortiz-Colón, S. M. Jiménez-Zafra, Gpt-3.5 para la nivelación de ele en un entorno universitario: Un estudio comparativo de zero-shot learning y fine-tuning, Procesamiento del Lenguaje Natural 75 (2025) 53–62. doi:10.26342/2025-75-4.
[19] M. V. Cantero, Aproximación a un posible uso de chatgpt para nivelar la expresión escrita en ele, in: F. M. Sirignano, R. Martínez Roig, A. López Padrón (Eds.), Enseñanza y aprendizaje en la era digital desde la investigación y la innovación, Octaedro, Barcelona, 2024, pp. 55–64.
[20] M. D. Bustamante-Rodríguez, A. A. Piedrahita-Ospina, I. M. Ramírez-Velásquez, Modelo para detección automática de errores léxico-sintácticos en textos escritos en español, TecnoLógicas 21 (2018) 199–209. doi:10.22430/22565337.788.
[21] C. A. Martínez González, Determinación automática del grado de dominio de una lengua extranjera usando modelos de lenguaje, Master's thesis, Benemérita Universidad Autónoma de Puebla, Puebla, Mexico, 2025. Facultad de Ciencias de la Computación, Maestría en Ciencias de la Computación.
[22] A. Miaschi, D. Brunato, F. Dell'Orletta, An nlp-based stylometric approach for tracking the evolution of l1 written language competence, Journal of Writing Research 13 (2021) 71–105. doi:10.17239/jowr-2021.13.01.03.
[23] A. Ferreira, G. Kotz, Ele-tutor inteligente: Un analizador computacional para el tratamiento de errores gramaticales en español como lengua extranjera, Revista Signos 43 (2010) 211–236. doi:10.4067/S0718-09342010000200002.
[24] A. Bonet-Jover, J. González-Barba, L. Chiruzzo, Overview of iberlef 2026: Natural language processing challenges for spanish and other iberian languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS.org, 2026. Forthcoming.
[25] J. Xue, X. Tang, L. Zheng, A hierarchical bert-based transfer learning approach for multi-dimensional essay scoring, IEEE Access 9 (2021) 125403–125415. doi:10.1109/ACCESS.2021.3110683.
[26] H. M. Alawadh, T. Meraj, L. Aldosari, H. T. Rauf, An efficient text-mining framework of automatic essay grading using discourse macrostructural and statistical lexical features, SAGE Open 14 (2024) 21582440241300548. doi:10.1177/21582440241300548.
[27] F. Nadeem, H. Nguyen, Y. Liu, M. Ostendorf, Automated essay scoring with discourse-aware neural models, in: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Florence, Italy, 2019, pp. 484–493. doi:10.18653/v1/W19-4450.
[28] M. Beseiso, O. A. Alzubi, H. Rashaideh, A novel automated essay scoring approach for reliable higher educational assessments, Journal of Computing in Higher Education 33 (2021) 727–746. doi:10.1007/s12528-021-09283-1.
[29] C. M. Ormerod, A. Malhotra, A. Jafari, Automated essay scoring using efficient transformer-based

