Measuring writing skills in Spanish as a foreign language with generative artificial intelligence

Medición de la habilidad escrita en español como lengua extranjera con inteligencia artificial generativa

María-Victoria Cantero Romero - *Universidad de Jaén, UJA (Spain)*
María-Teresa Martín-Valdivia - *Universidad de Jaén, UJA (Spain)*
Ana María Ortiz-Colón - *Universidad de Jaén, UJA (Spain)*
Salud María Jiménez-Zafra - *Universidad de Jaén, UJA (Spain)*

ABSTRACT

The emergence of Generative Artificial Intelligence (GAI)—particularly Large Language Models (LLMs) such as ChatGPT—is transforming the educational landscape, especially in the field of foreign language instruction. This article explores the potential of these technologies to automate the assessment of writing proficiency in Spanish as a Foreign Language (SFL), a task that is especially time-consuming at the beginning of university-level courses for Erasmus students. The study is based on three experiments conducted using the Spanish Learner Corpus compiled by the Instituto Cervantes. The first experiment applied a zero-shot learning approach by prompting the model with level descriptors from the Instituto Cervantes's Curriculum Plan. In the second and third experiments, the model was adjusted through fine-tuning using 90% and 80% of the corpus, respectively, with the remaining data reserved for testing and validation. The results indicate that the fine-tuned models significantly outperform the zero-shot configuration in identifying the correct proficiency levels of learner texts. These findings demonstrate that LLMs can be effectively employed to streamline the initial placement process in SFL courses, thus reducing the workload of instructors and improving efficiency. The study concludes that GAI can serve as a valuable complementary tool in multilingual and multicultural educational settings, provided its use is guided by sound pedagogical principles.

Keywords: language instruction; writing; artificial intelligence; teaching method innovations.

RESUMEN

La irrupción de la Inteligencia Artificial Generativa (IAG), y en particular de los Modelos de Lenguaje de gran tamaño (LLMs) como ChatGPT, está transformando el ámbito educativo, especialmente en la enseñanza de lenguas extranjeras. Este artículo analiza el potencial de estas tecnologías para automatizar la evaluación de la competencia escrita en español como lengua extranjera (ELE), una tarea especialmente laboriosa al inicio de los cursos universitarios dirigidos a estudiantes Erasmus. La metodología se basa en tres experimentos con el Corpus de Aprendices de Español del Instituto Cervantes. En el primero, se utilizó la técnica de *zero-shot learning*, proporcionando al modelo un *prompt* con los descriptores del Plan Curricular del Instituto Cervantes. En el segundo y tercer experimentos, se ajustó el modelo mediante *fine-tuning* con el 90 % y el 80 % del corpus, respectivamente, reservando el resto para validación y prueba. Los resultados muestran que los modelos ajustados son capaces de identificar el nivel de competencia escrita con una precisión significativamente superior al enfoque sin entrenamiento previo. Estos hallazgos evidencian que los LLMs pueden emplearse para agilizar procesos de evaluación inicial en cursos de ELE, reduciendo la carga docente y mejorando la eficiencia. Se concluye que la IAG representa una herramienta complementaria valiosa en contextos educativos multiculturales y multilingües, siempre que su uso esté guiado por criterios pedagógicos sólidos.

Palabras clave: enseñanza de lenguas; expresión escrita; inteligencia artificial; innovación pedagógica relevante.

How to cite: Cantero Romero, M.-V., Martín-Valdivia, M.-T., Ortiz-Colón, A. M., & Jiménez-Zafra, S. M. (2026). Measuring writing skills in Spanish as a foreign language with generative artificial intelligence [Medición de la habilidad escrita en español como lengua extranjera con inteligencia artificial generativa]. *RIED-Revista Iberoamericana de Educación a Distancia, 29*(1), 353-379. https://doi.org/10.5944/ried.45486

INTRODUCTION

The evolution of Artificial Intelligence has marked a turning point in all areas, especially in education (Aparicio Gómez, 2023; Hernández-León & Rodríguez-Conde, 2024; Zambrano Campozano, 2025). Advances in this discipline are opening up a field of new research and work with which to implement Artificial Intelligence, specifically generative AI, in the classroom. As Bolaño-García and Duarte-Acosta (2024) point out, generative Artificial Intelligence has gained attention in education because it improves the personalisation of learning and real-time feedback. Zapata Ros (2024) supports this idea of personalisation in addition to the availability of information. Likewise, Fajardo et al. (2023) highlight the use of these tools for personalising learning in university education, adapting it to the preferences and needs of each student through guided and virtual tutorials. García-Peñalvo et al. (2024) point out the importance of preparing both teachers and learners in the use of generative artificial intelligence, as it will be present in all aspects of life. Furthermore, as Moreno (2019) notes, it is important to highlight the potential of generative Artificial Intelligence to transform education by creating adaptive learning environments tailored to student performance, such as for students with special educational needs. In their study on generative Artificial Intelligence tools, Area-Moreira et al. (2024), in addition to the functions already mentioned, indicate that these tools can be used to automate tasks to support teachers in their work and even as anti-plagiarism systems. In line with the above, Chan and Tsi (2023) point to the use of generative Artificial Intelligence as a supplementary tool for teachers and not as a replacement for them. Moreno (2019) also points to three approaches to work with in education: generative Artificial Intelligence, educational robotics, and self-learning platforms. We will focus on the first of them in this study. Both Barroso-Osuna and Cabero-Almenara (2025) and Owan et al. (2023) identify a number of benefits of using Artificial Intelligence, specifically generative AI, in education, including the optimisation of teaching time and automated and accurate assessment. In this regard, Crespo Mendoza et al. (2024) point out that it can improve the accuracy and reliability of assessments.

As mentioned above, the main focus of this work is generative Artificial Intelligence, specifically Language Models (LLMs), which have provided a new tool for conducting studies aimed at improving teaching tasks. LLMs are natural language generation tools trained with a large amount of text (Wang, 2024). García-Peñalvo et al. (2024) identify various functions that LLMs can perform, such as supporting research, creating educational content, generating chatbots to interact with students by offering self-directed feedback, complementing search engines, paraphrasing text, teaching languages, and generating exams and questionnaires. It is in connection with these last two functions that the present study is framed: placement in language teaching, and specifically in Spanish as a foreign language.

Generative Artificial Intelligence offers the opportunity to carry out adaptive assessments for each student with immediate and specific feedback, suggesting possible solutions (Barroso-Osuna & Cabero-Almenara, 2025). Likewise, Language Models can be used to automate (García-Peñalvo, 2024) the correction of both multiple-choice tests and open-ended responses (Moreno, 2019). The advantages of using LLMs for assessment include efficiency (Area-Moreira et al., 2024) and plagiarism detection (García-Peñalvo et al., 2024). García-Peñalvo et al. (2024) also point out that LLMs improve teacher productivity.

With regard to the use of Language Models and language teaching, their usefulness is backed by various studies (Baskara & Mukarto, 2023; Salguero Romero, 2023; Wang, 2024). Baskara and Mukarto (2023) point out how ChatGPT is capable of generating realistic texts that bring students closer to the reality of the language. On the other hand, Hong (2023) points out the advantage of being able to use these tools to speed up exam correction, freeing teachers from workload and giving them the opportunity to focus on lesson preparation.

