REVISTA DE EDUCAÇÃO A DISTÂNCIA E ELEARNING

ISSN 2182-4967

Modelos de lenguaje en ELE: comparativa de ChatGPT, LLaMA 3 y Gemini para la nivelación y la retroalimentación de la expresión escrita

Modelos de linguagem em ELE: comparação entre ChatGPT, LLaMA 3 e Gemini para a nivelagem e o feedback da expressão escrita

Language models in ELE: comparison of ChatGPT, LLaMA 3 and Gemini for levelling and feedback on written expression

María Victoria Cantero Romero[1]

Resumen

La incorporación de modelos de lenguaje de gran tamaño ha abierto nuevas posibilidades para la enseñanza del español como lengua extranjera, especialmente en relación con la personalización del aprendizaje. En este contexto, la nivelación de la expresión escrita y la retroalimentación formativa se configuran como ámbitos clave para apoyar tanto la evaluación como la revisión de textos producidos por aprendientes. El objetivo de este estudio es comparar el rendimiento de ChatGPT, LLaMA 3 y Gemini en tareas de nivelación y retroalimentación de la expresión escrita español. La metodología se basa en el análisis de textos de aprendientes procedentes del Corpus de Aprendices de Español, a los que se aplicó un mismo *prompt* estructurado alineado con los descriptores del Marco Común Europeo de Referencia para las lenguas y del Plan Curricular del Instituto Cervantes.

Los resultados muestran perfiles diferenciados entre los modelos en términos de precisión, adecuación curricular y calidad de la retroalimentación, lo que permite valorar su potencial y sus limitaciones para una integración pedagógica fundamentada en la enseñanza de español como lengua extranjera.

Palabras clave: Inteligencia Artificial; Modelos de Lenguaje; Español Lengua Extranjera; nivelación escrita; retroalimentación

Resumo

A incorporação de modelos de linguagem de grande porte abriu novas possibilidades para o ensino do espanhol como língua estrangeira, especialmente no que diz respeito à personalização da aprendizagem. Neste contexto, a nivelagem da expressão escrita e o feedback formativo constituem-se como áreas-chave para apoiar tanto a avaliação como a revisão de textos produzidos pelos alunos. O objetivo deste estudo é comparar o desempenho do ChatGPT, LLaMA 3 e Gemini em tarefas de nivelamento e feedback da expressão escrita do espanhol. A metodologia baseia-se na análise de textos de alunos provenientes do Corpus de Aprendices de Español, aos quais foi aplicado um mesmo prompt estruturado alinhado com os descritores do Quadro Europeu Comum de Referência para as Línguas e do Plano Curricular do Instituto Cervantes.

[1] Universidad de Jaén, Espanha, vcantero@ujaen.es, https://orcid.org/0009-0008-7052-7322

Os resultados mostram perfis diferenciados entre os modelos em termos de precisão, adequação curricular e qualidade do feedback, o que permite avaliar o seu potencial e as suas limitações para uma integração pedagógica baseada no ensino do espanhol como língua estrangeira.

Palavras-chave: Inteligência Artificial; Modelos de Linguagem; Espanhol como Língua Estrangeira; nivelamento escrito; feedback

Abstract

The incorporation of large language models has opened up new possibilities for teaching Spanish as a foreign language, especially in relation to personalised learning. In this context, the levelling of written expression and formative feedback are key areas for supporting both the assessment and revision of texts produced by learners. The aim of this study is to compare the performance of ChatGPT, LLaMA 3 and Gemini in levelling and feedback tasks for written expression in Spanish language. The methodology is based on the analysis of texts written by learners from the Corpus de Aprendices de Español (Corpus of Spanish Learners), to which the same structured prompt was applied, aligned with the descriptors of the Common European Framework of Reference for Languages and the Instituto Cervantes Curriculum Plan. The results show different profiles among the models in terms of accuracy, curricular adequacy, and quality of feedback, which allows for an assessment of their potential and limitations for pedagogical integration based on the teaching of Spanish as a foreign language.

Keywords: Artificial Intelligence; Language Models; Spanish as a Foreign Language; written levelling; feedback

1. Introducción

En los últimos años, los modelos de lenguaje de gran tamaño (Large Language Models, LLM) han despertado un creciente interés en el ámbito educativo. En el contexto de la enseñanza de lenguas, estos modelos han sido explorados principalmente como herramientas de apoyo a la corrección y a la evaluación de producciones escritas. Esto ha abierto nuevas posibilidades para automatizar parcialmente tareas tradicionalmente dependientes del profesorado.

En la enseñanza del español como lengua extranjera (ELE), la evaluación de la expresión escrita constituye un ámbito especialmente complejo. La producción escrita implica la integración de múltiples dimensiones lingüísticas y discursivas, asimismo, su valoración requiere un análisis cuidadoso de la adecuación del texto a los descriptores curriculares de nivel. En este marco, la nivelación de la expresión escrita, entendida como la asignación de un nivel de competencia, a partir de textos producidos por aprendientes, desempeña un papel clave en la organización de cursos y en la planificación didáctica. Todo esto supone una tarea costosa en términos de tiempo y recursos docentes.

