Futuridad en el español de los *Large Language Models*

Futurity in the Spanish of Large Language Models

Autor1[1], Autor2[2], Autor3[2], Autor4[2,3]
[1] Universidad 1
[2]Universidad 2
[3]Universidad 3
mail contacto

Resumen: Este trabajo analiza el uso de las formas verbales de futuro en el español generado por modelos de lenguaje de gran tamaño (LLM) para evaluar si estos sistemas reproducen una variable lingüística en proceso de cambio: el desplazamiento progresivo del futuro morfológico (*cantaré*) por el futuro perifrástico (*voy a cantar*) en el español contemporáneo. A partir de un experimento controlado en formato tipo test, se examina la preferencia entre el futuro morfológico y el perifrástico en distintos contextos (prospectivos y epistémicos) por parte de diecisiete modelos de lenguaje. Los resultados muestran una tendencia general al uso del futuro morfológico aunque con notables diferencias entre modelos. El análisis revela no solo patrones de comportamiento lingüístico, sino también sesgos derivados de los datos y procesos de entrenamiento. Este estudio constituye una contribución al campo emergente de evaluación lingüística de los LLM en español.

Palabras clave: español, futuros, Large Language Models, evaluación.

Abstract: This study analyzes the use of future verb forms in Spanish generated by large language models (LLMs) in order to evaluate whether these systems reproduce a linguistic variable undergoing change: the progressive displacement of the morphological future (*cantaré*) by the periphrastic future (*voy a cantar*) in contemporary Spanish. Based on a controlled multiple-choice experiment, it examines the preference for the morphological future or the periphrastic future across different contexts (prospective and epistemic) in the output of seventeen contemporary language models. The results show a general tendency toward the morphological future, although with notable differences across models. The analysis reveals not only patterns of linguistic behavior but also biases stemming from training data and processes. This study contributes to the emerging field of linguistic evaluation of LLM in Spanish.

Keywords: Spanish, futures, Large Language Models, evaluation.

1 Introducción

El auge de los modelos de lenguaje ha traído consigo la necesidad de evaluar sus capacidades de manera sistemática, con el fin de impulsar mejoras continuas. Sin embargo, los modelos heredan numerosos sesgos lingüísticos como consecuencia de sus procesos y datos de entrenamiento (Muñoz-Basols *et al.,* 2024) o de la traducción, automática o manual, de sus *datasets* (Plaza *et al*., 2024; Plaza-del-Arco *et al*., 2020; Singh *et al*., 2024). Para minimizar estas limitaciones, se recomienda crear los conjuntos de datos directamente en la lengua seleccionada o adaptarlos manualmente, puesto que, con el tiempo, los textos generados por los modelos terminarán por ser parte del lenguaje producido, transmitido y entendido por humanos (Grieve *et al.,* 2025). Esta afirmación enlaza con la idea de "dialecto digital", una variedad lingüística emergente generada por los modelos, que, si bien puede resultar perfectamente comprensible, no necesariamente representa a ninguna comunidad real de hablantes (Muñoz-Basols *et al.,* 2024).

En efecto, nos encontramos ante un nuevo terreno de estudio en el ámbito del

procesamiento del lenguaje natural y de la sociolingüística, puesto que los LLM son, en sentido amplio, variedades del lenguaje (Grieve *et al.,* 2025). Ante este escenario, se reclama una acción coordinada para desarrollar modelos de lenguaje en español que preserven su complejidad y diversidad (Muñoz-Basols y Hernández Muñoz, 2019). De lo contrario, la IA podría poner en riesgo las tres PPP del español —su carácter policéntrico, polifónico y poliédrico (Moreno Fernández, 2000; Muñoz-Basols y Hernández Muñoz, 2019)— al invisibilizar normas, dialectos y perfiles de hablantes.