language models, arXiv preprint arXiv:2102.13136 (2021). doi:10.48550/arXiv.2102.13136.
[30] E. Kerz, D. Wiechmann, Y. Qiao, E. Tseng, M. Ströbel, Automated classification of written proficiency levels on the cefr-scale through complexity contours and rnns, in: Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, 2021, pp. 199–209. URL: https://aclanthology.org/2021.bea-1.21/.
[31] T. Gaillat, A. Simpkin, N. Ballier, B. Stearns, A. Sousa, M. Bouyé, M. Zarrouk, Predicting cefr levels in learners of english: The use of microsystem criterial features in a machine learning approach, ReCALL 34 (2022) 130–146. doi:10.1017/S095834402100029X.
[32] E. Volodina, I. Pilán, D. Alfter, Classification of swedish learner essays by cefr levels, in: S. Papadima-Sophocleous, L. Bradley, S. Thouësny (Eds.), CALL Communities and Culture—Short Papers from EUROCALL 2016, Research-publishing.net, Dublin, 2016, pp. 456–461. doi:10.14705/rpnet.2016.eurocall2016.606.
[33] A. Caines, P. Buttery, Reprolang 2020: Automatic proficiency scoring of czech, english, german, italian, and spanish learner essays, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 5614–5623. URL: https://aclanthology.org/2020.lrec-1.689/.
[34] J. Fields, K. Chovanec, P. Madiraju, A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe?, IEEE Access 12 (2024) 6518–6531. doi:10.1109/ACCESS.2024.3349952.
[35] V. J. Schmalz, A. Brutti, Automatic assessment of english cefr levels using bert embeddings, in: Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021), CEUR Workshop Proceedings, 2021, pp. 295–301. URL: https://aclanthology.org/2021.clicit-1.45/.
[36] A. Mizumoto, M. Eguchi, Exploring the potential of using an ai language model for automated essay scoring, Research Methods in Applied Linguistics 2 (2023) 100050. doi:10.1016/j.rmal.2023.100050.
[37] A. C. Lahuerta Martínez, The dynamics of syntactic and lexical complexity measures in academic writing, Ibérica 47 (2024) 251–274. doi:10.17398/2340-2784.47.251.
[38] B. Bulté, A. Housen, Conceptualizing and measuring short-term changes in l2 writing complexity, Journal of Second Language Writing 26 (2014) 42–65. doi:10.1016/j.jslw.2014.09.005.
[39] M. A. Meloni, Analisi dell'Interlingua e CEFR nell'Apprendimento della Lingua Inglese in studenti di scuole secondarie di secondo grado di Sassari, Ph.D. thesis, Università degli Studi di Sassari, Sassari, Italy, 2017. URL: https://iris.uniss.it/retrieve/e1dc1a2c-d623-1507-e053-3a05fe0ac7a3/Meloni__M_Analisi_Interlingua_e_CEFR.pdf.
[40] N. Ballier, T. Gaillat, A. Simpkin, B. Stearns, M. Bouyé, M. Zarrouk, A supervised learning model for the automatic assessment of language levels based on learner errors, in: M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou, J. Schneider (Eds.), Transforming Learning with Meaningful Technologies, volume 11722 of *Lecture Notes in Computer Science*, Springer, Cham, 2019, pp. 308–320. doi:10.1007/978-3-030-29736-7_23.
[41] L. Yan, L. Sha, L. Zhao, Y. Li, R. Martinez-Maldonado, G. Chen, X. Li, Y. Jin, D. Gašević, Practical and ethical challenges of large language models in education: A systematic scoping review, British Journal of Educational Technology 55 (2024) 90–112. doi:10.1111/bjet.13370.
[42] G.-G. Lee, E. Latif, X. Wu, N. Liu, X. Zhai, Applying large language models and chain-of-thought for automatic scoring, Computers and Education: Artificial Intelligence 6 (2024) 100213. doi:10.1016/j.caeai.2024.100213.
[43] S. Maity, A. Deroy, S. Sarkar, Exploring the capabilities of prompted large language models in educational and assessment applications, in: B. Paaßen, C. Demmans Epp (Eds.), Proceedings of the 17th International Conference on Educational Data Mining, International Educational Data Mining Society, Atlanta, GA, 2024, pp. 961–968. doi:10.5281/zenodo.12730013.
[44] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, G. Kasneci, Chatgpt for good? on opportunities and challenges of large language models for education, Learning and Individual

Differences 103 (2023) 102274. doi:10.1016/j.lindif.2023.102274.
[45] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, J. Ba, Large language models are human-level prompt engineers, in: Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 2023. URL: https://openreview.net/forum?id=92gvk82DE-.
[46] S. García-Méndez, F. de Arriba-Pérez, M. d. C. Somoza-López, A review on the use of large language models as virtual tutors, Science & Education 34 (2025) 877–892. doi:10.1007/s11191-024-00530-2.
[47] S. Sivarajkumar, M. Kelley, A. Samolyk-Mazzanti, S. Visweswaran, Y. Wang, An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study, JMIR Medical Informatics 12 (2024) e55318. doi:10.2196/55318.
[48] B. Chen, Z. Zhang, N. Langrené, S. Zhu, Unleashing the potential of prompt engineering for large language models, Patterns 6 (2025) 101260. doi:10.1016/j.patter.2025.101260.
[49] J. M. Imperial, A. Barayan, R. Stodden, R. Wilkens, R. M. Sánchez, L. Gao, M. Torgbi, D. Knight, G. Forey, R. R. Jablonkai, E. Kochmar, R. Reynolds, E. Ribeiro, H. Saggion, E. Volodina, S. Vajjala, T. François, F. Alva-Manchego, H. Madabushi, Universalcefr: Enabling open multilingual research on language proficiency assessment, arXiv preprint arXiv:2506.01419 (2025). doi:10.48550/arXiv.2506.01419.
[50] J. P. Wahle, T. Ruas, Y. Xu, B. Gipp, Paraphrase types elicit prompt engineering capabilities, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, FL, 2024, pp. 11004–11033. doi:10.18653/v1/2024.emnlp-main.617.
[51] S. DeVore, Exploring the ability of llms to classify written proficiency levels, Computer Speech & Language 90 (2025) 101745. doi:10.1016/j.csl.2024.101745.
[52] A. Leidinger, R. van Rooij, E. Shutova, The language of prompting: What linguistic properties make a prompt successful?, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 9210–9232. doi:10.18653/v1/2023.findings-emnlp.618.
[53] J. de la Rosa, E. G. Ponferrada, M. Romero, P. Villegas, P. González de Prado Salas, M. Grandury, Bertin: Efficient pre-training of a spanish language model using perplexity sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23. doi:10.26342/2022-68-1.