La nivelación en ELE se apoya habitualmente en marcos de referencia como el Marco Común Europeo de Referencia para las Lenguas (en adelante MCER), complementado, en determinados contextos, por documentos de referencia como el Plan Curricular del Instituto Cervantes (en adelante PCIC), que concreta y desarrolla los contenidos asociados a los distintos niveles. La aplicación de estos descriptores a producciones

reales exige una interpretación experta que no siempre resulta fácil, especialmente en contextos donde podemos encontrar un elevado número de estudiantes. A ello se suma la necesidad de proporcionar retroalimentación formativa que oriente la revisión de los textos y favorezca el desarrollo progresivo de la competencia escrita, una labor que incrementa aún más la carga docente.

En este contexto, los LLM se presentan como una posible herramienta de apoyo para la nivelación y la retroalimentación en ELE. Su capacidad para generar respuestas coherentes y detalladas plantean oportunidades para agilizar los procesos de evaluación inicial y complementar la intervención docente. No obstante, esto podría llevar una serie de errores como la sobreestimación del nivel real del aprendiente, la falta de coherencia con los descriptores curriculares y la generación de retroalimentación poco precisa. Por ello, resulta necesario evaluar de forma empírica y controlada el rendimiento de estos modelos en tareas específicamente situadas en el contexto de ELE.

A pesar del creciente número de estudios sobre el uso de LLM en evaluación automática, sigue siendo limitada la investigación comparativa centrada en la nivelación de la expresión escrita en ELE. En particular, son escasos los trabajos que enfrenten distintos modelos de lenguaje a un mismo corpus de producciones auténticas de aprendientes. Esta falta de evidencia dificulta la toma de decisiones fundamentadas sobre la integración de estas tecnologías en prácticas docentes reales.

Con el fin de contribuir a este ámbito, el presente estudio compara el rendimiento de ChatGPT, LLaMA 3 y Gemini en tareas de nivelación y retroalimentación de la expresión escrita en ELE. El objetivo general es analizar la adecuación de la asignación de nivel y la calidad de la retroalimentación generada por cada modelo. De manera específica, se persigue: (1) evaluar la adecuación de la clasificación de nivel en relación con los descriptores del MCER y del PCIC; (2) analizar la claridad y pertinencia de la retroalimentación ofrecida; y (3) discutir fortalezas, limitaciones y posibles aplicaciones didácticas de los LLM como herramientas de apoyo al profesorado.

2. Marco teórico

2.1. Modelos de lenguaje y evaluación automática en lenguas extranjeras

La investigación reciente sobre el uso de modelos de lenguaje de gran tamaño (LLM) en evaluación lingüística se ha centrado, principalmente, en el análisis de su rendimiento en tareas de corrección y calificación automática de producciones escritas. En el caso del español, estudios empíricos han examinado la eficacia de estos modelos en la evaluación de respuestas abiertas, poniendo de relieve tanto su capacidad para generar juicios consistentes como ciertas limitaciones en la sensibilidad a los distintos niveles de competencia lingüística (Capdehourat et al., 2025).

Asimismo, las revisiones sistemáticas sobre el uso educativo de los LLM, han permitido identificar un conjunto de aplicaciones recurrentes en contextos de enseñanza de

lenguas, entre las que destacan la evaluación automática, la generación de retroalimentación y el apoyo a la revisión de textos escritos (Emirtekin, 2025). Estas revisiones coinciden en señalar que, si bien los LLM muestran un alto potencial para automatizar parcialmente tareas evaluativas, su utilización en contextos educativos requiere un análisis cuidadoso de los criterios empleados y de su adecuación a los objetivos pedagógicos.

En este sentido, diversos trabajos subrayan que la fluidez discursiva de los LLM no constituye un indicador suficiente de calidad evaluativa. Especialmente, cuando se trata de tareas situadas en contextos específicos de enseñanza de lenguas extranjeras. Estudios recientes destacan la necesidad de evaluar el rendimiento de estos modelos mediante diseños empíricos controlados y tareas educativas concretas, atendiendo no solo al resultado generado, sino también tanto a su adecuación pedagógica y a la interacción con los aprendientes (Emirtekin, 2025; Han et al., 2024).

Este planteamiento resulta especialmente relevante para el ámbito del ELE, donde la evaluación de la expresión escrita implica la alineación con descriptores curriculares explícitos y con objetivos formativos bien definidos. En consecuencia, el análisis del uso de LLM en tareas como la nivelación y la retroalimentación exige un enfoque que combine evidencia empírica y criterios pedagógicos.

2.2. Evaluación y nivelación de la expresión escrita en ELE

En la enseñanza del español como lengua extranjera, la evaluación de la expresión escrita ocupa un lugar central debido a su carácter integrador y a su relevancia para la certificación y la progresión del aprendizaje. A diferencia de otras destrezas, la producción escrita exige valorar simultáneamente aspectos formales y funcionales del lenguaje, así como su adecuación a la tarea comunicativa propuesta, lo que la convierte en una de las habilidades más complejas de evaluar en contextos de enseñanza de lenguas (Sánchez-Jiménez, 2009). Dentro de este marco, la nivelación se concibe como el proceso de asignar un nivel de competencia a partir de las producciones del aprendiente, con el fin de orientar la enseñanza y facilitar la organización de los cursos, de acuerdo con descriptores curriculares como los propuestos en el MCER.

En este estudio, la nivelación de la expresión escrita se fundamenta en los descriptores del MCER, en la medida en que proporcionan una escala común de progresión (A1–C2) y orientaciones generales para la evaluación de la producción escrita (Consejo de Europa, 2001, 2020). No obstante, dado su carácter amplio, la aplicación de estos descriptores a producciones reales de aprendientes exige una interpretación contextualizada.