One of the challenges of language teaching, and in this case, teaching Spanish as a foreign language, is the appropriate level assessment of students when they begin Spanish courses. It is essential for a student to be at the right level of Spanish learning in order to progress appropriately, since if they are placed at a higher level than what is appropriate for them, they may become frustrated, and conversely, if they are placed at a lower level, they may lack motivation. That is why level assessment is key when starting language courses.

Until now, this level assessment has been carried out through multiple-choice tests or interviews (Biedma Torrecillas et al., 2012). These tests mainly focus on determining the student's level of oral and written expression and comprehension. Existing Spanish level tests, such as the Instituto Cervantes level test, consist of answering a series of written multiple-choice questions (true-false; matching or ordering) of increasing difficulty (Centro Virtual Cervantes, n.d.). This same typology is currently observed in international contexts, such as in the Spanish level tests at Columbia University and the University of Wisconsin–Madison, which are also based on multiple-choice exercises without including free written production (Columbia University, n.d.; University of Wisconsin–Madison, n.d.). However, Spanish language proficiency tests have not focused directly on written expression tests involving the writing of complete texts, but rather indirectly, through item responses, due to the limitations this entails, such as the lack of immediacy or the complexity of level assessment processes when dealing with large groups.

To solve this problem, this study will focus on the level assessment of the written expression test. As mentioned above, Language Models are capable of automating the grading of written work (García-Peñalvo et al., 2024), and they do so quickly, saving teachers time (Area-Moreira et al., 2024). This is the main reason why written expression tests have not been included in current level tests.

In this context, it is necessary to establish a conceptual basis for understanding the strengths and limitations of automated writing assessment and, in particular, the linguistic foundations that underpin the classification of levels A1–C1. The following section develops this theoretical framework, which will serve as a basis for the methodological proposal of this study.

This article is structured as follows. After the introduction and theoretical framework, the technological basis and choice of model, the objectives of the study and the methodology are presented. This is followed by the results, followed by a section on the pedagogical relevance of the model. Finally, the article concludes with a discussion and conclusions, to which a subsection on ethical aspects and licences for use has been added.

THEORETICAL FRAMEWORK

This section begins with a review of the state of the art in automated writing assessment and the linguistic foundations underpinning the level classification of the Common European Framework of Reference (CEFR) and the Instituto Cervantes Curriculum Plan (PCIC). This conceptual basis will serve to contextualise the study proposal and, finally, to present the objectives guiding the research.

Automated handwriting assessment: current developments and approaches

The automatic classification of learners' texts using language models is part of a broader tradition of automated writing assessment (AWE), the development of which has given rise to multiple tools and systems that should be taken into account in order to contextualise this work. AWE has evolved considerably in recent decades, becoming an increasingly common tool in educational contexts. Pioneering systems such as e-rater (Burstein et al., 2003), developed by ETS, have been widely used in standardised tests, using linguistic, grammatical and discursive metrics to estimate textual quality. Coh-Metrix (McNamara et al., 2014), meanwhile, allows for detailed analysis of cohesion, syntactic complexity and readability, providing a multifactorial approach to written discourse. Recent studies, such as that by Zhang (2021), offer systematic reviews of these systems, highlighting their transition from rule-based approaches to models driven by machine learning and natural language processing. Along the same lines, Wang et al. (2022) analyse current approaches to the evaluation of argumentative texts, focusing on discursive components such as reasoning structure, evidence, and organisation. Other proposals, such as Writing Mentor (Burstein et al., 2018), integrate automatic assessment with formative feedback, promoting self-regulation processes in academic writing. Likewise, tools such as Write & Improve, developed by the University of Cambridge, exemplify how it is possible to provide immediate automatic feedback on texts produced by foreign language learners, facilitating autonomous and guided learning (Cambridge English, n.d.).

Linguistic foundations of the A1–C1 level classification

The classification of texts produced by learners of Spanish as a foreign language at levels A1–C1 is based on the descriptors established by the Common European Framework of Reference for Languages (Council of Europe, 2002) and the Instituto Cervantes Curriculum Plan (Instituto Cervantes, 2006). These documents define in detail the linguistic, pragmatic and sociolinguistic competences associated with each level, providing a solid basis for assessment.

At level A1, the lexical repertoire is very limited and restricted to basic transactions and everyday expressions. Texts are very short and simple, with simple sentences and a low average number of words per utterance. The use of regular forms of the present indicative predominates and an elementary grammatical repertoire is employed (Council of Europe, 2002; Instituto Cervantes, 2006).

At level A2, learners can produce short texts that convey simple information on familiar topics. A larger lexical repertoire and slightly more complex structures are observed, incorporating past indicative tenses (preterite, imperfect and indefinite) and

some irregular forms of the present tense. The use of the affirmative imperative also appears (Council of Europe, 2002; Instituto Cervantes, 2006).

At level B1, the lexical repertoire is broader and allows for the creation of texts that fulfil a specific communicative task, maintaining a coherent structure. Grammatically, tenses such as the simple future, the simple conditional and the past perfect are handled with a certain fluency, in addition to introducing the present subjunctive and the negative imperative. Discourse shows greater cohesion and a wider variety of connecting devices (Council of Europe, 2002; Instituto Cervantes, 2006).

At level B2, the user has a broad and precise linguistic repertoire, capable of sustaining complex arguments and detailed descriptions. Compound and subordinate clauses are used fluently, as well as confident use of indicative tenses (present, past, future and conditional) and subjunctive tenses (present, imperfect, perfect and pluperfect). Textual cohesion is consistent and lexical nuances appropriate to different registers are used (Council of Europe, 2002; Instituto Cervantes, 2006).

At level C1, the linguistic and non-linguistic repertoire is sufficiently broad and flexible to handle any type of communicative transaction or interaction, even in demanding academic or professional contexts. The learner is able to produce long, complex texts with a clear, well-organised structure, using all tenses accurately and a wide range of syntactic and lexical resources (Council of Europe, 2002; Instituto Cervantes, 2006).

TECHNOLOGICAL BASIS AND MODEL SELECTION

This study used version 3.5 of the ChatGPT language model developed by OpenAI, released in November 2022 (OpenAI, 2022). Although this version is not open access, it allows fine-tuning through OpenAI's API. The choice of version 3.5 over the latest version is justified by the possibility of performing this process, which is not possible in the latest version of ChatGPT.

For this research, GPT-3.5 was chosen over more recent or open-source models due to a combination of technical, economic, and methodological factors. As this is a novel task—the automatic level assessment of written proficiency in Spanish as a foreign language—it was considered appropriate to evaluate the performance of a widely tested generalist model, such as GPT-3.5, both in its base configuration and through fine-tuning. This model, accessible through the OpenAI API, does not require advanced computational infrastructure and offers a robust architecture with good coverage of Spanish (Li, 2023; Pourpanah et al., 2023). In addition, it offers an adequate balance between performance, cost and response speed, key factors for validating the viability of the approach in this exploratory phase (Roumeliotis et al., 2024). In contrast, at the time the experiments were carried out, models such as GPT-4 involved a considerably higher economic cost and longer inference times, which reinforced the decision to use GPT-3.5 as the initial reference.

However, it should be noted that GPT-3.5 has limitations compared to more recent models, such as GPT-4, which offer more accurate contextual understanding, training with larger and more heterogeneous datasets, and greater reasoning ability. These characteristics make them particularly suitable candidates for future research, both in the automated assessment of written competence aligned with CEFR levels and in the generation of high-quality synthetic corpora for training specialised systems.