A. Prompts Used in the Experiments

This appendix presents the prompts used in the experiments described in Section 3. The prompts are reported verbatim to facilitate reproducibility.

A.1. Prompt for Run 2

This appendix presents the approach used in Run 2, which involves performing a few-shot prompt engineering strategy using GPT-4o-mini with 15 examples extracted from the UniversalCEFR corpus.

Prompt: {“role”: “system”, “content”: “”“You are an expert in assessing language proficiency according to the CEFR for students of Spanish as a foreign language. Assign only one of the following categories: “A1”, “A2”, “B1”, “B2”, or “C1”. Write the assigned category in all caps. Do not explain your decision.}

few_shot_examples = [{”text”: ”””Como estais ? ! Hola amigos ! ? Espero que todo esta bien ! ? Como eran vuestros vacaciones ? Este verano yo ha viajado mucho por el Europa . Era mi primera vez volando en el avion y cuando pensaba de este , me daba miedo . ! Pero pasaba mucho mejor que esperaba ! Ha ido con mi familia y amigos de mis padres . Hemos visitado Italia y Espana ! Fue increible ! Primer pais era Italia , alli hemos comido unos platos tipicos de Italia , eran ricisimos . Ademas , hemos visitado muchas museas diferentess y lugares historicas .”””, ”label”: ”A1”}, {”text”: ”””¿ Que tal ? un beso Espero que todo va bien por tu . Mi padre Rene esta mejor , no le duele mas la pierna Es pronto el