En el ámbito específico del ELE, esta interpretación se concreta mediante el PCIC, que desarrolla los niveles del MCER a través de descriptores más detallados y operativos para el español (Instituto Cervantes, 2006). La combinación de ambos marcos permite situar las producciones escritas en un continuo de nivel y reducir la arbitrariedad en la asignación, sin eliminar la necesidad de un juicio experto.

2.3. Retroalimentación formativa y automatización

La retroalimentación formativa desempeña un papel fundamental en el desarrollo de la competencia escrita, en la medida en que orienta al aprendiente en la revisión de sus textos y favorece la toma de conciencia sobre su propio proceso de aprendizaje. En contextos de enseñanza de lenguas, la eficacia de la retroalimentación depende no solo de la identificación de errores, sino también de la claridad, pertinencia y orientación pedagógica de los comentarios ofrecidos.

En el ámbito de la evaluación automatizada, la generación de retroalimentación constituye un reto adicional, ya que no basta con producir observaciones lingüísticamente correctas, sino que estas deben alinearse con objetivos formativos y criterios curriculares explícitos. Las revisiones recientes sobre el uso educativo de los modelos de lenguaje señalan que, si bien los LLM muestran un alto potencial para generar comentarios detallados en lenguaje natural, la calidad pedagógica de dicha retroalimentación varía en función del diseño de la tarea y de los criterios de evaluación empleados (Emirtekin, 2025; You et al., 2026).

Desde una perspectiva empírica, estudios centrados en la interacción entre aprendientes y modelos de lenguaje en tareas de escritura en lenguas extranjeras subrayan la necesidad de evaluar la retroalimentación generada no solo en términos de corrección formal, sino también atendiendo a su utilidad para el proceso de aprendizaje y a su adecuación al contexto educativo específico (Han et al., 2024). En este sentido, los LLM se presentan como una opción prometedora para complementar la retroalimentación docente, siempre que sus respuestas se ajusten a criterios curriculares y eviten formulaciones ambiguas o excesivamente generales.

2.4. Estudios previos sobre LLM y evaluación en ELE

En el ámbito específico del ELE, algunos trabajos recientes han comenzado a explorar el uso de LLM para la evaluación y nivelación de la expresión escrita. Los estudios de Cantero Romero (2024) y Cantero Romero y Ortiz Colón (2025) analizan el potencial de modelos como ChatGPT y LLaMA 3 para asignar niveles a producciones escritas de aprendientes, destacando tanto la fluidez de las respuestas generadas como ciertas tendencias a la sobreestimación del nivel real. Estos trabajos ponen de relieve la importancia del diseño del *prompt* y de la inclusión explícita de descriptores curriculares para mejorar la adecuación de la nivelación.

2.5. Limitaciones de la investigación actual y enfoque adoptado

En conjunto, la literatura revisada pone de manifiesto el potencial de los modelos de lenguaje de gran tamaño para apoyar tareas de evaluación y retroalimentación, pero también evidencia una falta de estudios empíricos comparativos en el ámbito del ELE que analicen de forma sistemática el comportamiento de distintos modelos bajo un mismo diseño metodológico (Capdehourat et al., 2025; Emirtekin, 2025; Leiva-Guerrero et al., 2026). En particular, resulta necesario evaluar su rendimiento en tareas de nivelación de la expresión escrita alineadas con el MCER y el PCIC, incorporando

criterios pedagógicos que permitan valorar no solo la precisión de la asignación de nivel, sino también la utilidad didáctica de la retroalimentación generada. Este vacío justifica la pertinencia del presente estudio y orienta el enfoque comparativo adoptado.

3. Metodología

3.1. Diseño del estudio

El estudio adopta un diseño empírico comparativo orientado a analizar el rendimiento de tres modelos de lenguaje de gran tamaño (ChatGPT, LLaMA 3 y Gemini) en tareas de nivelación de la expresión escrita en ELE. El diseño enfrenta a los tres modelos a un mismo conjunto de textos y a un mismo *prompt*, con el fin de garantizar la comparabilidad directa de los resultados.

3.2. Corpus

El corpus analizado está compuesto por 60 textos escritos por aprendientes de español como lengua extranjera, extraídos del Corpus de Aprendices de Español (CAES) (Palacios et al., 2019). Se seleccionaron producciones correspondientes a los niveles A1, A2 y B1, de acuerdo con la clasificación original del corpus, con una distribución equilibrada de 20 textos por nivel.

Los textos corresponden a producciones escritas breves y completas, sin anotaciones de corrección, elaboradas por aprendientes adultos en contextos formativos de ELE. Los textos del corpus fueron recogidos en distintos centros del Instituto Cervantes y centros asociados de manera controlada por los docentes de dichos centros (Palacios et al., 2019). La selección de los niveles A1–B1 responde a su relevancia en procesos de nivelación inicial, donde la asignación de nivel resulta especialmente determinante para la organización de cursos y la planificación didáctica.

3.3. Modelos analizados

El estudio compara el rendimiento de tres modelos de lenguaje ampliamente utilizados, seleccionados por su disponibilidad y por representar enfoques distintos en el desarrollo de LLM:

Se empleó ChatGPT como representante de los modelos de lenguaje comerciales de uso general, caracterizado por su alta fluidez discursiva y su amplia adopción en contextos educativos informales. El modelo se utilizó en su versión accesible públicamente en el momento del estudio (Kasneci et al., 2023).

