ISSN: 2386-9658 https://doi.org/10.21071/arf.v37i.18497

Rendimiento de modelos de IA generativa en tres funciones docentes: análisis de tareas con CRP en ELE

(Effectiveness of Generative AI Models in Three Teaching Functions: Task-Based Analysis of Prepositional Regime Complements in SFL)

WANYUN QI
https://orcid.org/0009-0007-5843-7191
qi.wanyun@163.com
Universidad de Valladolid https://ror.org/01fvbaw18

Fecha de recepción: 28 de julio de 2025
Fecha de aceptación: 19 de diciembre de 2025

Resumen: Ante la creciente demanda de profesionales de español en China, este estudio explora el uso de la inteligencia artificial generativa en la enseñanza de lengua española. Tomando los complementos de régimen preposicional (CRP) como medio de análisis, se evaluó el rendimiento de seis plataformas de IA en tres tareas docentes representativas: resolución de ejercicios, explicación gramatical y diseño de actividades. A partir del manual *Español Moderno*, ampliamente utilizado en universidades chinas, se seleccionaron y estandarizaron ejercicios sobre CRP, que fueron aplicados de forma uniforme a los modelos con instrucciones diferenciadas según las tareas docentes. La evaluación se basó en criterios cualitativos previamente definidos. Los resultados muestran que todos los modelos completan las tareas, pero con diferencias significativas en precisión, claridad y adecuación al contexto educativo. Cada plataforma presenta fortalezas y limitaciones según el tipo de tarea y algunas tienen dificultades para interpretar correctamente el formato o la intención pedagógica. El estudio ofrece datos empíricos útiles sobre las posibilidades y límites de estas herramientas, al tiempo que busca aportar orientaciones para su integración crítica en el aula de ELE en China.

Palabras clave: Inteligencia artificial generativa. Tareas docentes. Complementos de régimen preposicional. Enseñanza de ELE. Modelos de lenguaje.

Abstract: In response to the growing demand for Spanish professionals in China, this study explores the use of generative artificial intelligence in the teaching of Spanish as a foreign language (SFL). Using prepositional dependent constructions (PDC) as the object of analysis, the performance of six AI platforms was evaluated across three representative teaching tasks: exercise resolution, grammatical explanation, and activity design. Based on *Español Moderno*, a textbook widely used in Chinese universities, a set of standardized PDC exercises was selected and applied uniformly to all models, with task-specific instructions. The evaluation followed a set of predefined qualitative criteria. The results show that while all models completed the tasks, there were significant differences in accuracy, clarity, and pedagogical appropriateness. Each platform presents specific strengths and limitations depending on the task type, and some models showed difficulty in correctly interpreting task format or pedagogical intent. This study provides empirical insights into the potential and limitations of these tools and aims to offer guidance for their critical integration into SFL classrooms in China.

Keywords: Generative artificial intelligence (AI). Teaching Tasks. Prepositional Regime Complements. ELE Instruction. Language Models.

1. Introducción

En los últimos años, la inteligencia artificial generativa ha empezado a incorporarse progresivamente a la enseñanza de lenguas extranjeras, ofreciendo herramientas capaces de generar explicaciones, ejemplos y actividades con rapidez y precisión. Modelos como ChatGPT, Claude o DeepSeek abren nuevas posibilidades de apoyo para abordar fenómenos lingüísticos complejos.

Uno de estos fenómenos es el complemento de régimen preposicional (CRP), cuya adquisición representa un reto particular para los estudiantes sin estructuras equivalentes en su lengua materna, como es el caso del chino. Esta ausencia dificulta la transferencia interlingüística y complica la comprensión de su obligatoriedad sintáctica y su vinculación con preposiciones fijas. En este contexto, los modelos de IA pueden ofrecer recursos útiles tanto para los estudiantes como para los docentes, al facilitar recursos lingüísticos variados y apoyar la formulación de explicaciones o actividades.

Este estudio toma el CRP como punto de partida para analizar cómo seis modelos de IA desarrollados por OpenAI, Anthropic y DeepSeek responden ante tres tipos de tareas docentes representativas: (a) resolución de ejercicios sobre CRP; (b) explicación gramatical justificada de las respuestas; (c) diseño de

actividades orales aplicables en clase de ELE. Para ello, se construyó un corpus didáctico con 28 ejercicios extraídos del manual *Español Moderno*, ampliamente utilizado en las universidades chinas. Los ejercicios se aplicaron de forma uniforme a cada modelo, variando únicamente la instrucción proporcionada según el tipo de tarea docente.

El objetivo es evaluar el potencial funcional de estos modelos como herramientas complementarias para la enseñanza de estructuras gramaticales complejas. El artículo se organiza en seis partes, tras esta introducción, se expone el marco teórico (apartado 2), la metodología (apartado 3), los resultados por tipo de tarea (apartado 4), su análisis crítico (apartado 5) y las conclusiones generales (apartado 6).

2. Marco teórico

Este apartado se organiza en tres secciones. En primer lugar, se presentan los fundamentos lingüísticos del complemento de régimen preposicional (CRP) para contextualizar su funcionamiento gramatical. En segundo lugar, se abordan las principales dificultades que plantea su adquisición en el aula de español como lengua extranjera (ELE). Por último, se revisan estudios recientes sobre el uso de la inteligencia artificial (IA) generativa en la enseñanza de lenguas, con atención especial a su aplicación en tareas gramaticales.