STUDY OBJECTIVE

To carry out this study, the descriptors established by the Common European Framework of Reference for Languages (Council of Europe, 2002) and the Instituto Cervantes Curriculum Plan have been adapted in order to establish clear and operational criteria that allow a language model, in this case ChatGPT, to automatically classify the written productions of Spanish learners according to their level of competence.

Despite the aforementioned advances in automatic writing assessment, few studies have focused on the level assessment of written texts produced by foreign language learners and, specifically, on the specific case of teaching Spanish as a foreign language. Existing research tends to address automatic feedback or grading, but not the classification of written work according to CEFR or Instituto Cervantes Curriculum levels, especially in initial assessment tasks. This gap is particularly relevant in contexts such as universities, where the heterogeneity of international students, as in Erasmus programmes, requires efficient tools for level assessment. This study proposes an innovative solution based on generative language models (ChatGPT), which allows for the automatic classification of written texts by learners of Spanish as a foreign language (ELE) according to their level of proficiency, reducing the teaching load and improving the management of workload at the start of the academic year. Thus, this work not only complements previous research focused on textual improvement, but also broadens the scope of automated assessment to include initial diagnostic tasks in second language teaching contexts.

That is why, as the general objective of the study, we have set out to evaluate the effectiveness of the ChatGPT language model as an innovative tool for assessing the written expression level of Spanish learners. We also hope to achieve a series of specific objectives: i) to find out whether ChatGPT, with its prior training, is capable of carrying out level assessment adequately; ii) to verify whether ChatGPT is capable of level assessment if it is adjusted with a corpus levelled with the Instituto Cervantes Curriculum Plan and the Common European Framework of Reference for Languages; and iii) to determine the impact of ChatGPT on the efficiency of the level assessment of Spanish as a foreign language.

METHODOLOGY

To carry out this research, we used the Corpus of Spanish Learners (hereinafter CAES) (Palacios Martínez et al., 2019), developed by the Instituto Cervantes in collaboration with the University of Santiago de Compostela. We also used version 3.5 of the ChatGPT language model, developed by OpenAI (OpenAI, 2022).

Three types of experimental tests were carried out in this study: one using the zero-shot learning technique and the other two using model fine-tuning.

The fine-tuning procedure was carried out with a single training epoch, keeping the rest of the parameters at their default values according to the OpenAI API. To ensure reproducibility, a fixed random seed (value 42) was used, i.e., a reference value that allows experiments to be reproduced with the same results every time. No class balancing techniques were applied, as it was decided to preserve the actual distribution of the corpus, thus reflecting authentic levelling conditions in the classroom. In this way, the system's results are better suited to the challenges of real educational scenarios, without introducing artificial modifications to the representation of levels.

However, it is recognised that in future research, compensation strategies could be applied to compare their effect on the fairness and robustness of the model. Cross-validation was also not used, in line with the initial and experimental nature of the study.

Regarding the technical parameters of the training, the default values of OpenAI's API were maintained in aspects such as batch size (i.e., the number of examples processed at a time), learning rate (which indicates the speed at which the model adjusts its parameters during training), and loss functions (metrics that measure the difference between the model's prediction and the expected result). No additional regularisation techniques or early stopping strategies (early termination of training to avoid overfitting) were applied, as the main objective was to validate the feasibility of the approach rather than to optimise the model for maximum performance. This decision is in line with the exploratory nature of the research, which aims to test the applicability of the model to the task of automatic text levelling.

CAES Corpus

The CAES corpus was compiled by the University of Santiago de Compostela and funded by the Instituto Cervantes. Computerised data was collected from October 2011 to December 2020 from centres, mostly universities, in different countries such as Spain, the United States, Brazil, Egypt, Ireland and Portugal. The students who took part in the project had eleven different native languages (English, Mandarin Chinese, Portuguese, Arabic, Russian, German, French, Greek, Italian, Japanese and Polish).

This study used version 2.1 of the CAES corpus from March 2022, which updated the first data collection, which had a smaller amount of data, with a total of 1,423 participants, compared to the 2,544 participants in the 2022 update.

The corpus contains examples of different levels of Spanish according to the Common European Framework of Reference for Languages, from A1 to C1. At levels A1, A2 and B1, texts belonging to three different types of tasks that students had to write, ranging from 30 to 200 words in length, were collected. Levels B2 and C1 have a sample of two tasks per level, ranging from 275 to 500 words in length.

The topics identified at each level are as follows: At A1, the first task consists of an email about changing jobs, with a total of 728 samples; the second task consists of an email about their family, with 703 samples; the last task consists of a note about being late, with a sample of 705. At level A2, the first task is a biography, with 673 samples; the second task is booking a hotel room, with a sample of 603 texts; and the last task is writing a postcard about your holidays, with a sample of 701 texts. As it can be seen, around 2,000 examples were collected per level at the initial levels, providing a significant sample of these levels. Likewise, the texts' themes are matched to the functions and corresponding textual products of each level according to the Instituto Cervantes Curriculum Plan (Instituto Cervantes, 2006).

At level B1, there are also three different tasks. The first is to write a letter to a friend, with a sample of 528 texts; the second task is to write an email about a complaint to an airline, with a sample of 454; and the last task is to narrate a story, with a sample of 382. It should be noted that these tasks, like those at levels A1 and A2, are matched to the functions described in the Curriculum Plan.

At levels B2 and C1, the tasks are reduced to two. At level B2, the first task is to write an application for admission, with a total of 375 samples; while the second task

at level B2 is to write a text arguing the case for smoking in public places, with a sample size of 356.

At level C1, the first task consists of writing a complaint to a gas company, with a sample of 169; and in the second task, they must write a film review, with a sample of 184 texts. It can be seen how the sample size is significantly reduced at these levels as the number of tasks is reduced. It is also notable how the samples are smaller at level C1, with fewer than 400 samples at this level.

As in the previous levels, it can be seen that the tasks in levels B2 and C1 also correspond to the functions described in the Curriculum Plan.

The corpus was annotated by specialists in teaching Spanish as a foreign language at the University of Santiago de Compostela. Each text was classified into one of the CEFR levels according to the criteria established in the Instituto Cervantes's Curriculum and the guidelines defined in the project itself (Palacios Martínez et al., 2019; University of Santiago de Compostela, n.d.). To ensure the reliability of the classification, the texts were evaluated independently by several annotators and then reviewed jointly until a consensus was reached. This procedure ensures that level assessment is coherent and consistent, making the corpus a solid resource for research and automatic assessment of written production.

Table 1 presents a summary of the levels with respect to the tasks addressed and the total number of samples collected.

Cantero (2024) also conducted a study of this corpus, from which the following results can be drawn. At level A1, the average number of words per sentence is between 10.9 and 11.7, and the most frequent words are simple connectors and a limited lexicon. At level A2, the average number of words per sentence is higher, ranging from 12.5 to 14.6, with a simple, frequent lexicon. At the next level, B1, the average number of words per sentence increases to 12.0-16.5, and the vocabulary is broader and more complex than at previous levels. At level B2, the average number of words per sentence is even more complex, ranging from 17.7 to 21.5, and in terms of vocabulary, more complex connectors and more specialised vocabulary are used. Finally, at level C1, the average number of words per sentence is between 20.7 and 23.3, with complex and varied vocabulary.