LLaMA 3 se incluyó como ejemplo de modelo de lenguaje abierto, que permite una mayor experimentación y adaptación en entornos de investigación. Su inclusión responde al interés por analizar el potencial de modelos no comerciales en tareas de evaluación lingüística en ELE (Grattafiori et al., 2024).

Gemini se incorporó al estudio como representante de modelos comerciales recientes con capacidades avanzadas de razonamiento y generación de texto. Su selección permite contrastar el rendimiento de modelos de distintas arquitecturas bajo un mismo procedimiento experimental (Lang et al., 2024).

3.4. Diseño del *prompt*

El diseño del *prompt* constituye un elemento central en el uso de modelos de lenguaje de gran tamaño en tareas educativas, ya que condiciona de manera directa el tipo de respuesta generada por el modelo. Diversos estudios han mostrado que los LLM responden de forma sensible a la formulación explícita de la instrucción, al rol que se les asigna y a los criterios que delimitan la tarea, especialmente en contextos de evaluación y retroalimentación (Brown et al., 2020).

En el ámbito educativo, esta cuestión adquiere una relevancia particular, puesto que el modelo no actúa de forma neutral, sino en función del rol pedagógico que se le atribuye. Investigaciones recientes sobre el uso de LLM en tareas de escritura en lenguas extranjeras subrayan la necesidad de definir de manera explícita dicho rol y el objetivo de la tarea, con el fin de garantizar respuestas coherentes con las expectativas formativas y con los criterios de evaluación empleados (Han et al., 2024).

Con el objetivo de asegurar la comparabilidad de los resultados, se aplicó un único *prompt* a los tres modelos analizados. El *prompt* situaba explícitamente al modelo en el rol de docente de español como lengua extranjera y solicitaba la asignación de nivel de cada texto conforme al Plan Curricular del Instituto Cervantes (PCIC). El enunciado del *prompt* fue el siguiente:

*"*Eres docente de español como lengua extranjera, analiza e indica de qué nivel según el MCER son los textos que te doy al final. Aquí tienes los descriptores de los niveles A1, A2 y B1.

A1: Información general Identificar personas, lugares y objetos Describir personas, lugares, objetos y estados [...] A2: Información general Referirse a acciones y situaciones del pasado (Pretérito indefinido e imperfecto) Opiniones Conocimiento y grado de certeza Expresar y preguntar por el grado de (in)seguridad Obligación, permiso y posibilidad […] B1: Hablar del pasado Expresar acciones habituales. Describir situaciones pasadas. Expresar una acción ocurrida en una unidad de tiempo terminada. [...]

Ofrece retroalimentación sobre el texto de un estudiante de español como lengua extranjera. Ten en cuenta el nivel aproximado de competencia que refleja el texto y comenta: Qué aspectos lingüísticos y discursivos son adecuados para ese nivel. Qué elementos podrían trabajarse para avanzar al nivel siguiente según el MCER. Incluye ejemplos o correcciones breves solo cuando sean necesarios para explicar la sugerencia."

Junto con cada texto, se proporcionaron de forma explícita los descriptores curriculares correspondientes a los niveles A1, A2 y B1, con el propósito de orientar la decisión del modelo hacia criterios lingüísticos y discursivos definidos y evitar clasificaciones basadas únicamente en impresiones globales de fluidez. La salida solicitada incluía dos

elementos: (a) el nivel estimado de la producción escrita y (b) una justificación argumentada de la decisión adoptada, concebida como retroalimentación formativa.

La decisión de emplear un único *prompt* estructurado se apoya en investigaciones previas que han explorado la influencia del diseño de *prompts* en tareas de nivelación de la expresión escrita en ELE. En particular, estudios comparativos sobre distintas formulaciones de *prompt* en modelos de lenguaje como ChatGPT y LLaMA 3 han mostrado que la variación en la instrucción puede afectar tanto la asignación de niveles como la coherencia de las justificaciones proporcionadas, lo que refuerza la necesidad de controlar esta variable cuando se comparan distintos modelos en función de su rendimiento (Cantero Romero, 2024; Cantero Romero & Ortiz Colón, 2025).

3.5. Procedimiento

Cada uno de los 60 textos del corpus fue analizado de manera independiente por los tres modelos, utilizando exactamente el mismo prompt y los mismos descriptores curriculares. Las respuestas generadas se recopilaron y normalizaron para su análisis posterior.

El flujo de análisis seguido fue el siguiente:

texto del aprendiente → prompt con descriptores PCIC → respuesta del modelo → nivel estimado + justificación.

La asignación de nivel generada por cada modelo se comparó posteriormente con el nivel original del texto en el CAES.

3.6. Métricas de evaluación

Para evaluar el rendimiento de los modelos en la tarea de nivelación, se emplearon dos métricas cuantitativas:

- Exactitud (*accuracy*): proporción de textos cuya asignación de nivel coincide con la clasificación original del corpus.

- F1 ponderado por nivel: métrica que combina precisión y exhaustividad, ponderada según la distribución de los niveles, lo que permite valorar el equilibrio del modelo en la clasificación de A1, A2 y B1.

El uso conjunto de ambas métricas permite evaluar tanto el número total de aciertos como la consistencia del modelo en la asignación de los distintos niveles.