compleaño de mi hermano , ¡ 30 30+años ! madre mia vamos a hacer una fiesta y ! que fiesta ! Mi madre , Janette , trabaja siempre en el clinica , pero la clinica cierrada a el fin de esta año . Juntos viven siempre en el pueblo . Nadine”””, ”label”: ”A1”}, {”text”: ”””¿ Cómo estas chica ? Había regresado de México con mi madre . Fue muy divertido y en México había mucha sol . Mi madre y yo pasabamos mucho tiempo en la playa . ¡ Quiero regresar pronto ! La gente fue muy simpatica y pude practicar hablar en espaÑol casi todo el tiempo . Todos me ayudaron cuando no conocí las palabras si pude explicar las en íngles . Aprendía mucho durante mis vacaciones . También aprendí de los Mayans y todo que hayan en México . Ellos estaban muy moderna y inteligente .”””, ”label”: ”A1”}, {”text”: ”””Hola querida Sara , espero que estes muy bien y que todo en el trabajo vaya como tu quieras . No te vi en la universidad toda la semana y la verdad es que perdi tu numero de teléfono como muchos otros . Te estoy mandando este correo electrónico para pedir te una favor : es que la semana pasada , cuando iba a Marrakech para visitar a mi familia , tome mis notas de todos los cursos conmigo para trabajar un_poco , y en el camino de regreso , perdí mi equipaje en el avion . Es una situacion tan dificil y estoy muy avergonzada por pedir te eso pero eres la única persona con la que puedo contar . ¿ Puedes por_favor prestar me tus notas para fotocopiar las ? Esto me ayudaré mucho , no te puedes imaginar . Me puedes mandar un correo electrónico para decir me cuando y donde nos vemos (prefiero que sea en la universidad) o llamar me por teléfono (05554893) . Hasna .”””, ”label”: ”A2”}, {”text”: ”””Cuando llegué a el destino , le propuso que si se podía viajar en Tokyo juntos , y ella me dijo que sí . ¡ Era un momento hermoso para mí ! Nosotros paseamos por las calles en Shinjuku (Un lugar en Tokyo) , compramos unas cosas en las tiendas y tomamos varios tipos de las comidas japonesas . Nos discutimos mucho y me llevé muy bien con ella . Pero el tiempo pasaba rápidamente y teníamos que despedir nos . Creo que me divertí mucho , aunque estos días he estado todos los días pensando en ella . Antes no me encontraba con tanta buena chica como ella . Tengo su teléfono pero no quería llamar la . No me gusta el sentimiento triste . Espero tus respuestas . Un abrazo FRANCISCO”””, ”label”: ”A2”}, {”text”: ”””Querida Anita , ¡ Yo estoy muy contenta ! El ultimo año hice muchas cosas diferentes y buenas en mis vacaciones , quiero contar te . Yo fui con mi hermano a Egito y conoci las pyramides , la cultura , la comida , la música , la religion , las personas y el Nilo . Estuve en un barco por una semana en el Nilo , el sol me despiertó todos los días con su magía . Fui en el Vale_de_los_Renes , donde eston las múmias de muchas dinastias . Hacia más de 40 grados y sin árboles ... ¡ pero todo estaba perfecto ! En Cairo conoci su famoso museo , es muy grande y tiene muchas cosas para aprender . ¿ Y tus vacaciones como fueron ? Me conta . Hasta lluego , Maria_Alcina .”””, ”label”: ”A2”}, {”text”: ”””Hola Carlos ! Hace mucho que no te veo hombre ! Cómo está tu vida ? Trabajando muchísimo como siempre , no ? Continuas en el Instituto_Cervantes ahí de España ? Espero que sí , pues tu trabajo es buenísimo . Y por_favor , dí me lo ! Todavía vives con Maria ? ya estáis casados o sigues soltero ? Yo sigo soltera como siempre , tu sabes , hay cosas que no cambian . Pero tengo que contar te lo de Cláudia , está casada con Francisco , te acuerdas de Francisco ? El chico rubio de nuestra clase , el que no hablaba con nadie . Pues esta casada con el , y ella también ya no habla con nadie . Intento llamar la por teléfono pero sin respuestas . Es una lástima pues erámos muy amigas . Ya sabes , Carlos , que voy a viajar a Europa el próximo año . Voy a quedar me en Francia estudiando por seis meses . Después de esto tendré un més para quedar me donde quiera , por eso he pensado en visitar te . Sabes cuanto me gustaría conocer Madrid y , además , te echo de menos , pues hace mucho tiempo que no vuelves a el Brasil .”””, ”label”: ”B1”}, {”text”: ”””Querida Carla , Quiero preguntar por tu y de tu familia Como estas ? Como esta tus padres y tu pequeno hermano , Hugo ? Y tu trabajo ? Eras siempre una estudiante o has encontrado un trabajo ? Por_favor . Puedes preguntar a tu novio si puedes prestar me un coche de su sociedad ? Es simple para un o dos meses . Me haces falta , creas un vacio interior en mi y me siento sola sin tu cerca de mi porque eras mi amiga preferida . Si quieres que haga una cosa para ti , di me , una ayuda en tu trabajo de clase , o profesional , si quieres que me ocupe de tu gato o que pregunte por una persona . Claro , si tienes el deseo de pasar un momento en mi casa , pregunta me . Todo lo que quieres mi amiga ! Pero , tengo una cosa a preguntar te , ayer , he perdido mi coche porque una persona me la ha robada . Puedes contestar a mi carta enviando una otra carta . Vivo a el 42 Rue_Sanche_de_Pomiers , 33000 BORDEAUX .”””, ”label”: ”B1”}, {”text”: ”””En carnaval de ese año (2011) Bruno mi amigo de clase , fue para Olinda y bien estava lá conm su novia , sus amigos y estava muy feliz y bebendo mucho como de costume hahahahaha . Pero