Table 1
Sample summary

Level	Task	Exhibit
A1	Email regarding change of job	728
	Family email	703
	Note: arriving late	705
A2	Biography of a person you admire	673
	Book a hotel room	603
	Holiday postcard	701
B1	Letter to a friend	528
	Funny story	382
	Airline claim	454
B2	Application for admission	375
	Smoking in public places	356
C1	Film review	184
	Gas company complaint	169

Prompt used

Once the corpus had been analysed, a specific prompt was developed to carry out the tests with the model. As Morales-Chan (2023) points out, a good prompt can guarantee the success of the task. Therefore, it is important to define the objective and provide sufficient context.

The following prompt was used in the three tests carried out in this study —zero-shot learning (Test 1) and fine-tuning (Tests 2 and 3)— in order to maintain methodological consistency in the evaluation criteria. The prompt[1] design was based on the linguistic descriptors of the Cervantes Institute Curriculum Plan and the Common European Framework of Reference for Languages.

Tú eres un experto lingüista especializado en enseñanza de español como lengua extranjera. Tu tarea es indicar el nivel de español como lengua extranjera de los textos siguiendo el Plan Curricular del Instituto Cervantes.

Aquí tienes una descripción de los distintos niveles.

Niveles A1 y A2 Transacciones básicas relacionadas con su entorno.

A1: Repertorio limitado de léxico, textos muy breves y sencillos, un promedio de 10 palabras por oración. Formas regulares del presente de indicativo.

A2: Textos breves con información sencilla, un promedio de 12 palabras por oración. Tiempos verbales del pasado de indicativo: Pretérito perfecto, imperfecto e indefinido.Formas irregulares de presente de indicativo. Imperativo afirmativo.

Niveles B1 y B2 desenvolverse con textos sobre temas de su interés gustos y preferencias.

B1: Vocabulario amplio pero sencillo, realizar textos con una tarea concreta. Presente de indicativo, pretérito perfecto, imperfecto e indefinido de indicativo, futuro simple, condicional simple, pretérito pluscuamperfecto de indicativo, presente de subjuntivo. Imperativo negativo.

B2: Repertorio lingüístico amplio, oraciones subordinadas. Tiempos verbales de indicativo: presente, pretérito perfecto, imperfecto, indefinido, futuro simple y compuesto, condicional simple y compuesto, pretérito pluscuamperfecto. Tiempos verbales de subjuntivo: presente, pretérito imperfecto, pretérito perfecto y pluscuamperfecto.

C1 transacciones de todo tipo. Disponen de un repertorio de recursos lingüísticos y no lingüísticos lo suficientemente amplio y rico. Pueden enfrentarse a una amplia serie de textos extensos y complejos. Todos los tiempos verbales de indicativo y de subjuntivo el presente, pretérito perfecto, imperfecto y pluscuamperfecto.

Ahora vas a recibir un TEXTO y teniendo en cuenta lo explicado anteriormente y los errores gramaticales indica al final de tu respuesta con la etiqueta 'NIVEL:' el nivel del TEXTO (A1, A2, B1, B2 o C1).

TEXTO: "..."

With this prompt, brief descriptive information has been added for each level, following the analysis of the corpus mentioned above. Likewise, following the Instituto Cervantes Curriculum Plan (Instituto Cervantes, 2006), a description of the verb

[1] English version of the prompt in Appendix 2

tenses used in each of the levels and the types of texts has been included, following the indications of the textual products. In this way, the model is provided with a broader context so that its response is more accurate and tailored to the needs requested.

Testing with the ChatGPT Language Model

A total of three different tests were carried out in the study to assess the model's ability to assess the level of the texts in the corpus.

The first test was zero-shot learning. In this process, the model does not receive specific examples, but rather relies on its prior knowledge. To carry out this test, only the prompt mentioned in the previous section was used.

In tests two and three, fine-tuning was performed. This process consists of specialising a pre-trained model to perform a specific task, adapting it to a specific set of data provided to it. We obtained this dataset from the CAES corpus, given that it provides with a set of clear examples for the model. To perform fine-tuning, the model is equipped with an input-output specifying the input and the type of output we want it to provide. In this case, the prompt is the one mentioned in the previous section. Furthermore, the output requested was the level of Spanish. Likewise, to perform fine-tuning, part of the dataset was reserved to verify the response.

Tests 2 and 3 differ from each other in the division of the corpus. For the second test, the corpus was divided into 90%, 5% and 5%. Ninety per cent of the corpus was used to train the model, 5% to validate it and the remaining 5% to test it. In the third test, an 80%–20% division was made, using 80% of the corpus for training and 20% for testing.

RESULTS

To analyse the results, three evaluation measures widely used in classification tasks have been employed:

- **Precision:** indicates the percentage of examples that the model classified at a certain level and that actually belong to that level.
- **Recall**: indicates the proportion of examples of a specific level that the model correctly identified. For example, the percentage of A1-level texts detected as A1 out of the total number of A1 texts in the corpus.
- **F1-score**: A single value that combines accuracy and coverage through its harmonic mean, providing a balanced measure of performance. This metric is particularly useful when it is important for the model not only to be correct, but also to detect all possible cases in each category.

The results of the three experiments are presented below:

Zero-shot learning experimentation

As mentioned above, in this technique, the model is not provided with any examples; it is carried out using the prompt developed. In this case, the results obtained are as follows:

Table 2
ZSL experimental results

Tags	Precision	Recall	F1-score
A1	0.9375	0.1402	0.2439
A2	0.3333	0.7677	0.4648
B1	0.2400	0.2609	0.2500
B2	0.4286	0.8110	0.1364
C1	0.0000	0.0000	0.0000

Table 2 shows that, at level A1, accuracy is high, i.e. texts classified as level A1 have a high probability of being at this level. However, coverage is quite low, as the model has detected only a small percentage of texts in the corpus as A1. Therefore, the F1 score is low.

At level A2, the opposite phenomenon to that described at level A1 occurs. Accuracy is low, so it is less successful, but coverage is high. Therefore, we can say that at this level, although it detects A2 texts to a greater extent, it has low accuracy when it comes to detecting the correct level.

In relation to level B1, both factors, precision and coverage, are low. At this level, the model has problems both detecting B1-level texts and identifying them at their correct level.

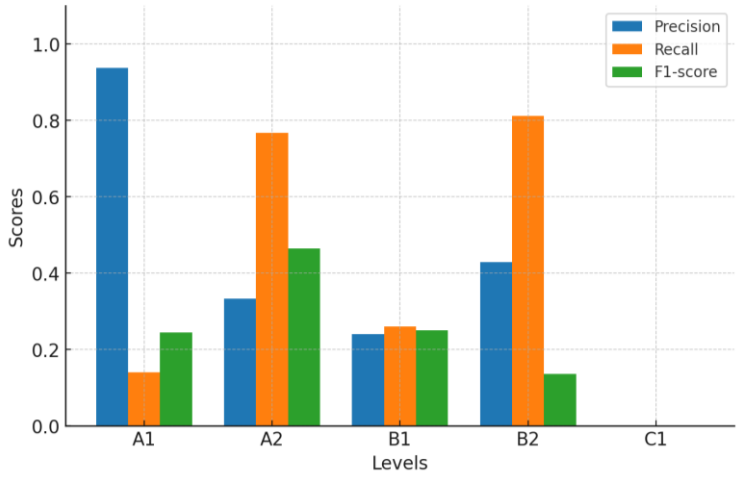
With regard to level B2, coverage is high, detecting most B2-level texts, but precision is moderate, with most classifications being incorrect.