3.7. Evaluación cualitativa de la retroalimentación y consistencia

Además de las métricas cuantitativas empleadas para evaluar la precisión de la nivelación, se llevó a cabo una evaluación cualitativa del rendimiento de los modelos, centrada en la calidad de la retroalimentación generada y en la coherencia del proceso

de clasificación. Esta evaluación se concibió como un complemento interpretativo a los resultados cuantitativos y se basó en la observación sistemática de las respuestas producidas por cada modelo.

Para ello, se definieron cuatro dimensiones pedagógicas, alineadas con los objetivos del estudio y con criterios habituales en la evaluación de la expresión escrita en ELE:

a) Fluidez discursiva, entendida como la naturalidad, coherencia y corrección formal del discurso generado por el modelo.

b) Adecuación a los descriptores del PCIC, referida al grado en que la asignación de nivel y la justificación ofrecida se ajustan a los criterios curriculares correspondientes.

c) Claridad de la retroalimentación, relativa a la explicitud, precisión y utilidad pedagógica de los comentarios proporcionados para orientar la revisión del texto.

d) Consistencia en la clasificación, entendida como la coherencia interna entre el nivel asignado y los argumentos empleados para justificar dicha asignación.

Cada una de estas dimensiones se valoró mediante una escala ordinal de cinco puntos, donde 1 corresponde a un rendimiento muy bajo y 5 a un rendimiento muy alto. La puntuación asignada a cada modelo se derivó de una síntesis cualitativa de las respuestas generadas a lo largo del conjunto de textos analizados, atendiendo a patrones recurrentes y evitando valoraciones basadas en casos aislados.

Esta evaluación cualitativa no tiene un carácter inferencial ni pretende sustituir a las métricas cuantitativas, sino caracterizar los perfiles de funcionamiento de los modelos desde una perspectiva pedagógica y facilitar la interpretación integrada de los resultados obtenidos en la tarea de nivelación y retroalimentación.

4. Resultados

Los resultados del estudio se presentan de forma estructurada en cuatro secciones. En primer lugar, se exponen los resultados globales de la nivelación; a continuación, se describen los resultados desagregados por nivel (A1, A2 y B1); posteriormente, se presenta el análisis cualitativo de la retroalimentación; y, finalmente, se sintetizan los resultados por modelo.

4.1. Resultados globales de la nivelación

En primer lugar, se analizó el rendimiento global de los tres modelos de lenguaje en la tarea de nivelación de la expresión escrita, comparando el nivel asignado por cada sistema con la clasificación original de los textos del corpus CAES. Para ello, se calcularon las métricas de exactitud y F1 ponderado, tal como se describe en la sección de Metodología.

Tabla 1.
Resultados globales de exactitud y F1 ponderado por modelo

Modelo	Exactitud	F1 ponderado
ChatGPT	0.390	0.409
Gemini	0.763	0.755
LLaMA 3	0.051	0.088

Fuente: elaboración propia

Los resultados muestran diferencias claras entre los modelos analizados. Gemini obtiene los valores más elevados tanto en exactitud como en F1 ponderado, lo que indica un rendimiento global consistente en la asignación de niveles. ChatGPT presenta un rendimiento intermedio, mientras que LLaMA 3 muestra valores muy bajos en ambas métricas, lo que limita su fiabilidad para la tarea de nivelación automática bajo el diseño metodológico aplicado.

4.2. Resultados por nivel

El análisis desagregado por nivel (tabla 2) permite observar patrones diferenciados en el comportamiento de los modelos en la tarea de nivelación de la expresión escrita. En el nivel A1, los tres modelos obtienen sus mejores resultados relativos, si bien con diferencias marcadas entre ellos. Gemini alcanza los valores más elevados, seguido de ChatGPT, mientras que LLaMA 3 presenta un rendimiento más limitado.

Tabla 2.
Resultados por nivel de F1 por modelo

Nivel	ChatGPT	Gemini	LlaMa 3
A1	0.54	0.89	0.27
A2	0.25	0.61	0.00
B1	0.50	0.78	0.00