2.1. Los CRP desde una perspectiva lingüística

Según la *Nueva gramática de la lengua española* (2009: §36), el complemento de régimen preposicional (CRP) es una función sintáctica desempeñada por un grupo preposicional que depende obligatoriamente del verbo. Esta relación directa con el núcleo verbal lo distingue de los complementos circunstanciales, cuyo uso es opcional y no necesario para completar el significado del verbo.

La formulación más sistemática del CRP se debe a Alarcos Llorach (1966, 1990, 1994), quien observó que ciertos verbos requieren una preposición específica para completar su significado, lo que justifica su tratamiento como complemento obligatorio y no como modificador libre.

Desde entonces, otros autores han profundizado en este fenómeno desde distintas perspectivas. Bosque (1983) cuestiona los límites del CRP como categoría gramatical y señala su grado variable de integración con el verbo. Rojo (1990) propone criterios para distinguirlo de los complementos adverbiales. Santiago-Guervós (2007) ofrece una clasificación terminológica especialmente útil en contextos didácticos. Más recientemente, Casanova Romero (2021) ha actualizado su caracterización, destacando su estabilidad estructural y frecuencia en diferentes registros.

Este marco lingüístico permite comprender mejor la naturaleza del CRP y resulta clave para su abordaje pedagógico en la enseñanza del español como lengua extranjera.

2.2. Dificultades de adquisición y enseñanza de los CRP en ELE

El aprendizaje del CRP representa un reto particular para los estudiantes cuya lengua materna, como el chino, carece de estructuras equivalentes. Esta ausencia obstaculiza la transferencia interlingüística y complica la comprensión de su obligatoriedad, su valor semántico y su comportamiento sintáctico (Canales Muñoz 2021; Casanova Romero 2021).

Los errores más frecuentes observados en corpus de aprendices incluyen la omisión de la preposición, el uso de formas incorrectas o la sustitución por estructuras no adecuadas (Alexopoulou 2006). Estos errores no solo reflejan un conocimiento insuficiente de las estructuras verbales, sino también una comprensión limitada de sus funciones comunicativas.

Desde una perspectiva pedagógica, estas dificultades exigen una intervención que combine el análisis metalingüístico con el uso contextualizado. La simple memorización de listas verbales no basta si no se acompaña de actividades que promuevan la reflexión y la aplicación en situaciones reales. Como señala Corder (1967/1992), la corrección de errores debe orientar al estudiante hacia la reconstrucción activa del sistema lingüístico.

En este contexto, se abren posibilidades para introducir herramientas digitales como apoyo a la enseñanza. La capacidad de los modelos de IA para generar respuestas, formular explicaciones y diseñar actividades plantea interrogantes sobre su aplicabilidad en fenómenos gramaticales complejos como el CRP.

2.3. Aplicaciones de la IA generativa en la enseñanza de lenguas

La inteligencia artificial generativa se ha incorporado recientemente al ámbito educativo como un recurso flexible para apoyar tanto a docentes como a estudiantes. Modelos como ChatGPT permiten automatizar tareas habituales en clase: generar ejemplos, explicar estructuras gramaticales o crear ejercicios adaptados al nivel del estudiante.

Hernández (2021) señala que estos modelos pueden facilitar el acceso a contenidos complejos y fomentar la autonomía del aprendiz. Ribes-Lafoz y Navarro Colorado (2023) exploran su integración en educación superior para mejorar la producción escrita y apoyar la corrección de errores. Fernández (2024) propone actividades concretas con ChatGPT para trabajar el léxico en niveles B1 y B2. Además, la revisión de Li *et al.* (2024) identifica tres líneas principales de

uso: diseño de materiales, asistencia metalingüística y simulación de interacciones orales.

En conjunto, estos estudios coinciden en que la IA puede aportar beneficios reales cuando se emplea con objetivos pedagógicos claros, aunque también advierten sobre la necesidad de evaluar su adecuación a cada entorno de enseñanza.

3. Diseño metodológico

Este estudio adopta un enfoque cuasiexperimental con el objetivo de evaluar el desempeño de seis plataformas de IA generativa en tareas docentes relacionadas con los complementos de régimen preposicional (CRP) en la enseñanza del español como lengua extranjera (ELE). El propósito no es medir el impacto en el aprendizaje del alumnado, sino analizar la capacidad funcional de los modelos para responder a distintos tipos de tareas docentes. Por esta razón, el diseño no incluye grupo de control y se basa en la aplicación de entradas normalizadas.

El corpus experimental está compuesto por 28 ejercicios extraídos del manual *Español Moderno*, de uso habitual en universidades chinas. Los mismos ejercicios se aplicaron en tres fases sucesivas. En cada fase solo varió la instrucción proporcionada a los modelos, con el fin de simular tres tareas docentes diferentes: resolución de ejercicios, explicación gramatical y diseño de actividades orales.

3.1. Plataformas de IA y condiciones de prueba

Se seleccionaron seis modelos de IA generativa pertenecientes a tres compañías diferentes (OpenAI, Anthropic y DeepSeek), cada una representada por una versión básica y una versión avanzada:

Compañía	**Versión básica**	**Versión avanzada**	**Modalidad de acceso**
OpenAI	GPT-4o mini	GPT-4o (versión completa)	Gratuita / De pago
Anthropic	Claude Sonnet 3.7	Claude Pro (Sonnet 3.7 con pensamiento extendido y acceso a internet)	Gratuita / De pago
DeepSeek	DeepSeek	DeepSeek+ (con pensamiento extendido y acceso a internet)	Ambas gratuitas

Tabla 1. Modelos de IA evaluados

Las pruebas se realizaron entre el 27 de abril y el 2 de mayo de 2025[1]. Todas las plataformas fueron evaluadas en condiciones estables de conexión y con el mismo conjunto de instrucciones y materiales.