Considerando que cada modelo es un productor de textos, una de las tareas que pueden realizarse para evaluar el español de los modelos de lenguaje es la de evaluar variables lingüísticas en proceso de cambio de los hablantes reales para entender si estos cambios siguen las mismas tendencias en el habla artificial y de qué manera. Esto abre una vía de análisis que no solo contribuye al entendimiento del comportamiento lingüístico de los modelos, sino que también permite reflexionar sobre el poder normativo que podrían tener estas tecnologías en la estandarización o marginación de determinadas formas de habla.

Una de las variables interesantes de estudio en los modelos de lenguaje es el uso del futuro en español. En español conviven dos formas principales para expresar futuro: el futuro morfológico (*amaré*) y el perifrástico (*voy a amar*), pero sus usos son complejos y no se limitan a la referencia temporal, ya que también expresan otros muchos valores de carácter epistémico y conjetural (*Llaman a la puerta, será el cartero*). A esta variación funcional se suma la variación geográfica, que muestra diferencias en su uso según el área hispanohablante. En efecto, en muchas variedades del español (fundamentalmente las americanas), y especialmente en el plano oral, es cada vez más común usar el futuro perifrástico (*voy a hacer*) en lugar del futuro morfológico (*haré*), que queda relegado a contextos más formales o escritos (Sedano 2005, RAE-ASALE 2009[2025]: §23.14c). Se observa, en definitiva, que el futuro perifrástico (FP) presenta una mayor extensión en el español de América que en el de España, donde el futuro morfológico (FM) conserva aún cierta vitalidad.

Sin embargo, hay otros factores que pueden contribuir a la elección de una u otra forma, como son los factores intraoracionales. Por ejemplo, se ha descrito que el FP se suele emplear más que el FM cuando el primero va acompañado por complementos adverbiales que indican conexión temporal con el momento de la enunciación (Blas Arroyo 2000). Así, el hablante recurre al FP cuando la acción expresada es cercana o muy cercana al momento de la enunciación (*Hoy voy a ir al cine*), y esta distancia se puede medir atendiendo a la existencia de especificación adverbial y a las características del adverbio o expresión temporal de la cláusula (Sedano 1994). Precisamente, la atención a estos factores permitirá en este trabajo comprobar si los modelos distinguen matices relevantes en la distribución contextual de ambas formas, morfológica y perifrástica, y, por tanto, si representan patrones de uso del español descritos previamente para hablantes humanos.

Todo ello justifica la relevancia del estudio de las formas de futuro en el español generado por grandes modelos de lenguaje, cuyos resultados y limitaciones dependerán intrínsicamente de las características variacionales de sus corpus de entrenamiento. Se plantean, así, tres cuestiones fundamentales:

1. ¿Qué proporción de futuros (morfológicos vs. perifrásticos) utilizan los modelos para expresar futuridad?
2. Al igual que los humanos, ¿son capaces los modelos de relacionar los valores epistémicos con el futuro morfológico?
3. ¿Qué revela esta proporción sobre los datos y el proceso de entrenamiento?

2 Objetivos e hipótesis

El objetivo general de este estudio es el de analizar el uso de las formas verbales de futuro en el español generado por los LLM, con el fin último de entender cómo utilizan esta variable de cambio en el español actual y qué posibles sesgos presentan los datos de entrenamiento de los que se nutren estos modelos. Para ello, se plantean los siguientes objetivos específicos:

O1: Examinar qué formas de futuro emplean los LLM para expresar prospección en español, a partir de su clasificación entre las formas morfológica (*amaré*) y perifrástica (*voy a amar*) para determinar las proporciones de uso de cada una de ellas.

O2: Evaluar si los modelos son capaces, igual que los humanos, de haber adquirido un conocimiento del futuro morfológico como medio de expresión de valores epistémicos y conjeturales.

O3: Identificar sesgos en los datos de entrenamiento de los modelos, inferidos a partir de la variación en el uso de las formas de futuro en español.

A partir de estos objetivos, se formulan las siguientes hipótesis de investigación:

H1: Los modelos de lenguaje en español muestran una preferencia general por el uso del futuro perifrástico frente al futuro morfológico para la expresión de eventos prospectivos. Al ser el perifrástico la tendencia de cambio en español, este será representativo en los datos de entrenamiento y, por tanto, también el más utilizado por los modelos.