Finally, the case of level C1 is striking, as both accuracy and coverage are 0; it does not detect any texts at this level.

Furthermore, as this was an experiment using a prompt, the model not only gave the level of the text in its responses, but also added comments on each one about the errors found. Complete examples of these responses, including the original texts and the corrections proposed by the model, are presented in Appendix 1.

Figure 1 complements this information by showing a comparison of the accuracy, coverage and F1-score values for each level evaluated in the zero-shot learning configuration.

Figure 1
Results by level in zero-shot learning experimentation



Fine-tuning experimentation 90-5-5

In this second experiment, as mentioned above, the corpus was fine-tuned and divided into 90%, 5% and 5% for training, validation and testing of the model. The results of this experiment are shown below:

Table 3
Fine-tuning experimentation 90-5-5

Tags	Precision	Recall	F1-score
A1	0.9905	0.9720	0.9811
A2	0.9519	1.0000	0.9754
B1	1.0000	0.9565	0.9778
B2	1.0000	1.0000	1.0000
C1	1.0000	1.0000	1.0000

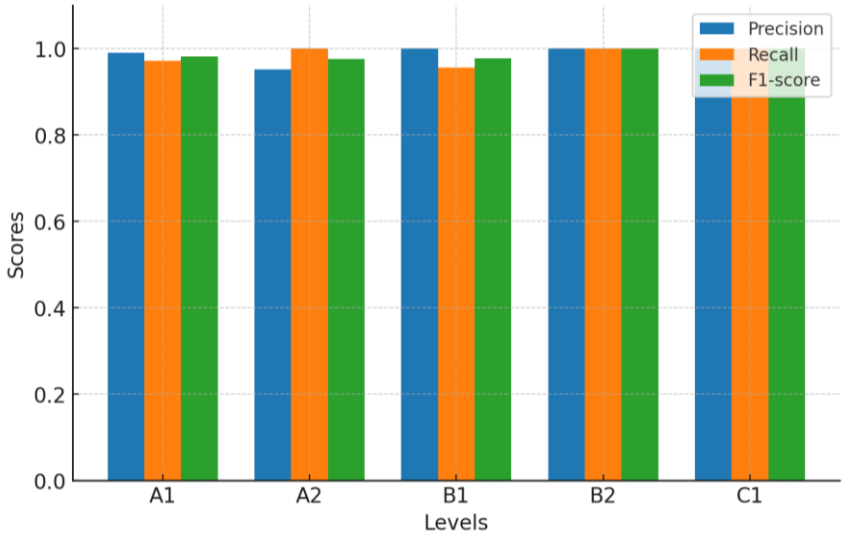
Table 3 shows that the results have higher levels of accuracy and coverage than those obtained with zero-shot learning. At level A1, we can see that the model makes almost no errors and detects almost all texts at level A1.

At level A2, as we can see, the model detects all texts at level A2 with very high accuracy.

With regard to level B1, the model correctly predicts all texts and also has high coverage, with a small percentage undetected.

Finally, with levels B2 and C1, the results show that the model detects all texts at these levels and is correct on all occasions. However, the results of this experiment show that the model can correctly predict all levels with values close to or equal to 1. Figure 2 complements this information by showing a comparison of the accuracy, coverage and F1-score values for each level evaluated in the 90-5-5 fine-tuning configuration.

Figure 2
Results by level in the 90-5-5 fine-tuning experiment



Fine-tuning experimentation 80-20

In the third experiment, fine-tuning was performed by dividing the corpus into two percentages, 80% for training and 20% for model validation. The results obtained are shown below.

In this experiment, as shown in Table 4, at level A1, both precision and coverage show a high level of accuracy and text detection.

Levels A2 and B1 show a similar result to A1, although with a slightly lower result in accuracy at level A2 and coverage at level B1.

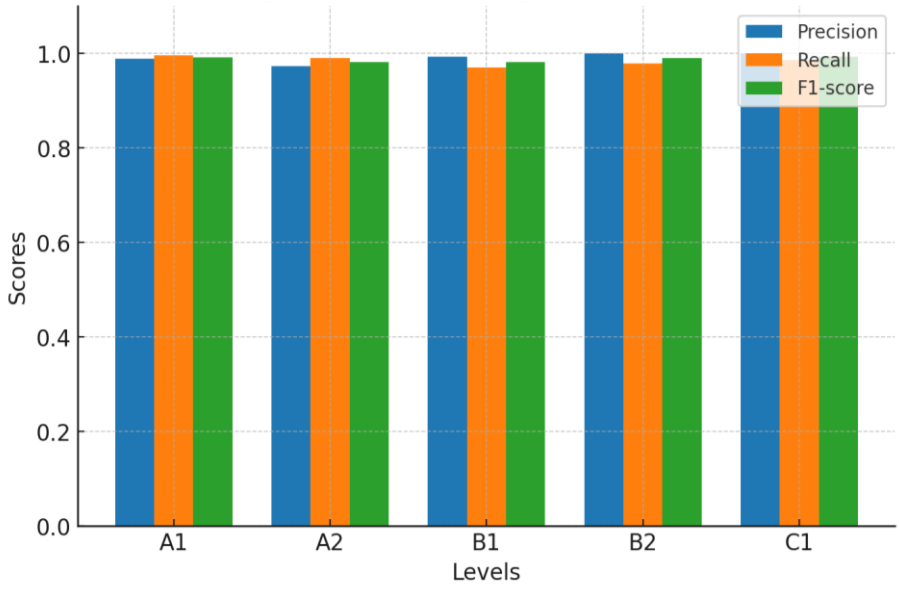
With regard to levels B2 and C1, we observe that the accuracy of the model is excellent, as it correctly predicts these levels, although the coverage is moderately lower.

Table 4
Fine-tuning experimentation 80-20

Tags	Precision	Recall	F1-score
A1	0.9884	0.9953	0.9918
A2	0.9727	0.9899	0.9812
B1	0.9925	0.9707	0.9815
B2	1.0000	0.9795	0.9896
C1	1.0000	0.9859	0.9929

Figure 3 complements this information by showing a comparison of the precision, coverage, and F1-score values for each level evaluated in the 80-20 fine-tuning configuration.

Figure 3
Results by level in the 80-20 fine-tuning experiment



Statistical analysis

To assess the statistical significance of the results, 95% confidence intervals (95% CI) were estimated for the Macro-F1 values using **bootstrapping**, a resampling technique with replacement that allows confidence intervals to be calculated without assuming a specific statistical distribution. One thousand replicates were performed to obtain these intervals, which indicate the range within which the true value of Macro-F1 is expected to fall with a 95% probability. Additionally, hypothesis tests were performed to verify the results. The zero-shot learning experiment obtained a Macro-F1 value of 0.2190 with a confidence interval CI95% = [0.1785, 0.2617], while the fine-tuning experiments achieved values close to perfection: a Macro-F1 value of 0.9869 with a confidence interval CI95% = [0.9748, 0.9959] on the 90-5-5 split and a Macro-F1 value of 0.9874 (95% CI = [0.9812, 0.9933]) on the 80-20 split. The comparisons were made using non-parametric bootstrap-based hypothesis tests—i.e., statistical contrasts that do not require assuming a specific data distribution and are based on multiple random resamples. These analyses showed that both fine-tuning experiments statistically significantly outperformed the zero-shot experiment ($\Delta^1 \approx 0.77$; $p < 0.001$ in both cases). In contrast, no significant differences were observed between the two fine-tuning experiments ($\Delta = -0.0005$; 95% CI = [−0.0133, 0.0113]; $p = 0.964$). These results confirm that fine-tuning substantially improves the model's ability to assess written competence in ELE.