3.2. Ejercicios gramaticales[2]

El corpus incluye 28 ejercicios centrados en el uso de los CRP, cada ejercicio presenta una situación concreta en la que el estudiante debe identificar, completar o seleccionar la forma adecuada del CRP correspondiente.

Los ejercicios se clasifican en cuatro tipos funcionales:

Tipo	**Descripción breve**	**Elementos evaluados**	**Nº de ejercicios**
1	Selección sin contexto	Identificación del régimen preposicional sin apoyo semántico; solo una opción es correcta	10
2	Selección con contexto	Comprensión sintáctica y semántica contextualizada; solo una opción es correcta	10
3	Completar huecos con opciones	Dominio estructural y morfológico (verbo + preposición + conjugación)	5
4	Selección con distractores válidos	Todas las opciones son formalmente correctas; solo una es adecuada según el contexto	3

Tabla 2. Tipos de ejercicios utilizados en el estudio

La distribución de los ejercicios responde a la frecuencia y al nivel de complejidad de los ejercicios en el manual de origen: los tipos 3 y 4 requieren una elaboración más específica y aparecen con menor frecuencia en los materiales originales. Los mismos ejercicios se reutilizaron en las tres funciones docentes del estudio (resolución, explicación y diseño), modificando únicamente la instrucción según la tarea simulada.

[1] Debido a una interrupción nacional del suministro eléctrico el 28 de abril de 2025 en España, algunas sesiones fueron reprogramadas al día siguiente, sin afectar a la integridad del conjunto de datos.

[2] Los ejercicios utilizados en este estudio proceden del cuestionario diseñado para la tesis doctoral de la autora, actualmente en desarrollo en la Universidad de Valladolid. En este artículo se presenta una selección adaptada de dichos ejercicios, conforme al objetivo y enfoque metodológico específicos del estudio.

3.3. Establecimiento de un criterio de referencia nativo

Con el fin de establecer un valor de referencia, se recopilaron las respuestas de 33 hablantes nativos de español (nacionalidad española), quienes resolvieron el mismo conjunto de 28 ejercicios. La información sociodemográfica se registró de forma descriptiva, con el fin de fijar una tasa de acierto nativa como parámetro de comparación para los modelos de IA.

Este grupo no participó en la experimentación con IA y se utilizó únicamente como base comparativa para la tasa media de aciertos y los patrones generales de respuesta. Conviene subrayar que este estudio se centra en el análisis del rendimiento de plataformas de IA en tareas docentes. Los hablantes nativos actúan solo como grupo de referencia, ya que el objetivo final es apoyar la enseñanza de ELE a sinohablantes.

3.4. Tareas docentes simuladas, instrucciones y criterios de evaluación

El estudio se organizó en tres fases, cada una asociada a una tarea docente distinta, manteniendo constantes los ejercicios lingüísticos.

3.4.1. Tipos de tareas docentes

Se simularon tres tareas representativas del aula de ELE: (a) resolución directa, centrada en la selección de la respuesta correcta; (b) explicación gramatical, que requiere justificar la respuesta y descartar las opciones incorrectas; (c) diseño de actividades, orientado a la creación de tareas orales aplicables al aula.

Las dos primeras tareas pueden corresponder tanto a docentes como a estudiantes, mientras que la tercera se asocia principalmente a la planificación didáctica.

3.4.2. Diseño de instrucciones

Para cada tarea se elaboró una instrucción específica, aplicada de forma constante a los 28 ejercicios. En la primera fase se presentó únicamente el ejercicio. En la segunda se solicitó explícitamente la explicación de la respuesta. En la tercera se pidió el diseño de una actividad oral dirigida a estudiantes de ELE.

Las instrucciones fueron redactadas de manera clara y concisa, con el fin de reproducir condiciones reales de uso educativo.

3.4.3. Criterios de evaluación

Cada tipo de tarea fue evaluado mediante un conjunto de criterios definidos antes del análisis y aplicados de forma sistemática por un mismo revisor, lo que garantiza la coherencia interna y la comparabilidad entre plataformas.

Rol docente	Criterios establecidos del análisis
Resolutor directo	Exactitud en la respuesta; estabilidad de respuesta; claridad expositiva; ejemplificación; tono didáctico; estructura textual
Asistente explicativo	Estructura y completitud; precisión lingüística; claridad expresiva; tono didáctico; calidad y adecuación de los ejemplos
Diseñador de actividades	Estructura y completitud; extracción y uso de CRP; adecuación al objetivo; claridad expositiva; tono didáctico; aplicabilidad; creatividad

Tabla 3. Resumen de funciones docentes simuladas y sus criterios de evaluación

En el caso de la resolución directa, se priorizó la precisión en la selección y la claridad de las respuestas. Para el rol de asistente explicativo, se valoró especialmente la estructura argumentativa, la coherencia lingüística y la ejemplificación adecuada. En cuanto al diseño de actividades, se evaluó la adecuación al contexto de aula, el uso pertinente de los CRP y el potencial didáctico de las propuestas.

Algunas dimensiones permiten una valoración cuantitativa (como la exactitud), mientras que otras se basan en una evaluación cualitativa. Los descriptores detallados de cada nivel se presentan en el Anexo I. Este sistema de evaluación está diseñado para identificar diferencias funcionales entre modelos en contextos reales de enseñanza de ELE.