H2: Como resultado del entrenamiento, los modelos identifican los contextos conjeturales en los que se requiere el futuro morfológico.

H3: Existen sesgos lingüísticos derivados de los distintos procesos, costes y datos de entrenamiento de los modelos, lo que se refleja en una diferente proporción de futuros arrojada por cada uno de ellos.

3 Metodología

Para la obtención de respuestas por parte de los modelos, se ha creado un *dataset* en forma de test de opción múltiple, que es un método común para evaluar la capacidad de los modelos, ya que puede cubrir diferentes aspectos, desde el conocimiento en diferentes temas hasta el razonamiento de problemas matemáticos (Guo *et al.,* 2023). El *dataset* final[1] es un conjunto de 165 oraciones cuya forma verbal debe completar el modelo, al que se le proporcionan dos opciones: la forma morfológica del futuro y la forma perifrástica.

Así, una primera parte se destina a oraciones en las que se espera como respuesta un futuro prospectivo (100/165), cuya respuesta correcta puede ser una forma verbal morfológica de futuro o una forma perifrástica (1); y, una segunda parte, formada por oraciones en las que se espera un valor conjetural o epistémico (65/165), cuya respuesta correcta solo puede ser una forma verbal morfológica de futuro (2). Las características del test se resumen en la tabla 1.

(1) Estoy en medio de un proceso jurídico interminable. Un proceso que yo _____.
   a. ganaré
   b. voy a ganar

(2) Llaman a la puerta. No sé… ____ el cartero.
   a. será
   b. va a ser

Tipo	Total	Opciones posibles	Opciones correctas
Prospectivo	100	A) futuro morfológico, B) futuro perifrástico	A) futuro morfológico, B) futuro perifrástico
Epistémico	65	A) futuro morfológico B) futuro perifrástico	A) futuro morfológico

Tabla 1: Resumen de las características de la prueba.

En total, se han seleccionado diecisiete modelos para la obtención de resultados, que se reparten de la siguiente manera con el objetivo de asegurar una muestra representativa del panorama actual en el desarrollo de modelos de lenguaje: cuatro modelos de Google (Gemini-2.5-flash, Gemma-3-27b-it, Gemma-3-12b-it y Gemma-3-4b-it), tres modelos de Anthropic (Claude-3.5-Haiku, Claude-Opus-4 y Claude-Sonnet-4), un modelo de Meta (Llama-3.3-70b-instruct), un modelo de DeepSeek (DeepSeek-rl-0528), un modelo de Mistral (Mistral-nemo), tres modelos de Qwen (Qwen-3-32b, Qwen-3-14b y Qwen-3-8b) y cuatro modelos de OpenAI (GPT-4o, GPT-4.1, GPT-4.1-mini y GPT-4.1-nano). Las especificaciones de cada modelo, así como sus costes por token de entrada y de salida, se detallan en la tabla 2:

Modelo	Tamaño	Coste input / output ($/1M tokens)
Gemini-2.5-flash	-	0.15 / 0.60
Gemma-3-27b-it	27B	0.10 / 0.18
Gemma-3-12b-it	12B	0.05 / 0.10
Gemma-3-4b-it	4B	0.02 / 0.04
Claude-3.5-Haiku	~100B	0.80 / 4.00
Claude-Opus-4	-	15.00 / 75.00
Claude-Sonnet-4	-	3.00 / 15.00
Llama-3.3-70b-instruct	70B	0.05 / 0.18
Deepseek-r1-0528	~30-50B	0.50 / 2.15
Mistral-nemo	8-10B	0.01 / 0.012
Qwen3-32B	32B	0.10 / 0.30

[1] Se incluirá el enlace a los datos en la versión final. No se incluye para preservar el anonimato de los autores.

Qwen3-14B	14B	0.06 / 0.24
Qwen3-8B	8B	0.034 / 0.138
GPT4o	-	2.50 / 10.00
GPT-4.1	-	2.00 / 8.00
GPT41mini	-	0.40 / 1.60
GPT41nano	-	0.10 / 0.40

Tabla 2: Relación de modelos, especificaciones y costes por *token*.