PEDAGOGICAL RELEVANCE OF THE MODEL AND ITS APPLICATION IN BOTH FACE-TO-FACE AND DISTANCE LEARNING CONTEXTS

The results obtained show that a language model adjusted with a specialised corpus can achieve very high performance in the automatic classification of written texts according to CEFR levels. This capability has clear potential to facilitate teaching, especially in the initial stage of student placement, constituting a pedagogical innovation in the use of artificial intelligence for language teaching.

In a real learning environment, the system could be integrated into a Learning Management System (LMS) as a diagnostic assessment module. The flow of use would be simple: the student submits a text, the model automatically classifies it by level and presents the result to the teacher using a rubric aligned with the descriptors of the Instituto Cervantes Curriculum Plan (PCIC). In this way, teachers could assign students to the course corresponding to their actual level, optimising the adjustment of groups and avoiding gaps that could affect learning progress.

In future implementations, and based on more exhaustive analyses, the system could generate more detailed reports based on the PCIC, identifying specific linguistic areas to be reinforced (e.g. use of verb tenses, textual cohesion or lexical repertoire). This information would allow the course to be tailored not only to the student's level, but also to their main weaknesses, guiding the teaching programme towards improving these aspects.

In distance education and online learning contexts, integration into an LMS would have the same diagnostic function, automatically classifying students from their first access. This would allow for efficient group organisation even in non-face-to-face environments, which is especially relevant in massive courses or virtual programmes with continuous enrolment.

Compared to tools such as Write & Improve (Cambridge English, n.d.), which focus mainly on corrective feedback and formative assessment, the proposal presented here offers a complementary approach: automatic classification by CEFR levels. This feature, combined with the possibility of direct integration into educational platforms, reinforces its value as a teaching support tool for optimising initial placement and streamlining course planning.

In all cases, the system is designed as a teaching aid and not as a replacement, with the aim of streamlining initial assessment tasks and freeing up time for higher value-added activities, such as individualised feedback or progress monitoring.

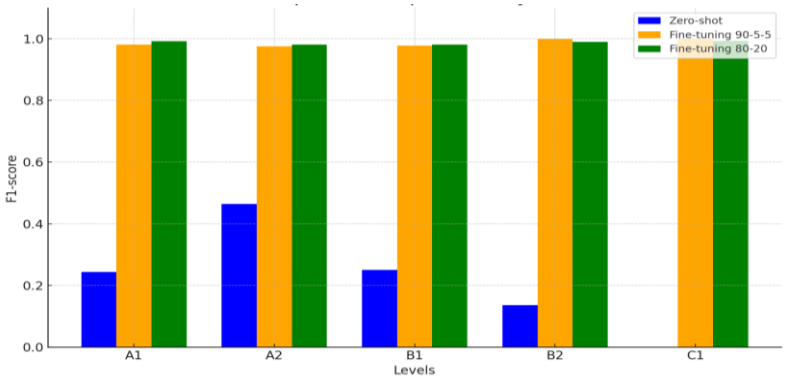
Among the possible future improvements are expanding the corpus to include more genres and topics, adding additional lexical and pragmatic descriptors, experimenting with other language models, and adapting the system to other languages and educational levels. It would also be relevant to evaluate its real impact on student performance and motivation, as well as its smooth integration into different LMSs and educational contexts.

DISCUSSION AND CONCLUSIONS

Once the tests have been carried out and the results obtained, there is a clear contrast between the model's performance in the zero-shot learning configuration and in the fine-tuning experiments. In the former, the model does not achieve adequate accuracy at any of the levels, with only a relatively high value in A1, albeit with very low coverage. At levels A2 and B2, coverage is medium-high (0.76 and 0.81), indicating that the model detects most of the texts at those levels, but with low accuracy. Level C1 is particularly problematic, with zero values for both accuracy and coverage. These data confirm that, with its original pre-training, the model is not capable of reliably assessing the level of ELE texts according to the Instituto Cervantes Curriculum Plan.

In contrast, the two fine-tuning tests (90-5-5 and 80-20) provided accuracy and coverage values very close to 1 at all levels, with particular strength at B2 and C1. The comparison between the two experiments suggests that a larger volume of training data can improve performance, although even with less data the model maintains a high predictive capacity. This behaviour coincides with that observed in previous work on model adaptation for specific natural language processing tasks ((García-Peñalvo, 2024; García-Peñalvo et al., 2024), where customisation of the system significantly increases its effectiveness.

Figure 4
Comparison of experimental results



Compared to tools such as Write & Improve (Cambridge English, n.d.), which focus on formative feedback and error detection, this study offers a complementary approach: automatic classification by CEFR levels, which could be integrated into learning environments to optimise the initial placement of students and streamline teaching. Likewise, studies such as those by Burstein et al. (2003, e-rater) and McNamara et al. (2014, Coh-Metrix) had already shown the usefulness of combining linguistic features and automatic metrics to evaluate writing; our results reinforce this line of thinking, showing that a model adjusted with specific corpora achieves very high performance levels.

On the other hand, in the zero-shot learning experiment, the model not only indicated the texts' level but also generated comments on errors and suggestions for correcting the texts. Although this functionality is not relevant to the main objective of this study, it could be explored in future research as a teaching support resource to help students identify and correct their errors.

In terms of limitations, the study was conducted on a single corpus due to the scarcity of specialised and level-assessed ELE corpora that meet homogeneous criteria and are managed by experts. Furthermore, the classification was based on grammatical descriptors present in the prompt (e.g., verbal repertoire, average number of words per sentence, use of certain tenses and modes), without systematically integrating complex syntactic structures, specific lexical resources, or pragmatic aspects such as appropriateness or discursive coherence, which are particularly relevant at intermediate and advanced levels.

As lines of future work, we propose:

1. Design and validate new corpora for Spanish as a foreign language, level-assessed by experts, with greater thematic diversity and representativeness of levels.
2. Explore the application of the method with other language models to compare their performance.
3. Expand the set of linguistic descriptors, incorporating syntactic, lexical, and pragmatic indicators.
4. Experiment with controlled variations of the prompt to evaluate their impact on classification.
5. Validate the system with authentic student texts in real educational contexts and integrate the tool into learning management systems (LMS).

Overall, the results achieved confirm that generative artificial intelligence, and in particular language models adjusted with specific data, can become an effective resource for streamlining the initial level assessment of ELE students, supporting teaching work, and optimizing teaching-learning processes.

Within this reflection, it is also pertinent to address the ethical and legal aspects involved in the use of learner corpora, which are discussed in the following subsection.

Ethical and licensing issues

This study used the Corpus of Spanish Learners (CAES), which is available online for academic and research purposes (University of Santiago de Compostela, n.d.). Its texts have been anonymised beforehand so that they do not include any identifiable personal data, and they were compiled in regulated educational contexts

and with expert validation, which guarantees the appropriate treatment of the information.

From an ethical point of view, it is necessary to consider that the expansion or combination of the corpus with other resources must address both clarity in usage licences and the prevention of biases associated with mother tongue or sociocultural context. These factors are essential to ensure fair and reproducible classification of texts and, ultimately, to guarantee the responsible use of artificial intelligence in educational environments.