4. Análisis de los resultados

Este apartado presenta los resultados obtenidos por las seis plataformas en las tres tareas planteadas: resolución de ejercicios, explicación gramatical y diseño de actividades. La evaluación se organiza por tarea, según los criterios definidos previamente (véase el apartado 3.4.3). No se realiza comparación cruzada entre tareas.

4.1. Evaluación del rendimiento en la tarea 1: resolución de ejercicios

4.1.1. Objetivo y planteamiento de la tarea

En esta fase, los modelos actuaron como resolutores directos. Recibieron los 28 ejercicios sin instrucciones adicionales, con el fin de observar su capacidad para seleccionar la opción correcta de manera espontánea.

Los criterios principales fueron: (a) precisión, entendida como coincidencia con la respuesta establecida; (b) estabilidad, es decir, coherencia en responder en distintos momentos. También se observaron otras dimensiones cualitativas, no solicitadas pero relevantes: (c) claridad expositiva, (d) ejemplificación, (e) tono didáctico y (f) estructura textual.

4.1.2. Comparación de aciertos entre IA e hispanohablantes

Los resultados se comparan con los obtenidos por 33 hablantes nativos de España, quienes completaron los mismos ejercicios.

Tipo de ejercicio/ Resultado	**Hispanohablantes**	**IA global (promedio)**
Promedio total	86,69 %	94,78 %
Tipo 1: selección sin contexto	93,93 %	100,00 %
Tipo 2: selección con contexto	96,37 %	96,67 %
Tipo 3: completación estructural	50,93 %	80,00 %
Tipo 4: elección con distractores	89,90 %	100,00 %

Tabla 4. Porcentaje de aciertos de hispanohablantes *versus* IA por tipo de ejercicio

Aunque el promedio de aciertos de las plataformas (94,78 %) supera al del grupo nativo (86,69 %), esta diferencia no debe interpretarse de forma lineal. El análisis de chi-cuadrado según los tipos de ejercicio no mostró diferencias significativas ($\chi^2 = 15$, $p = 0{,}241$). En cambio, el análisis sobre la distribución global de aciertos sí reveló una diferencia estadísticamente significativa ($\chi^2 = 45{,}043$, $p = 0{,}003$), atribuida a la alta concentración de respuestas correctas del 100 % en las plataformas, frente a una distribución más variada en el grupo nativo.

Este resultado sugiere que, aunque las plataformas tienen un alto rendimiento medio, su estilo de respuesta es menos flexible, lo que podría limitar su eficacia en contextos intermedios o con ambigüedad pragmática.

4.1.3. Análisis del resultado

Se toma como ejemplo el ejercicio "casarse […] alguien" para ilustrar el proceso de la evaluación. Esta pregunta fue seleccionada por tres motivos: pertenece al tipo más frecuente (selección sin contexto), incluye un verbo preposicional de uso claro, y no presenta opciones distractoras que pudieran generar ambigüedad. Aunque solo se muestra un caso, la misma metodología se aplicó de forma sistemática a todas las respuestas.

GPT-4o ofrece una explicación clara: "El verbo casarse se construye con la preposición con...". Claude Sonnet presenta una formulación similar. Claude Pro usa un estilo más formal: "cuando va seguido de una persona". GPT-4o mini responde: "La preposición correcta es con", sin desarrollo adicional. DeepSeek organiza la respuesta con subtítulos, pero en tono técnico. DeepSeek+ emplea inglés: "The reflexive verb casarse requires the preposition con...", lo que reduce la accesibilidad para aprendices sinohablantes.

Este ejemplo sirve de base para el sistema de evaluación empleado en las tablas siguientes. Por razones de espacio, no se presentan todas las respuestas, pero los criterios se aplicaron de forma sistemática en los tres roles docentes.

Plataforma	Exac.	Clar.	Estab.	Ejemp.	Tono	Estruc.
GPT-4o	100 %	Alta	Alta	Alta	Alta	Alta
GPT-4o mini	95 %	Media	Media	Media	Media	Media
Claude Sonnet	95 %	Alta	Media	Alta	Alta	Alta
Claude Pro	90 %	Alta	Media	Media	Media	Alta
DeepSeek	95 %	Baja	Baja	Media	Baja	Media
DeepSeek+	90 %	Baja	Baja	Baja	Baja	Baja

Tabla 5. Evaluación cualitativa del rendimiento de las plataformas en la tarea 1

Nota: Para facilitar la presentación, en el encabezado de la tabla se han utilizado las siguientes abreviaciones: Exac. = Exactitud; Clar. = Claridad expositiva; Estab. = Estabilidad; Ejemp. = Ejemplificación; Tono = Tono pedagógico; Estruc. = Estructura.

4.1.4. Conclusión de la tarea 1

Todas las plataformas lograron buenos niveles de precisión. GPT-4o fue el único modelo con 100 % de aciertos y estabilidad total. Claude Sonnet y Claude Pro también presentaron respuestas consistentes. GPT-4o mini mantuvo un rendimiento aceptable, aunque menos elaborado. DeepSeek y DeepSeek+ ofrecieron respuestas técnicas, con menor claridad y tono pedagógico.

Aunque no se solicitaba justificación, varios modelos la incorporaron de manera espontánea, lo que sugiere cierta utilidad pedagógica.

4.2. Evaluación del rendimiento en la tarea 2: asistencia en la explicación

4.2.1. Objetivo y planteamiento de la tarea

En esta tarea, los modelos actuaron como asistentes explicativos. Se les pidió justificar la elección correcta para cada ejercicio, explicando la construcción verbo y preposición, el motivo de la elección según el enunciado y uno o dos ejemplos ilustrativos.