Fuente: elaboración propia

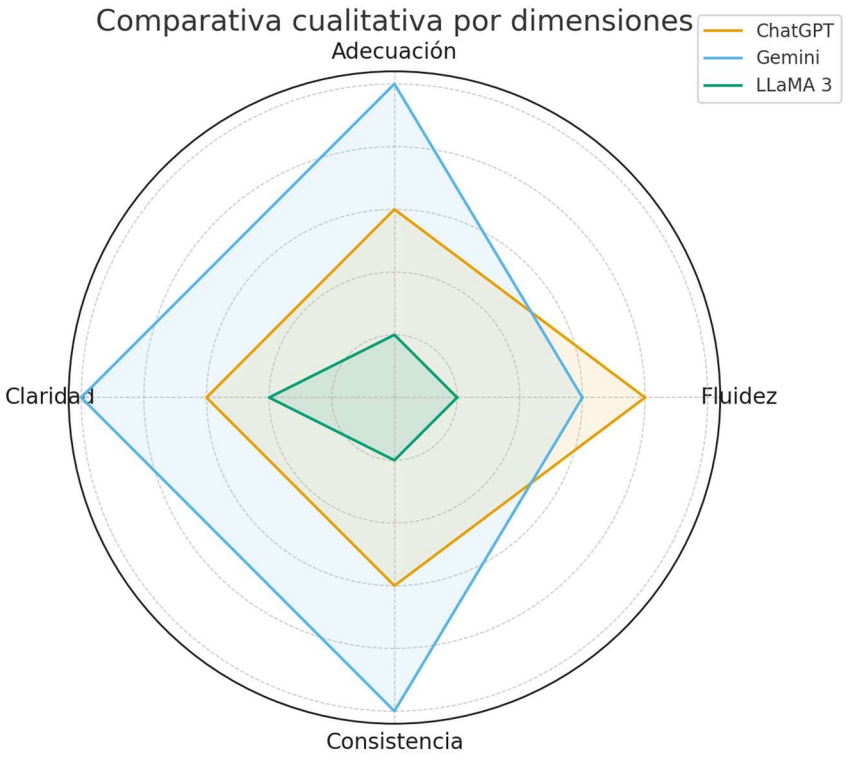
En los niveles A2 y B1, las diferencias entre modelos se acentúan. Gemini mantiene valores de F1 moderados o altos en ambos niveles, lo que indica una mayor estabilidad en la aplicación de los descriptores curriculares a lo largo de la progresión. ChatGPT muestra un descenso progresivo de su rendimiento conforme aumenta la complejidad del nivel, mientras que LLaMA 3 no logra discriminar de forma fiable entre niveles intermedios, con valores de F1 prácticamente nulos. Estos resultados evidencian que la progresión intermedia supone un reto mayor para la nivelación automática mediante modelos de lenguaje.

4.3. Resultados de la retroalimentación: análisis cualitativo por dimensiones

Con el fin de analizar la calidad de la retroalimentación formativa generada por los modelos, se llevó a cabo una evaluación cualitativa basada en cuatro dimensiones pedagógicas: fluidez discursiva, adecuación a los descriptores del PCIC, claridad de la retroalimentación y consistencia en la clasificación, conforme a lo descrito en el apartado de Metodología.

La Figura 1 presenta una comparativa cualitativa por dimensiones mediante un gráfico radar, que permite visualizar de forma integrada el perfil pedagógico de cada modelo a partir de una escala de valoración de 1 (muy baja) a 5 (muy alta).

Figura 1.
Comparativa cualitativa de la retroalimentación por dimensiones pedagógicas



Fuente: elaboración propia

La representación muestra perfiles claramente diferenciados. ChatGPT destaca especialmente en fluidez discursiva, aunque presenta puntuaciones más moderadas en adecuación curricular y consistencia. Gemini muestra un perfil equilibrado y elevado en todas las dimensiones analizadas, mientras que LLaMA 3 obtiene valoraciones bajas de forma sistemática, en coherencia con su bajo rendimiento en la tarea de nivelación.

4.4. Resultados por modelo

El análisis de los resultados por modelo permite sintetizar los perfiles de funcionamiento observados en la tarea de nivelación y en la generación de retroalimentación formativa.

ChatGPT presenta un comportamiento caracterizado por una elevada fluidez discursiva y una buena coherencia textual en sus respuestas. Sin embargo, muestra una tendencia sistemática a sobreestimar el nivel de las producciones escritas, especialmente en textos de transición entre niveles, lo que repercute en una menor precisión de la clasificación y en una retroalimentación a menudo formulada en términos generales y poco vinculados a los descriptores curriculares.

Gemini destaca por un rendimiento estable y consistente tanto en la asignación de niveles como en la calidad de la retroalimentación ofrecida. Sus respuestas muestran una clara adecuación a los descriptores del PCIC y una coherencia elevada entre el nivel asignado y los comentarios proporcionados, lo que se refleja en un perfil equilibrado desde el punto de vista pedagógico.

Por su parte, LLaMA 3 presenta limitaciones significativas en la discriminación entre niveles, especialmente en los niveles intermedios. Las respuestas generadas muestran una escasa alineación con los descriptores curriculares y una elevada variabilidad interna, lo que reduce su utilidad para tareas de nivelación y retroalimentación de la expresión escrita en ELE bajo el diseño metodológico aplicado.

5. Discusión

Los resultados de este estudio confirman que los modelos de lenguaje de gran tamaño presentan comportamientos diferenciados cuando se aplican a tareas de nivelación de la expresión escrita y generación de retroalimentación formativa en el ámbito del español como lengua extranjera. En conjunto, los datos sugieren que el rendimiento de los LLM depende en gran medida de su capacidad para alinearse con descriptores curriculares explícitos y de la complejidad inherente a la progresión entre niveles, especialmente en los estadios intermedios.

Desde el punto de vista de la nivelación, los resultados globales muestran que Gemini ofrece un rendimiento más estable y preciso que los otros modelos analizados, mientras que ChatGPT presenta una fiabilidad intermedia y LLaMA 3 muestra limitaciones significativas. Esta diferencia apunta a que no todos los LLM responden del mismo modo cuando se les exige una clasificación ajustada a criterios curriculares, incluso cuando el diseño del prompt se mantiene constante. En particular, la dificultad de ChatGPT para discriminar con precisión entre niveles intermedios parece estar relacionada con una tendencia a apoyarse en indicadores globales de fluidez discursiva, que no siempre se corresponden con los descriptores específicos del MCER y del PCIC.