su día no fue tan bueno como estoy hablando , o que ocorrió fue que otro hombre se quedó enamorado de él y estava cerca de Bruno todo lo tiempo , y él no percebió nada a princípio . Quería tirar lo hombre y matá lo tambíen , pero sus amigos y su novia no dejaran ! Bruno perdió su término de carnaval , se fue para su casa y ponió a durmir y olvidar de el ocorrido ! ! ! Pero un otro amigo (que se llama Oseas) , también tuvo una experiencia similar , fue quando yo y él fuimos a un ' show ' en Chevrolet_Hall , bien havia uns diez hombres (' gays ') cerca de un couche y nosotros estávamos cerca de lo mismo couche ... cuando vimos , los ' gays ' estavan a llamar mi amigo , pensavan que él también era de el mismo club kkkkkkkkkkkkkkkk . Fue muy " diferente " esa experiencia que mi amigo tuve ! Y fue también muy similar a história de Bruno en carnaval ."""", "label": "B1"}, {"text": """Y voy a graduar de mi universidad en dos años . ¿ Hay algunas posibilidades para obtener la residencia universaria ? ¡ Gracias por su respuesta de_antemano ! En_cuanto_a mí , soy estudiante de la facultad de traducción e interpretación de la Universidad_Estatal_Lingüística_de_Moscú , como yo he dicho . Me gusta estusdiar mucho y aprender muchas cosas nuevas y eso es la primera razón porque quiero seguir estudiando y , por_supuesto , ir a a el país de la lengua que estoy aprendiendo . Soy una persona muy trabajadora , y me gusta hacer el trabajo de principio a fin , sin posponer todo para más tarde . Siempre quiero hacer todo a el más alto nivel ya_que es fundamental en cada trabajo - Entonces , encuentro fácilmente el contacto con la gente , y me parece que eso es muy relevante en mi profesión . Entonce , voy a despedir me . Es un placer a escribir y tratar a entrar en vuestro Universidad . Y me espero mucho que va a contactar me ."""", "label": "B2"}, {"text": """Lev_Zykov . Un saludo , me llamo Lev_Zykov y me interesa su programa de posgrado de lengua y literatura española , porque me interesan los estilos en los que la literatura se ha escrito , también me gustaría saber más la literatura española y la lengua española , porque hay mucho que todavía no se en esta dirección , por_lo_tanto , me interesan las fechas cuando este programa se empieza y cuando se termina , también me gustaría saber condiciones de admisión , cuanto cuesta la participación en el programa y si el precio de alojamiento está incluido en el precio de el programa . Tengo 7 publicaciónes en ingenieria y un patento , pero me gustaría saber más en filología y literatura . Soy activo , me interesan muchas cosas de este mundo , especialmente cuando los explican bueno , también me gusta la cultura española y latinoamericana , yo bailo el tango argentino y me gusta mucho el arte . ¡ Muchas gracias por la atención !"""", "label": "B2"}, {"text": """El cigarro es uma droga que contiene muchas substancias que causan vicio y adicción a las personas . Por ser una droga lícita , hay un gran numero de personas que fuman y muchas vezes ni estan preocupada con su salud ni con la salud de las otras persona . En muchos lugares de el mundo existem reglas a cerca_de la utilizacíon de el cigarro . Una de estas reglas se traduz en poder fumar en lugares publicos o no . En mi opinión , no deveria ser permitido que las personas fumasen en lugares publicos , pues se aprobarmos la liberacíon de el uso de el cigarro en publico , estaremos ayudando , a estas personas a hacer daño a ellas mismas y aún a las personas que circulan en estes ambientes comunes . El humo producido por el cigarro puede causar cancer de pulmón , de lengua , de laringe , además acarretar daños a el corazón , desgatar los dientes , producir substancias inflamatorias en los vasos sanguineos , depresión , abuso de otras drogras , etc. Diante_de tantos problemas y maleficios que el cigarro puede causar , hago la seguiente pregunta : Por que las personas fuman ? Será que tienem consciencia de las consequencias de el uso de el cigarro ? Será que no imaginan que estan haciendo daño a si mismas y tambien a las otras personas con que conviven ? Concluo esta mi opinión , decindo que para mi , no debe ser permitido fumar en lugares publicos , porque ademas de hacer daño a los individuos en general , también estaremos haciendo daño a el medio_ambiente . Intenten cambiar este habito por otro saludable ! Se preocupen no solamente con sus vidas , pero tambien con la de las otras personas , con la naturaleza ! Fumar es un tipo de suicidio asistido , digan no a esta adiccíon !"""", "label": "B2"}, {"text": """Acudí a su compañía la dirección de la cual ví en una página web y personalmente llegué a su oficina para asugurar me de que tenían todos los documentos necesarios y que la compañía estaba acreditada . Allí una empleada suya me contó de las tarifas , me mostró unos cuantos comentarios de su clientela , que de_hecho eran más que buenos , y me aseguró que no iba a tener nigunos problemas con gas y electricidad una vez firmado el acuerdo con su compañia . Los precios me parecieron un_poco más altos de lo normal pero con la evaluación tan positiva de su empleada pensé que valía la pena probar lo . La primera semana todo funcionaba sin contratiempos : la luz se encendía y el gas tampoco fallaba , no