La evaluación se realizó con base en cinco criterios previamente definidos para esta tarea, según lo establecido en el apartado 3.4.3: (a) estructura y completitud, (b) precisión lingüística, (c) claridad expositiva, (d) tono docente y (e) calidad de los ejemplos. Estos criterios fueron formulados teniendo en cuenta las necesidades didácticas específicas de los estudiantes sinohablantes, para

quienes los CRP representan un área de dificultad frecuente en la producción y comprensión escrita. Todas las respuestas fueron evaluadas aplicando de forma constante los criterios definidos para esta tarea.

4.2.2. Análisis del resultado

La tabla siguiente resume los resultados obtenidos de esta tarea.

Plataforma	Estruct. y compl.	Precisión	Claridad	Tono	Ejemplos
GPT-4o	Alta	Alta	Alta	Alta	Alta
GPT-4o mini	Media	Media	Media	Media	Media
Claude Sonnet	Alta	Alta	Alta	Alta	Alta
Claude Pro	Alta	Alta	Alta	Media	Alta
DeepSeek+	Baja	Media	Baja	Baja	Baja
DeepSeek	Baja	Media	Baja	Baja	Baja

Tabla 6. Evaluación cualitativa del rendimiento de las plataformas en la tarea 2

Nota: Estruct. y compl. = Estructura y completitud; Precisión = Precisión lingüística; Claridad = Claridad expresiva; Tono = Tono didáctico; Ejemplos = Calidad de ejemplos.

Claude Sonnet y Claude Pro siguieron el formato propuesto, con explicaciones claras y bien organizadas. GPT-4o ofreció razonamientos correctos, aunque con lenguaje técnico en algunos casos. GPT-4o mini cumplió con lo básico, pero sin mucha profundidad. DeepSeek y DeepSeek+ mostraron falta de cohesión y uso ocasional del inglés.

En los ejercicios del tipo 4, los modelos no lograron diferenciar entre la corrección gramatical y la adecuación pragmática al contexto, lo que se analizará con más detalle en el apartado de discusión (véase 5.2).

4.2.3. Conclusión de la tarea 2

Claude Sonnet y Claude Pro fueron los más sólidos en esta tarea, con explicaciones bien estructuradas, adecuadas y útiles. GPT-4o razonó bien, pero con menor consistencia pedagógica. GPT-4o mini ofreció respuestas simples y claras. DeepSeek y DeepSeek+ mostraron limitaciones importantes en coherencia y tono docente. Todos los modelos fallaron al manejar ítems con sensibilidad pragmática, lo cual limita su eficacia en tareas de explicación más complejas.

4.3. Evaluación del rendimiento en la tarea 3: diseño de actividades didácticas

4.3.1. Objetivo y planteamiento de la tarea

En esta ocasión, las plataformas asumieron el rol de diseñadoras de actividades didácticas, elaborando tareas orales centradas en los CRP previamente evaluados, dirigidas a estudiantes universitarios chinos de Filología Hispánica.

En esta fase, se pidió a los modelos diseñar actividades orales basadas en los CRP evaluados. Las instrucciones fueron: crear una tarea oral completa (sin solución o explicación); usar de forma activa el CRP correspondiente; adaptarla a estudiantes chinos de Filología Hispánica.

Las tareas generadas se analizaron según los siete criterios definidos previamente en el apartado 3.4.3: estructura y completitud, extracción y uso de CRP, adecuación al objetivo, claridad expositiva, tono docente, aplicabilidad en el aula y creatividad.

4.3.2. Análisis del resultado

En la tabla se resumen los resultados siguiendo los criterios establecidos.

Plataforma	**Estruct. y compl.**	**Uso de CRP**	**Objetivo**	**Claridad**	**T. / A. / C.**
GPT-4o	Alta	Media	Media	Alta	Media
GPT-4o mini	Alta	Media	Media	Media	Media
Claude Sonnet	Alta	Alta	Alta	Alta	Alta
Claude Pro	Alta	Alta	Alta	Alta	Alta
DeepSeek+	Media	Baja	Media	Baja	Baja
DeepSeek	Media	Baja	Media	Media	Baja

Tabla 7. Evaluación cualitativa del rendimiento de las plataformas en la tarea 3

Nota: En el encabezado de la tabla se han empleado abreviaciones para facilitar la presentación: Estruct. y compl. = Estructura y completitud; Uso de CRP = Extracción y uso de CRP; Objetivo = Adecuación al objetivo; Claridad = Claridad expositiva; T. / A. / C. = Tono didáctico, aplicabilidad en aula y creatividad en el diseño.

Claude Sonnet y Claude Pro presentaron propuestas completas, con objetivo claro, desarrollo, consigna y cierre. GPT-4o diseñó actividades aplicables, pero menos variadas. GPT-4o mini fue más general. DeepSeek y DeepSeek+ no integraron bien el CRP y usaron lenguaje técnico.

4.3.3. Conclusión de la tarea 3

En esta tarea, Claude Sonnet y Claude Pro lograron diseñar actividades completas, funcionales y adaptadas al aula. GPT-4o cumplió los requisitos básicos del diseño, aunque con menor variedad en las propuestas. GPT-4o mini ofreció actividades funcionales pero más generales. DeepSeek y DeepSeek+ mostraron limitaciones en la integración del contenido gramatical y en la claridad de las instrucciones.