El análisis por nivel refuerza esta interpretación. El mejor rendimiento de los modelos en A1 sugiere que la identificación de rasgos propios de niveles iniciales resulta más accesible para los LLM, mientras que la progresión hacia A2 y B1 introduce una complejidad mayor, caracterizada por la coexistencia de estructuras propias de distintos niveles en una misma producción escrita.

En lo que respecta a la retroalimentación formativa, los resultados cualitativos ponen de manifiesto que la fluidez discursiva del modelo no garantiza necesariamente una retroalimentación pedagógicamente útil. ChatGPT genera comentarios naturales y coherentes desde el punto de vista textual, pero con frecuencia recurre a formulaciones generales que limitan su valor orientador. Por el contrario, Gemini ofrece una retroalimentación más estructurada y alineada con los descriptores curriculares, lo que refuerza su potencial como herramienta de apoyo en procesos de revisión guiada. Estos hallazgos coinciden con estudios previos que subrayan la necesidad de evaluar los LLM en tareas educativas concretas y no únicamente en función de su rendimiento general.

Estas diferencias se aprecian con mayor claridad al observar ejemplos concretos de retroalimentación generada por los modelos, que se reproducen a continuación de forma literal.

Ejemplo 1. Retroalimentación generada por ChatGPT (Texto ID_002)

> Nivel asignado: A2
> Aspectos adecuados al nivel: Expresa gustos, deseos y planes; léxico variado y frases coordinadas.
> Para avanzar al nivel siguiente: Introducir pasado simple para narrar experiencias y usar mejor la concordancia: lugar favorito.
> Recomendaciones: Expresa gustos, deseos y planes; léxico variado y frases coordinadas. Introducir pasado simple para narrar experiencias y usar mejor la concordancia: lugar favorito.

Ejemplo 2. Retroalimentación generada por Gemini (Texto ID_002)

> Nivel asignado: A1
> Aspectos adecuados al nivel: Diferenciación de grados de gusto (me encanta, me gusta mucho, me gusta bastante). Uso de descriptores físicos y adjetivos.
> Para avanzar al nivel siguiente: Avanzar a A2 (pasado): introducir el pretérito indefinido. Mejorar concordancia y vocabulario: soy un poco baja (no baha), tengo muchas aficiones. Mejorar el uso de artículos: mi lugar favorito.
> Recomendaciones: Diferenciación de grados de gusto y uso de adjetivos descriptivos. Avanzar a A2 (pasado): introducir el pretérito indefinido y mejorar concordancia, léxico y artículos.

Ejemplo 3. Retroalimentación generada por LLaMA 3 (Texto ID_002)

> Nivel asignado: A1
> Aspectos adecuados al nivel: Presentación personal, descripción de características físicas, expresión de gustos y preferencias, uso de vocabulario básico.
> Para avanzar al nivel siguiente: Uso de conectores y marcadores discursivos (además, sin embargo, por lo tanto), mayor precisión léxica y uso de tiempos verbales (pretérito indefinido, imperfecto).

Estos ejemplos ilustran los distintos perfiles de retroalimentación observados y refuerzan las diferencias identificadas en el análisis cualitativo entre los modelos.

Los resultados obtenidos también aportan elementos relevantes para la reflexión metodológica. En primer lugar, confirman la importancia del diseño del prompt y de la inclusión explícita de descriptores curriculares como mecanismo para orientar el comportamiento del modelo hacia criterios pedagógicos definidos. En segundo lugar, ponen de relieve que la comparación entre modelos debe realizarse bajo condiciones controladas, evitando la variación simultánea de múltiples factores que dificulten la interpretación de los resultados. En este sentido, el enfoque comparativo adoptado permite identificar perfiles de funcionamiento diferenciados y contribuye a reducir el vacío existente en la investigación empírica sobre el uso de LLM en ELE.

Finalmente, es necesario señalar algunas limitaciones del estudio. El análisis se ha centrado en un corpus acotado a niveles iniciales e intermedios, y la evaluación cualitativa de la retroalimentación se basa en una síntesis interpretativa, lo que invita a futuras investigaciones a ampliar tanto el rango de niveles como los procedimientos de validación. Asimismo, sería pertinente explorar diseños longitudinales que permitan evaluar el impacto real de la retroalimentación generada por LLM en el desarrollo de la competencia escrita, así como el potencial de prompts adaptativos ajustados al nivel del aprendiente.

6. Conclusiones y líneas futuras

Los resultados obtenidos permiten confirmar que los LLM pueden desempeñar un papel relevante como herramientas de apoyo en contextos educativos, siempre que su uso se articule a partir de criterios pedagógicos explícitos y bajo supervisión docente.

La principal aportación del trabajo reside en la comparación empírica de distintos modelos a partir de un diseño metodológico controlado y alineado con descriptores curriculares del MCER y del PCIC. Este enfoque ha permitido identificar perfiles de funcionamiento diferenciados y poner de manifiesto que la eficacia de los LLM en ELE no depende únicamente de su fluidez discursiva, sino de su capacidad para ajustarse a criterios de evaluación lingüística y ofrecer una retroalimentación pedagógicamente útil.

Asimismo, el estudio contribuye a reducir el vacío existente en la investigación sobre evaluación automatizada en ELE, un ámbito todavía poco explorado desde perspectivas comparativas y empíricas. Al integrar métricas cuantitativas y un análisis cualitativo de la retroalimentación, se ofrece una visión más completa del potencial y de las limitaciones de los modelos en tareas educativas concretas.