Acknowledgements

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00) and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project ROMANET (CERV-2024-CHAR-LITI-101215052), funded by the European Union under the Citizens, Equality, Rights and Values programme, and Project HEART-NLP-UJA (PID2024-156263OB-C21) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU. The research work conducted by Salud María Jiménez-Zafra is part of the grant RYC2023-044481-I, supported by MICIU/AEI/10.13039/501100011033 and by ESF+.

NOTES

1. To quantify the differences between models, the symbol Δ (delta) was used, which is conventionally employed in statistics to indicate the difference between two values. Thus, Δ = Macro-F1(fine-tuning) – Macro-F1(zero-shot) ≈ 0.77 means that the difference in Macro-F1 between the experiments being compared is approximately 0.77.

REFERENCES

Aparicio Gómez, W. O. (2023). La inteligencia artificial y su incidencia en la educación: transformando el aprendizaje para el siglo XXI. *Revista Internacional de Pedagogía e Innovación Educativa*, *3*(2), 217-229. https://dialnet.unirioja.es/servlet/articulo?codigo=9624350

Area-Moreira, M., Del Prete, A., Sanabria-Mesa, A. L., & Sannicolás-Santos, M. B. (2024). No todas las herramientas de IA son iguales: análisis de aplicaciones inteligentes para la enseñanza universitaria. *Digital Education Review*, *45*, 141-149. https://doi.org/10.1344/der.2024.45.141-149

Barroso-Osuna, J., & Cabero-Almenara, J. (2025). Potencialidades de la inteligencia artificial en la personalización de la educación. In P. Román-Graván, J. Barroso-Osuna, J. Cabero-Almenara, & C. Llorente-Cejudo (Eds.), *Visiones sobre la integración educativa de la inteligencia artificial* (1st ed.). Dykinson. https://doi.org/10.14679/4177

Baskara, R., & Mukarto, M. (2023). Exploring the implications of ChatGPT for language learning in higher education. *Indonesian Journal of English Language Teaching and Applied Linguistics*, *7*(2), 343-358. https://doi.org/10.21093/ijeltal.v7i2.1387

Biedma Torrecillas, A., Chamorro Guerrero, M. D., Lozano, G., & Sánchez Cuadrado, A. (2012). Diseño y validación de las pruebas de nivel del CLM de la Universidad de Granada. In *Actas del VII Congreso ACLES: Multilingüismo en los centros de lengua universitarios: evaluación, acreditación, calidad y política lingüística* (pp. 26-37). ACLES. https://dialnet.unirioja.es/servlet/libro?codigo=501925

Bolaño-García, M., & Duarte-Acosta, N. (2024). Una revisión sistemática del uso de la inteligencia artificial en la educación. *Revista Colombiana de Cirugía, 39*(1), 51-63. https://doi.org/10.30944/20117582.2365

Burstein, J., Elliot, N., Beigman Klebanov, B., Madnani, N., Napolitano, D., Schwartz, M., Houghton, P., & Molloy, H. (2018). Writing mentor: Writing progress using self-regulated writing support. *Journal of Writing Analytics*, *2*, 285-313. https://doi.org/10.37514/JWA-J.2018.2.1.12

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (2003). E-rater as a diagnostic tool for writing instruction. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations* (pp. 79-81). Association for Computational Linguistics.

Cambridge English. (n. d.). *Write & Improve*. https://writeandimprove.com

Cantero, M. V. (2024). 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* (pp. 55-64). Octaedro.

Centro Virtual Cervantes. (n. d.). *Ítem de respuesta cerrada*. https://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/diccionario/itemrespuestacerrada.htm

Chan, C. K. Y., & Tsi, L. H. Y. (2023). *The AI revolution in education: Will AI replace or assist teachers in higher education* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2305.01185

Columbia University, Department of Latin American and Iberian Cultures. (n. d.). *Spanish placement exam*. Recuperado el 25 de julio de 2025, de https://laic.columbia.edu/content/spanish-second-language-placement-exam

Council of Europe. (2002). *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación*. Instituto Cervantes; Ministerio de Educación, Cultura y Deporte. https://cvc.cervantes.es/ensenanza/biblioteca_ele/marco/cvc_mer.pdf

Crespo Mendoza, R., Rodríguez López, W., Montenegro Patrel, M., & Tomalá Tomalá, G. (2024). IA: una herramienta para asistir a los docentes en la evaluación de los estudiantes. *Conocimiento Global, 9*(2), 305-323. https://doi.org/10.70165/cglobal.v9i2.423

Fajardo, G. M., Ayala, D. C., Arroba, E. M., & López, M. (2023). Inteligencia artificial y la educación universitaria: una revisión sistemática. *Magazine de las Ciencias: Revista de Investigación e Innovación, 8*(1), 109-131. https://doi.org/10.33262/rmc.v8i1.2935

García-Peñalvo, F. J. (2024). Cómo afecta la inteligencia artificial generativa a los procesos de evaluación. *Cuadernos de Pedagogía, (549)*.

García-Peñalvo, F. J., Llorens-Largo, F., & Vidal, J. (2024). La nueva realidad de la educación ante los avances de la inteligencia artificial generativa. *RIED-Revista Iberoamericana de Educación a Distancia, 27*(1), 9-39. https://doi.org/10.5944/ried.27.1.37716

Hernández-León, N., & Rodríguez-Conde, M. J. (2024). Inteligencia artificial aplicada a la educación y la evaluación educativa en la universidad: Introducción de sistemas de tutorización inteligentes, sistemas de reconocimiento y otras tendencias futuras. *Revista de Educación a Distancia (RED), 24*(78), Artículo 6. https://doi.org/10.6018/red.594651

Hong, W. C. H. (2023). The impact of ChatGPT on foreign language teaching and learning: Opportunities in education

and research. *Journal of Educational Technology and Innovation, 5*(1), 38-53. https://doi.org/10.61414/jeti.v5i1.103

Instituto Cervantes. (2006). *Plan curricular del Instituto Cervantes: Niveles de referencia para el español* (3 vols.). Biblioteca Nueva. https://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/

Li, Y. (2023). *A practical survey on zero-shot prompt design for in-context learning* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2309.13205

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

Morales-Chan, M. A. (2023). *Explorando el potencial de ChatGPT: Una clasificación de prompts efectivos para la enseñanza*. Universidad Galileo. https://biblioteca.galileo.edu/tesario/handle/123456789/1348

Moreno, R. D. (2019). La llegada de la inteligencia artificial a la educación. *Revista de Investigación en Tecnologías de la Información, 7*(14), 260-270. https://doi.org/10.36825/riti.07.14.022

OpenAI. (2022). *ChatGPT (versión 3.5) [Artificial intelligence language model]*. https://openai.com

Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O., & Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *EURASIA Journal of Mathematics, Science and Technology Education, 19*(8), em2307. https://doi.org/10.29333/ejmste/13428

Palacios Martínez, I., Barcala Rodríguez, F. M., & Rojo, G. (2019). El corpus de aprendices de español (CAES) y sus aplicaciones para la enseñanza y aprendizaje del español como lengua extranjera. In M. Blanco, H. Olbertz, & V. Vázquez Rozas (Eds.), *Corpus y construcciones: Perspectivas hispánicas* (pp. 273-301). Universidade de Santiago de Compostela (Verba, Anexo 79). https://doi.org/10.15304/9788417595876

Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., & Lim, C. P. (2023). A review of generalized zero-shot learning methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(4), 4051-4070. https://doi.org/10.1109/TPAMI.2022.3182926

Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2024). Next-generation spam filtering: Comparative fine-tuning of LLMs, NLPs, and CNN models for email spam classification. *Electronics, 13*(11), 2034. https://doi.org/10.3390/electronics13112034

Salguero Romero, P. (2023). *La traducció pedagògica i l'ús de ChatGPT-3 a classes d'anglès com a segona llengua per a nens i nenes* [Trabajo de fin de grado, Universitat Autònoma de Barcelona]. Repositorio UAB. https://ddd.uab.cat/record/279383

Universidad de Santiago de Compostela. (n. d.). *Corpus de aprendices de español (CAES)*. https://galvan.usc.es/caes

University of Wisconsin–Madison, Testing and Evaluation Services. (n. d.). *Spanish placement test*. University of Wisconsin–Madison. https://testing.wisc.edu/centerpages/spanishtest.html

Wang, L. (2024). *El uso de ChatGPT y Gemini para la preparación de actividades de comprensión oral de CLE: Un estudio comparativo* [Trabajo de fin de máster, Universitat Autònoma de Barcelona]. Repositorio UAB. https://ddd.uab.cat/record/300353

Wang, X., Lee, Y., & Park, J. (2022). *Automated evaluation for student argumentative writing: A survey* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2205.04083

Zambrano Campozano, R. F. (2025). Impacto de la inteligencia artificial en la transformación digital de la educación. *Serie Científica de la Universidad de las Ciencias Informáticas, 18*(1), 177-192. https://scielo.sld.cu/scielo.php?script=s

ci_arttext&pid=S2306-24952025000100177

Zhang, S. (2021). Review of automated writing evaluation systems. *Journal of China Computer-Assisted Language Learning, 1*(1), 170-176. https://doi.org/10.1515/jccall-2021-2007

Zapata Ros, M. (2024). IA generativa y ChatGPT en educación: Un reto para la evaluación y ¿una nueva pedagogía? *Revista Paraguaya de Educación a Distancia (REPED), 5*(1), 12-44. https://doi.org/10.56152/reped2024-vol5num1-art2

APPENDICES

Appendix 1. Complete examples of model responses.

Corpus text:
"Hola Eva.
Que hace dos meses, hemos estudiado espanol en madrid.
He vuelt a Japon la semana pasada.
Glacias por tu ayuda.
Yo escribo mi familia, mi mujer, hija y hijo.
Mi mujer trabaya en Idioma.
Hija trabaya en una tienda de palo, hijo en un enterprise electrolico.
Vivimos en yokohama cerca de tokyo.
Vamos a ir a madrid que mes viene
Por favor, vernos en madrid.
Hasta plonto." (CAES, 2022)
Model response:
Errores encontrados en el texto:
- "Que hace dos meses, hemos estudiado espanol en madrid." La forma correcta sería "Hace dos meses, estudiamos español en Madrid."
- "He vuelt a Japon la semana pasada." La forma correcta sería "He vuelto a Japón la semana pasada."
- "Glacias por tu ayuda." Debería ser "Gracias por tu ayuda."
- "Yo escribo mi familia, mi mujer, hija y hijo." Se debería decir "Te escribo sobre mi familia: mi mujer, mi hija y mi hijo."
- "Mi mujer trabaya en Idioma." La forma correcta sería "Mi mujer trabaja en Idiomas."
- "Hija trabaya en una tienda de palo, hijo en un enterprise electrolico." Debería ser "Mi hija trabaja en una tienda de ropa, y mi hijo en una empresa electrónica."
- "Vamos a ir a madrid que mes viene." La forma correcta sería "Vamos a ir a Madrid el mes que viene."
- "Hasta plonto." Debería ser "Hasta pronto."

NIVEL: A2 (OpenAI, 2022)
—
Corpus text:
"El 24 de Noviembre
Estimados señores
El 22 de noviembre pasado, viajé con Tonta Airlines de Badaros a Madrid con el vuelo TA 222.
Estoy aquí en Madrid hasta el 27 para algunas citas profesionales.
Cuando embarqué en Badaros, la azafata me dijo que mi maleta estaba demasiado grande y que tenia que facturarla.
La maleta esta de marca Samsonite, de modelo XYZ1234, color negra, con mis nombre y dirección muy visibles.
Cuando llegamos a Madrid, la maleta no fue encontrada.
Hace a dos dias que estoy esperando pero nadie de su compañía me ha llamado con información.
Evidentemente, no puedo quedar sin cambiarme o lavarme me.

Entonces compré algunas cosas y me parece que Tonta Airlines tenga que indemnizarme.

Yo incluido la lista de mis compras y les pide que se pongan en contacto para decirme si sean de acuerdo de pagar el 50% des estas compras.

Se pueden contactarme por correo electronico o llamarme en el movil 660 12 34 45 y evidentemente, en cualquier caso que Tonta Airlines encuentre mi maleta, me daría mucha alegria.

Un saludo" (CAES, 2022)

Model response:

El texto presenta varios errores gramaticales y algunos elementos propios de un nivel de español como lengua extranjera B1:

- Uso incorrecto de tiempos verbales: "tenia" en lugar de "tenía".
- Errores en el uso de los artículos: "la maleta esta" en lugar de "la maleta está".
- Uso incorrecto de pronombres: "Hace a dos días" en lugar de "Hace dos días".
- Errores de concordancia de género y número: "me parece que Tonta Airlines tenga que indemnizarme" debería ser "me parece que Tonta Airlines debe indemnizarme".

NIVEL: B1 (OpenAI, 2022)

Appendix 2. English version prompt.

You are a linguist expert specializing in teaching Spanish as a foreign language. Your task is to indicate the level of Spanish as a foreign language in the texts, following the Instituto Cervantes Curriculum Plan.
Here is a description of the different levels.
Levels A1 and A2 Basic transactions related to their environment.
A1: Limited vocabulary, very short and simple texts, an average of 10 words per sentence. Regular forms of the present indicative.
A2: Short texts with simple information, an average of 12 words per sentence. Past indicative tenses: present perfect, imperfect, and indefinite. Irregular forms of the present indicative. Affirmative imperative.
Levels B1 and B2: ability to handle texts on topics of interest, tastes, and preferences.
B1: Extensive but simple vocabulary, writing texts with a specific task. Present indicative, past perfect, imperfect, and indefinite indicative, simple future, simple conditional, past perfect indicative, present subjunctive. Negative imperative.
B2: Extensive linguistic repertoire, subordinate clauses. Indicative verb tenses: present, past perfect, imperfect, indefinite, simple and compound future, simple and compound conditional, past perfect. Subjunctive verb tenses: present, imperfect, past perfect and past perfect.
C1 transactions of all kinds. They have a sufficiently broad and rich repertoire of linguistic and non-linguistic resources. They can deal with a wide range of long and complex texts. All indicative and subjunctive verb tenses: present, past perfect, imperfect, and pluperfect.
Now you are going to receive a TEXT. Taking into account the above and the grammatical errors, indicate at the end of your answer with the label 'LEVEL:' the level of the TEXT (A1, A2, B1, B2, or C1).
TEXT: "…"

Date of reception: 1 June 2025
Date of acceptance: 9 September 2025
Date of approval for layout: 8 October 2025
Date of publication in OnlineFirst: 28 October 2025
Date of publication: 1 January 2026