5. Discusión

Este apartado presenta los hallazgos principales del estudio, organizados en tres partes: el rendimiento general de las plataformas, su adecuación a los distintos roles docentes simulados y los factores externos que afectan su aplicabilidad en el aula. La discusión se basa en los resultados descritos en el apartado anterior, con referencia tanto a los datos cuantitativos como a las observaciones cualitativas por criterio.

5.1. Desempeño general de las plataformas

El análisis del rendimiento se ha centrado en tres funciones docentes simuladas: resolución directa, explicación asistida y diseño de actividades. Cada función fue evaluada según un conjunto de criterios preestablecidos, como la precisión, la claridad expositiva, la adecuación didáctica, la estructura textual y el tono docente, etc.

Claude Pro y Claude Sonnet ofrecieron los resultados más sólidos. Claude Pro se destacó por su organización clara y propuestas listas para el aula. Claude Sonnet combinó claridad, creatividad y un estilo cercano al discurso del profesor. GPT-4o mostró respuestas precisas y bien formuladas, aunque con un estilo técnico menos adaptado al entorno educativo. GPT-4o mini fue más directo y rápido, pero con desarrollos limitados. DeepSeek y DeepSeek+ cumplieron con los requisitos mínimos, aunque presentaron errores frecuentes, uso de inglés y escasa adaptación a estudiantes de ELE.

Aunque la media general de aciertos de las plataformas fue superior a la de los hablantes nativos (94,78 % frente a 86,69 %), el análisis por chi-cuadrado mostró una diferencia significativa en la distribución ($\chi^2 = 45{,}043$, $p = 0{,}003$). Esta diferencia se debe a la alta concentración de aciertos del 100 % en algunas plataformas, mientras que las respuestas humanas fueron más variadas. Este patrón sugiere que los modelos funcionan bien en ejercicios cerrados, pero muestran menor flexibilidad ante contextos ambiguos o intermedios. Esta limitación debe ser considerada al valorar su uso en entornos reales de enseñanza.

5.2. Recomendaciones según tipo de tarea docente

Los resultados indican que cada plataforma rinde de forma distinta según la tarea. Por ello, se recomienda seleccionar el modelo en función del objetivo didáctico específico.

En tareas de respuesta directa, como la selección del CRP correcto, GPT-4o ofreció la mayor precisión y estabilidad. Claude Sonnet y Claude Pro también obtuvieron buenos resultados, con diferencias menores en la presentación.

En tareas explicativas, Claude Sonnet fue el más claro, con un estilo accesible para estudiantes. Claude Pro ofreció explicaciones bien organizadas, aunque algo largas. GPT-4o razonó de forma adecuada, pero con un tono técnico que puede dificultar la comprensión.

En los ejercicios con mayor sensibilidad pragmática, como los de tipo 4, todos los modelos fallaron en diferenciar entre corrección gramatical y adecuación contextual. Por ejemplo, en "Dudamos […] la veracidad…," aunque "dudar en" "dudar entre" son válidas desde un punto de vista gramatical, solo "dudar de" es pragmáticamente adecuada. Sin una indicación explícita sobre el tipo de criterio esperado, todas las plataformas trataron otras opciones como errores, lo que revela una falta de competencia contextual. Por esta razón, al usar este tipo de ejercicios en clase, el docente debe explicar claramente la lógica del ejercicio para evitar interpretaciones erróneas.

En el diseño de actividades orales, Claude Pro presentó las propuestas más completas, claras y aplicables. Claude Sonnet fue más creativo, lo que puede ser útil en clases comunicativas. GPT-4o diseñó actividades funcionales, aunque menos variadas. GPT-4o mini ofreció tareas simples pero aceptables. DeepSeek y DeepSeek+ mostraron deficiencias en la integración del contenido gramatical y en la redacción de instrucciones.

En resumen, GPT-4o es más útil para tareas que requieren precisión. Claude Sonnet se adapta mejor a actividades de explicación. Claude Pro destaca en el diseño de tareas docentes. Es importante no aplicar la misma herramienta a todas las situaciones, sino adaptarla según el contexto y objetivo.

5.3. Factores de accesibilidad y condiciones de uso

Más allá del rendimiento lingüístico, también deben considerarse las condiciones técnicas de uso, como el tiempo de respuesta, el acceso a la plataforma, los límites de uso diario y las restricciones por país.

Respecto a la velocidad, GPT-4o mini fue el más rápido, con un promedio de 8.49 segundos. Le siguieron GPT-4o (16,28 s), Claude Pro (17,11 s) y Claude Sonnet (19,72 s), todos con velocidades medias. DeepSeek tuvo un tiempo medio

de 23,66 segundos, clasificado como medio-lento, mientras que DeepSeek+ fue el más lento, con 54,15 segundos.

En cuanto al acceso, DeepSeek y DeepSeek+ son gratuitos, sin límites de uso ni bloqueos por red, lo que los hace accesibles para estudiantes con menos recursos. Claude Sonnet también es gratuito, pero con una limitación importante: tras diez mensajes, es necesario esperar unas cuatro horas para seguir usándolo. GPT-4o mini también es gratuito, aunque no está disponible en todos los países. Por el contrario, GPT-4o (20 USD/mes) y Claude Pro (15 USD/mes) son versiones de pago que ofrecen mayor estabilidad y continuidad.

Cabe señalar que, en China continental, tanto la serie GPT como la serie Claude presentan restricciones de red, lo que dificulta su uso normal sin herramientas técnicas adicionales. Esta barrera limita su aplicabilidad en contextos institucionales sin infraestructura digital adecuada.