De cara a futuras investigaciones, resulta pertinente profundizar en el diseño de prompts adaptativos que tengan en cuenta el nivel del aprendiente y los objetivos de aprendizaje, así como desarrollar análisis discursivos más finos que permitan evaluar aspectos específicos de la competencia escrita más allá de la asignación global de nivel. Igualmente, se hace necesaria una reflexión continuada sobre las implicaciones éticas y la gobernanza del uso de la inteligencia artificial en ELE, con el fin de garantizar una

integración responsable, transparente y pedagógicamente fundamentada de estas tecnologías en la enseñanza de lenguas.

Referencias

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Cantero Romero, M. V. (2024). Aproximación a un posible uso de ChatGPT para nivelar la expresión escrita en ELE. En F. M. Sirignano, R. Martinez-Roig, & A. López-Padrón (Eds.), Enseñanza y aprendizaje en la era digital desde la investigación y la innovación (pp. 55–64). Octaedro. https://doi.org/10.36006/16433-1

Cantero Romero, M. V., & Ortiz Colón, A. M. (2025). Aproximación al uso del modelo de lenguaje LLaMA 3 para nivelar la expresión escrita en ELE. En A. Palacios-Rodríguez, S. Domene-Martos, R. Piñero Virué, & J. Fernández-Cerero (Eds.), Innovación educativa: Perspectivas y experiencias para la transformación del aprendizaje (pp. 170–185). Dykinson.

Capdehourat, G., Amigo, I., Lorenzo, B., & Trigo, J. (2025). On the effectiveness of LLMs for automatic grading of open-ended questions in Spanish [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2503.18072

Consejo de Europa. (2001). Marco común europeo de referencia para las lenguas: Aprendizaje, enseñanza, evaluación. Instituto Cervantes, Ministerio de Educación, Cultura y Deporte, & Anaya.

Consejo de Europa. (2020). Marco común europeo de referencia para las lenguas: Aprendizaje, enseñanza, evaluación. Volumen complementario. https://rm.coe.int/marco-comun-europeo-de-referencia-para-las-lenguas-aprendizaje-ensenan/1680a52d53

Emirtekin, E. (2025). Large language model-powered automated assessment: A systematic review. Applied Sciences, 15(10), Article 5683. https://doi.org/10.3390/app15105683

Google DeepMind. (s. f.). Gemini 2.5 Pro [Large language model]. https://deepmind.google

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … van der Maaten, L. (2024). The Llama 3 herd of models [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2407.21783

Han, J., Yoo, H., Myung, J., Kim, M., Lim, H., Kim, Y., Lee, T., Hong, H., Kim, J., Ahn, S., & Oh, A. (2024). LLM-as-a-tutor in EFL writing education: Focusing on evaluation of student-LLM interaction. En S. Kumar, V. Balachandran, C. Y. Park, W. Shi, S. A. Hayati, Y. Tsvetkov, N. Smith, H. Hajishirzi, & D. Jurgens (Eds.), Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U) (pp. 284–293). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.customnlp4u-1.21

Instituto Cervantes. (2006). Plan curricular del Instituto Cervantes: Niveles de referencia para el español. Biblioteca Nueva.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Lang, G., Triantoro, T., & Sharp, J. H. (2024). Large language models as AI-powered educational assistants: Comparing GPT-4 and Gemini for writing teaching cases. Journal of Information Systems Education, 35(3), 390–407. https://jise.org/Volume35/n3/JISE2024v35n3pp390-407.html

Leiva-Guerrero, M. V., Araya Zamorano, I., Escobar Collins, R., & Silva Castro, F. (2026). Retroalimentación de aprendizajes con inteligencia artificial generativa en estudiantes universitarios. RIED–Revista Iberoamericana de Educación a Distancia, 29(1), 241–267. https://doi.org/10.5944/ried.45547

Meta AI. (2025). LLam 3 [Modelo de lenguaje de gran tamaño]. https://llama.meta.com/

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. En M. Blanco, H. Olbertz, & V. Vázquez Rozas (Eds.), Corpus y construcciones: Perspectivas hispánicas (pp. 273–301). Universidade de Santiago de Compostela. https://doi.org/10.15304/9788417595876

OpenAI. (s. f.). ChatGPT-5 [Large language model]. https://www.openai.com

Sánchez-Jiménez, D. (2009). La expresión escrita en la clase de ELE. MarcoELE, Revista de Didáctica Español Lengua Extranjera, Suplemento 8. https://academicworks.cuny.edu/ny_pubs/217

You, M., Zhang, J., Lu, C., Alves, A. C. F. de A. R., & Zuo, J. (2026). Explorando o uso de prompts por aprendizes e a qualidade do feedback de IA na escrita em português como segunda língua. Texto Livre, 19, e63188. https://doi.org/10.1590/1983-3652.2026.63188

Nota sobre la utilización de inteligencia artificial

En la elaboración de este artículo se utilizaron herramientas de inteligencia artificial generativa como apoyo puntual a tareas de reformulación lingüística y mejora de claridad expositiva. El diseño del estudio, la selección del corpus, el análisis de los datos, la interpretación de los resultados y la redacción final del manuscrito fueron realizados íntegramente por los autores, que asumen la plena responsabilidad académica y ética del contenido.

Recebido 14/02/2026
Aceite 02/08/2026
Publicado 29/08/2026