me podía quejar . Todo comenzó el 18_de_noviembre . Fue entonces cuando se me cortaron las luces por primera vez . Pensé que era un incidente aislado y que no volvería a pasar , pero lo mismo pasaba una y otra vez . En total he tenido 5 cortes de electrecidad en un mes ! No sé si es una practica común para su compañía pero yo me niego a pagar tanto dinero en vano ! Y encima ayer recibí la factura y por alguna razón las cifras que ví no coincidían con las que me habían prometido . Y no esperen que lo quede así ! Si su compañía no arregla el funcionamiento de la electrecidad y de el gas y si no me paga la indemnización dentro_de una semana , me veré obligado a acudir a las autoridades e incluso a el corte si es necesario . Espero que no tarden en responder . Con respeto , Ramón_Vidal""", "label": "C1"}, {"text": """Estimado Sres. , Me llamo Nicolas_Berhorst , economista , vivo en el barrio norte de la ciudad de Rosario hace 20 años . Escribo este correo pues estoy harto de aguardar por el teléfono oyendo aquella terrible música de ascensor en la cual tengo que escuchar a cada vez que me pasan por todos los empleados de su empresa , pero nadie puede resolver a mi problema . ¡ Lo que pasa es que así_que nuestra presidente privatizó esta empresa , y ustedes tomaron su administración , la cuenta llegó as las nubes ! ¡ Antes yo pagaba 50 pesos por mes en la cuenta de energía , y ahora la misma subió para 200 pesos ! ¡ Si uno hace las cuentas , es 300 % de aumento ! Santa_Fé entera debe estar mordiéndo se de rabia ahora . Y yo no me recuerdo que las aguas de el Rio_Paraná secaron de un mes para acá . Entonces me parece que los precios no pueden subir todo eso . ¿ Yo imagino como será en el invierno cuando tendremos que encender la calefacción , como vamos hacer para pagar ? ¡ Lo que me deja mas loco que una cabra es que cuando supimos de ese cambio , nos fue prometido que los precios iban a subir un_poco , pero no 300 % ! ¡ Yo fue engañado , y sé seguramente que no soy el único ! Es como comprar un producto que no te sirve porque es de baja calidad . ¿ Pregunto a usted , te parece justo cobrar un absurdo de un precio por La energía y córta la por 5 días en un único mes ? Es 17 % de el precio que no debía ser cobrado de nosotros . ¡ Es una calamidad !""", "label": "C1"}, {"text": """Mi nombre es Ana_María_Gómez_Martínez . Le escribo para quejar me de la calidad de servicios que me proporciona su compañía . Tengo todas las razones para estar muy discontenta . Y no conozco ninguna circunstancia que que sirva de excusa para su incapacidad de hacer su trabajo como se debe . Para empezar , la primera factura que he recibido sea mucho más alta que la provista . Hace un mes su empleado me prometió por teléfono que sería 50 euros y el pago que debo efectuar en realidad es 77 euros . ¿ No les parece sorprendiente ? 103 Podría consentir se lo , si el sericio fuera perfecta . Pero es ni siquiera satisfactorio . Me indigna el hecho de que de los 30 días de el mes haya tenido una vida tranquila solo durante los 25 debido_a los cortes de electricidad que han tenido en otros 5 . Primero , casi todos los productos en mi nevera no servieron para nada tras el corte más largo ."""", "label": "C1"}]

A.2. Prompt for Run 3

This appendix presents the approach used in Run 3, which involves performing a few-shot prompt engineering strategy using GPT-4o-mini with the five official examples provided on the competition website, one for each CEFR level.

few_shot_examples = [{"text": """¡Hola! ¿Qué tal?

Me llamo Carlos. Soy italiano, de Firenze. Vivo en Madrid. Soy estudiante. Tengo 19 años.

Soy alto, tengo el pelo negro, corto y liso.

En mi tiempo libre juego al fútbol, escucho música, voy al cine y salgo con mis amigos. Me gusta mucho viajar, escuchar música y comer.

Mi ciudad es famosa por la comida.

Me gusta estudiar español también, porque España es muy interesante y bonita.

No me gusta mucho estudiar por la noche.

Gracias. ¡Adiós!""", "label": "A1"},

{"text": """¡Hola!

Mis últimas vacaciones fueron en verano. Mi familia y yo viajamos a Tailandia por dos semanas.

Casi todos los días fuimos a la playa cerca del hotel y nadamos en el mar. También tuvimos varias excursiones,por ejemplo estuvimos en templo "Wat Chalong". Es muy grande y con una vista muy bonita. Está situado en la montaña.