En conjunto, se observa que ninguna plataforma es superior en todos los aspectos. Cada una presenta fortalezas y limitaciones que deben valorarse según el entorno y el tipo de tarea. La elección debe ser estratégica y no generalizada. Además de la calidad lingüística, se debe valorar la sostenibilidad técnica y económica de la herramienta para su integración real en cursos de ELE.

Conclusión

Este estudio ha analizado el rendimiento de seis plataformas de inteligencia artificial en tareas docentes, utilizando los complementos de régimen preposicional (CRP) como medio del análisis. Los resultados muestran que todas las plataformas completaron las tareas asignadas, aunque con diferencias notables en cuanto a precisión, claridad y adecuación didáctica. Además, el desempeño varió según la naturaleza de la tarea, lo que confirma la necesidad de una aplicación diferenciada en función del objetivo pedagógico.

En términos generales, Claude Sonnet y Claude Pro ofrecieron un lenguaje más adecuado en tareas de explicación y diseño, mientras que GPT-4o destacó por su precisión en tareas de respuesta directa. DeepSeek y DeepSeek+ mostraron ciertas limitaciones, pero cumplen con las funciones básicas y presentan ventajas relevantes como el acceso libre, sin restricciones de red ni límites de uso, lo que los hace viables en contextos con recursos limitados.

Es importante señalar que ninguna plataforma sobresale en todos los aspectos. Cada modelo presenta fortalezas y debilidades que deben considerarse al integrarlos en entornos educativos. La elección debe responder a los objetivos específicos de cada tarea, al perfil del alumnado y a las condiciones técnicas de uso. Además de la calidad lingüística, se deben valorar factores como la

sostenibilidad económica y la viabilidad técnica para garantizar una integración real y continua en cursos de ELE.

Este trabajo no pretende establecer un ranking entre las plataformas, sino identificar sus puntos fuertes en distintos contextos docentes, con el fin de orientar su uso como herramienta de apoyo en la enseñanza del español en universidades chinas. La meta no es encontrar un único modelo superior, sino proponer una combinación estratégica según las necesidades concretas del aula.

Los datos analizados ofrecen una base inicial para comprender el potencial de la inteligencia artificial en la enseñanza de lenguas. Aunque este estudio se ha centrado en tareas simuladas y ha priorizado un enfoque cualitativo, sería conveniente que futuras investigaciones incorporaran análisis cuantitativos y se llevaran a cabo en aulas reales. Así se podrá evaluar con mayor profundidad el papel de estas herramientas en entornos educativos diversos.

Referencias bibliográficas

ALARCOS LLORACH, Emilio, "Verbo transitivo, verbo intransitivo y estructura del predicado". En: *Archivum*, 16, 1966, pp. 5-17.

——, *La noción de suplemento*. Gobierno de La Rioja, Consejería de Educación, Cultura y Deportes, 1990.

——, *Gramática de la lengua española*. Madrid: Espasa Calpe, S. A., 1994.

ALEXOPOULOU, Angélica, "Los criterios descriptivo y etiológico en la clasificación de los errores del hablante no nativo: una nueva perspectiva". En: *Porta Linguarum*, 5, 2006, pp. 17-36. Disponible en línea en: https://dialnet.unirioja.es/servlet/articulo?codigo=1709311

BOSQUE, Ignacio, "Dos notas sobre el concepto de 'suplemento' en la gramática funcional". En: *Dicenda: Estudios de lengua y literatura españolas*, 2, 1983, pp. 147-156.

CANALES MUÑOZ, L. A., "El complemento de régimen preposicional en la enseñanza de español como lengua extranjera". En: *E-leeando*, 21, 2021, pp. 1-107. Disponible en línea en: https://ebuah.uah.es/dspace/handle/10017/49133

CASANOVA ROMERO, V., *El complemento de régimen verbal: construcción y distribución en español actual*. Tesis doctoral, Université de Montréal, 2021.

CORDER, S. P., "La importancia de los errores del que aprende una lengua segunda". En: *La adquisición de las lenguas extranjeras*, 1992, pp. 31-40. Visor. Trabajo original publicado en 1967. Disponible en línea en: https://eric.ed.gov/?id=ED019903

FERNÁNDEZ, M. G., "Inteligencia artificial y léxico: Una propuesta con ChatGPT para niveles B1 y B2 en la enseñanza del español como lengua

extran-jera". En: *RILEX, Revista sobre investigaciones léxicas*, 2024. Disponible en línea en: https://revistaselectronicas.ujaen.es/index.php/RILEX/article/view/9111

HERNÁNDEZ, J. C. E., "La Inteligencia Artificial y la Enseñanza de lenguas: una aproximación al tema". En: *Decires*, 21, 25, 2021, pp. 29-44. Disponible en línea en: https://decires.cepe.unam.mx/index.php/decires/article/view/3

LI, B., *et al.*, "A systematic review of the first year of publications on ChatGPT and language education: Examining research on ChatGPT's use in language learning and teaching". En: *Computers and Education: Artificial Intelligence*, 2024, 100266. Disponible en línea en: https://doi.org/10.1016/j.caeai.2024.100266

REAL ACADEMIA ESPAÑOLA, *Nueva gramática de la lengua española (vol. 2)*. Madrid: Espasa, 2009.

RIBES-LAFOZ, M.; NAVARRO COLORADO, B., "Aprovechamiento de *ChatGPT* en la enseñanza de lengua extranjera en educación superior". En: Ortega-Sánchez, D.; López-Padrón, A. (eds.), *Educación y sociedad: claves interdisciplinares*. Barcelona: Octaedro, 2023, pp. 1264-1275. Disponible en línea en: http://hdl.handle.net/10045/139198

ROJO, G., "En torno a los complementos circunstanciales". En: *Lecciones del I y II Cursos de Lingüística Funcional*. Oviedo: Universidad de Oviedo, 1985, pp. 181-191.

——, "Sobre los complementos adverbiales". En: *Homenaje al profesor Francisco Marsá*, 1990, pp. 153-171.

SANTIAGO-GUERVÓS, F. J. d., *El complemento (de régimen) preposicional*. Madrid: Arco Libros, 2007.

ISSN: 2386-9658 https://doi.org/10.21071/arf.v37i.18497

Anexo

Criterios cualitativos de evaluación por funciones docentes

Rol docente	Criterio	Nivel	Descripción
Resolutor directo	Exactitud en la respuesta	alta	La respuesta es correcta y se ajusta plenamente a la norma gramatical del CRP.
		media	La respuesta es parcialmente correcta o muestra dudas estructurales menores.
		baja	La respuesta es incorrecta o muestra errores gramaticales evidentes.
	Estabilidad de respuesta	alta	El modelo mantiene coherencia en respuestas a lo largo del corpus.
		media	Se detectan variaciones puntuales o inconsistencias leves.
		baja	Las respuestas cambian de forma significativa sin justificación.
	Claridad expositiva	alta	El contenido se presenta de forma clara, directa y comprensible para aprendices.
		media	La explicación es entendible, pero incluye frases ambiguas o confusas.
		baja	El contenido es difícil de entender o usa tecnicismos innecesarios.
	Ejemplificación	alta	Incluye ejemplos relevantes, variados y bien contextualizados.
		media	Incluye ejemplos básicos o repetitivos con valor limitado.
		baja	No incluye ejemplos o los ejemplos son irrelevantes o incorrectos.
	Tono didáctico	alta	El modelo simula un tono de enseñanza claro y adecuado para el aula.

		media	El tono es neutral o informal, pero no inadecuado.
		baja	El tono es técnico, impersonal o inapropiado para el contexto educativo.
	Estructura textual	alta	La respuesta sigue un orden lógico y contiene introducción, desarrollo y cierre.
		media	La estructura está presente pero poco marcada o incompleta.
		baja	La organización textual es caótica o inexistente.
Asistente explicativo	Estructura y completitud	alta	La propuesta presenta una estructura completa y coherente, con todos los elementos clave (objetivo, desarrollo, consigna, evaluación, etc.) claramente definidos.
		media	La estructura está presente, pero es parcial o poco detallada; algunos elementos clave están ausentes o desarrollados de forma superficial.
		baja	La propuesta carece de una estructura clara o presenta una organización fragmentada e incompleta, lo que dificulta su aplicación didáctica.
	Precisión lingüística	alta	Uso gramatical y léxico correcto, sin errores evidentes. Explicaciones claras y coherentes.
		media	Presenta algunos errores menores que no impiden la comprensión general.
		baja	Contiene errores gramaticales, léxicos o semánticos que afectan la claridad y la corrección.
	Claridad expresiva	véanse criterios anteriores	

	Tono didáctico	véanse criterios anteriores	
	Calidad y adecuación de los ejemplos	alta	Los ejemplos son correctos, relevantes para el verbo y su uso con CRP, y están adaptados al nivel de los estudiantes. Apoyan de forma clara la comprensión de la explicación.
		media	Los ejemplos son correctos pero genéricos o poco contextualizados. Aunque ilustran el uso del CRP, su utilidad didáctica es limitada.
		Baja	Los ejemplos son incorrectos, irrelevantes o confusos. No se ajustan al contexto del verbo o pueden inducir a error, dificultando el aprendizaje.
Diseñador de actividades	Estructura y completitud	alta	La propuesta presenta una estructura completa y coherente, con todos los elementos clave (objetivo, desarrollo, consigna, evaluación, etc.) claramente definidos.
		media	La estructura está presente, pero es parcial o poco detallada; algunos elementos clave están ausentes o desarrollados de forma superficial.
		baja	La propuesta carece de una estructura clara o presenta una organización fragmentada e incompleta, lo que dificulta su aplicación didáctica.
	Extracción y uso de CRP	alta	Todos los CRP son integrados activamente en la propuesta, con función clara y contextualizada.
		media	Solo se usan parcialmente o sin integrar de forma explícita en la actividad.
		baja	Los CRP no se usan o aparecen desconectados del diseño de la actividad.

	Adecuación al objetivo	alta	La actividad responde directamente al objetivo planteado y se ajusta al tipo de ítem.
		media	Relación parcial con el objetivo; puede cumplirlo, pero de forma genérica.
		baja	Desajuste claro entre la actividad y el objetivo o el tipo de ejercicio.
	Claridad expositiva	véanse criterios anteriores	
	Tono didáctico	véanse criterios anteriores	
	Aplicabilidad	alta	Actividad viable, bien estructurada, con instrucciones claras y posible de implementar.
		media	Viable con ajustes; algunas partes requieren aclaración o adaptación.
		baja	Difícilmente aplicable en un aula real; falta de claridad o coherencia operativa.
	Creatividad	alta	Propuesta original, variada, que supera estructuras convencionales.
		media	Muestra ligeros elementos de innovación dentro de un formato común.
		baja	Respuesta estandarizada, sin elementos diferenciadores ni valor añadido.