Por otro día estuvimos en la excursión del mercado flotante. La verdad,estuvo un poco extraño pero interesante.

El lugar más memorable para mi familia y yo fue la isla ”Phi Phi”. Este lugar con una playa muy hermosa que tiene agua muy clara y arena blanca hizo relajarnos mucho.

Mi familia y yo tenemos las impresiones muy bonitas de Tailandia! Este país no se puede comparar con otros porque es muy diferente.

Creo que nuestras próximas vacaciones estarán en Italia porque tenemos muchas ganas de visitarlo.

Con mucho amor,Ana.”””, ”label”: ”A2”},

{”text”: ”””Estimados señores:

Me llamo Natalia. Soy clienta de su compania desde hace muchos años y nunca he tenido ningun problema antes.

Pero 3 semanas antes de 15 septiembre de 2018 su compania perdio mi equipaje cuando yo regresaba a Barcelona.

Yo hice una reclamacion en el aeropuerto,pero su compañia no quiso tomar ninguna responsabilidad. Hasta ahora no he recibido ni mi maleta,ni ninguna informacion sobre donde esta.

Espero que haya pasado algun error y puedan ayudarme solucionar este problema.

Mi maleta era de color negro,mediana,90 l,de marca American Tourister. Adentro habia muchas cosas importantes para mi,por ejemplo ropa,documentos y regalos para mi familia.

El completo listo de cosas pueden encontrar en mi reclamacion de 15 septiembre de 2018.

Por si a caso pueden poner en contacto conmigo por mi correo electronico o por telefono.

Espero que me respondan lo mas pronto posible.

Su clienta, Natalia”””, ”label”: ”B1”}, {”text”: ”””Estimados señores de la dirección de la universidad:

Me llamo Laura Martinez y tengo 23 años. He estudiado filología española en la universidad de Sevilla y terminé mis estudios hace dos meses.

Las lenguas siempre han sido algo que me interesan mucho desde mi infancia y no podría imaginar estudiar otra cosa diferente.

Como mis estudios de grado ya han terminado, me gustaría mucho continuar mi formación con un postgrado y el programa de lengua y literatura española de su universidad en Chile me parece muy interesante.

Durante mis estudios de grado hice prácticas en un instituto español y también escribí un pequeño trabajo sobre el uso del español entre los jovenes y los cambios que aparecen en la lengua hoy en día.

Lo que me gustaría hacer despues del postgrado es trabajar como profesora en un instituto porque enseñar es algo muy interesante para mi.

Sería muy bueno si me podría dar algunas informaciones sobre el programa de postgrado.

Primero, no estoy segura sobre la duración del programa. Despues, me gustaría saber si existen algunas condiciones de admisión, por ejemplo experiencia profesional o algunas notas especiales en la universidad. Tambien me interesaría recibir informacion sobre la organización del programa porque vivo actualmente en España y tengo que preparar muchas cosas antes de ir a Chile.

Otra cosa que me preocupa es el coste del programa. Me interesaría también saber si hay posibilidad de becas o residencia universitaria porque mi familia tiene que gastar mucho dinero para mis estudios y una ayuda financiera sería muy importante para mi.

Finalmente, mi última pregunta es sobre los servicios que ofrece la universidad para los estudiantes extranjeros.

Creo que este programa de postgrado sería perfecto para mi porque la lengua y la literatura española me interesan mucho.

He recibido dos cartas de recomendación de mis profesores en la universidad y mi nota final de grado es 8,7.

Soy una persona muy trabajadora y responsable que intenta siempre hacer las cosas lo mejor posible cuando quiere conseguir algo.

Quiero agradecerles mucho por su atención. Espero su respuesta pronto.

Atentamente, Laura”””, ”label”: ”B2”},

{”text”: ”””Estimados señores:

Me llamo Carlos Mendez y vivo en la calle Mayor 27, Alcázar de San Juan, Ciudad Real. Les escribo para comentar la situación que he experimentado con su compañia durante las últimas semanas.

El día 3 de abril firmé un contrato con su empresa para recibir los servicios de gas y electricidad en mi vivienda. Sin embargo, desde el principio han aparecido varios problemas que no esperaba encontrar.

En primer lugar, los datos que habiamos acordado por telefono no coinciden con los que aparecen en la factura que recibi hace unos días. Me sorprendió bastante porque durante nuestra conversación todo parecia claro y bien explicado.

Además, cuando intenté aclarar la situación llamando a su servicio de atención al cliente, la persona que me atendió no me dio muchas explicaciones y la impresión que tuve fue que queria terminar la conversación lo antes posible. Incluso me dijo que probablemente era un error del sistema pero no me explicó que debía hacer en este caso.

Otro problema importante es que durante este corto periodo he tenido varios cortes de electricidad. En total han sido aproximadamente cuatro días en los que no he tenido luz durante varias horas seguidas. Despues la electricidad vuelve a funcionar normalmente, pero al cabo de un tiempo vuelve a pasar lo mismo.

La situación resulta bastante complicada para mi familia porque tengo dos hijos pequeños y sin electricidad ni gas es dificil organizar la vida diaria de manera normal. Por ejemplo, algunas tardes hemos tenido que estar en casa sin luz y preparar la comida se vuelve bastante complicado.

Por todo esto, me gustaría pedirles que revisen mi contrato y las condiciones del servicio. Creo que sería necesario comprobar los precios que aparecen en la factura porque no corresponden con lo que habiamos acordado anteriormente.

Tambien agradecería que pudieran informarme sobre el motivo de los cortes de electricidad y si existe alguna solución para evitar que sigan ocurriendo en el futuro.

No quiero exagerar la situación, pero sinceramente esperaba un servicio más estable por parte de su empresa.

Espero que puedan revisar mi caso y darme una respuesta lo antes posible.

Atentamente, Carlos”””, ”label”: ”C1”}

]

A.3. Prompt for Run 4

This appendix presents the approach used in Run 4, which involves performing a few-shot prompt engineering strategy using GPT-4o-mini with explicit instructions related to linguistic features associated with proficiency development in Spanish as a foreign or additional language and with CEFR proficiency levels.

Prompt: {”role”: ”system”, ”content”: """Eres un experto clasificador de nivel de lengua usando CERF, para aprendices del español como lengua extranjera. Asigna únicamente con una de las siguientes categorías: ”A1”, ”A2”, ”B1”, ”B2” or ”C1”. Considera las letras mayúsculas para escribir la cateogría asignada. No expliques tu decisión.

Instrucciones:

Rasgo lingüístico | Indicador CEFR | Qué observar | Feature computacional

Tiempos verbales | A1–A2: presente / B1+: pasados / B2+: variedad | Uso de pasado, subjuntivo, condicional | Frecuencia por tiempo verbal (conteo y proporción) Variedad verbal | A1: limitado / B2+: amplio repertorio | Cantidad de formas verbales distintas | Número de formas verbales únicas Subjuntivo | B1–B2 en adelante | Presencia y uso correcto | % de verbos en subjuntivo Longitud de oración | A1: corta / B2+: más larga | Número de palabras por oración | Promedio de palabras por oración Subordinación | B1+: subordinadas / B2+: complejas | Uso de ”que”, ”aunque”, ”cuando” | Número de subordinadas por oración Profundidad sintáctica | B2–C1: estructuras anidadas | Oraciones con múltiples niveles | Profundidad del árbol sintáctico Conectores básicos | A1–A2: y, pero, porque | Uso limitado | Conteo de conectores básicos Conectores avanzados | B2+: sin embargo, por lo tanto | Variedad discursiva | Conteo y diversidad de conectores Diversidad léxica | A1: baja / B2+: alta | Repetición vs variedad | Type-Token Ratio (TTR) Léxico sofisticado | B2–C1 | Uso de palabras abstractas o poco frecuentes | Frecuencia de

palabras raras (según corpus) Errores gramaticales | A1–B1: frecuentes | Concordancia, género, verbos | Errores por cada 100 palabras Tipos de error | Varían por nivel | Tipo de error dominante | Clasificación de errores (género, tiempo, etc.) Cohesión | B1+: uso de referencia | Pronombres, sustituciones | Frecuencia de pronombres y anáforas Repetición léxica | A1–A2: alta | Repetición de palabras | Ratio de repetición léxica Estructura discursiva | B2+: organización clara | Introducción, desarrollo, cierre | Detección de estructura (heurística o modelo) Longitud del texto | Generalmente crece con nivel | Número total de palabras | Conteo total de tokens Densidad léxica | B2+: mayor densidad | Más contenido que función | Ratio palabras léxicas / totales Complejidad global | Aumenta con nivel | Mezcla de factores | Índices combinados (ej: complejidad sintáctica + léxica)

"""}]

B. Online Resources

The source code corresponding to all experiments is available in a GitHub repository at:

- GitHub