Los *prompts* utilizados para cada tipo de futuro son ligeramente distintos, de forma que para los futuros prospectivos se especifica que el evento de la oración se desarrolla en el futuro (3), mientras que para los conjeturales (4) no:

(3) "Teniendo en cuenta que el evento se desarrolla en el futuro, completa la siguiente oración de forma que te resulte natural en español de España. Ajústate solo a las opciones dadas".

(4) "Completa la siguiente oración de la forma que te resulte natural en español de España. Ajústate solo a las opciones dadas".

Se ha seleccionado la variedad del español peninsular por garantizar mayor consistencia experimental, por su cercanía a la variedad lingüística de los investigadores y por ser una de las variedades mejor reconocidas y procesadas por los modelos actuales (Mayor-Rocher et al., 2025). Además, las oraciones fueron extraídas del Corpus de Referencia del Español Actual (CREA), que permite precisamente filtrar los datos por variedad. Dado que estos datos se encuentran en línea, el conjunto de oraciones extraídas se complementó con un *subset* de oraciones paralelas sintácticamente, diseñadas para comprobar que los modelos pudiesen generalizar a datos nuevos, no presentes en un posible entrenamiento previo. Por otro lado, aunque que la mayor variabilidad en los usos del futuro se encuentra en el plano oral, las oraciones recopiladas del corpus y las creadas paralelamente a él pertenecen al plano escrito del lenguaje. Esta elección se justifica porque los modelos de lenguaje, incluso los multimodales, operan fundamentalmente con texto escrito.

Los prompts y las oraciones con sus opciones se organizaron en un archivo Excel, y la interacción con los modelos GPT se realizó mediante la API de OpenAI en modo *batch*. Esta modalidad permite procesar múltiples solicitudes en JSON de forma diferida sin afectar la funcionalidad. Además, se utilizó OpenRouter para acceder vía API a diversos modelos de distintos proveedores, configurando todos con temperatura cero para tener resultados lo más deterministas posibles. El resto de los parámetros se ha mantenido en su configuración predeterminada. El experimento se desarrolló en Python, con el código disponible en TODO[2], y los datos se analizaron en R (v. 4.3.1) usando RStudio, permitiendo procesar, visualizar y evaluar relaciones entre variables.

Evaluar el uso del futuro en los modelos de lenguaje mediante un *dataset* tipo test ofrece varias ventajas: permite un control experimental riguroso al aislar la variable de interés, garantiza reproducibilidad y comparabilidad entre modelos, facilita la identificación de respuestas y es escalable para grandes volúmenes de datos o pruebas de generalización. Entre sus limitaciones destacan la reducción del contexto lingüístico, la focalización en el español peninsular escrito, y la restricción a dos opciones de respuesta, lo que excluye otras formas posibles de expresar futuridad. A pesar de estas limitaciones, el experimento permite analizar cómo los modelos usan los futuros y evaluar su relación con los usos reales y con el Sesgo Lingüístico Digital.

4 Análisis de los resultados

El análisis se organiza en tres bloques principales. En primer lugar, se abordan los futuros prospectivos, para los cuales se han considerado variables como la especificación adverbial y la distancia temporal respecto al evento enunciado. Después, se examinan los futuros enfocados en los usos epistémicos y conjeturales del futuro morfológico. Por último, se analiza la correlación entre el coste computacional de los modelos y las respuestas generadas.

4.1 Futuros con valor prospectivo

La primera parte de la prueba se centra en los futuros con valor prospectivo, es decir, aquellos casos en los que los modelos deben expresar una acción venidera sin matiz conjetural, del tipo *Han dicho los jugadores que finalmente no ____ el partido del sábado.* En estos contextos, se ha evaluado la preferencia de los modelos entre el uso del futuro morfológico (FM) y el futuro

[2] Se incluirá el enlace a los datos en la versión final. No se incluye para preservar el anonimatode los autores.

perifrástico (FP). Como se observa en la Figura 1, algunos de los modelos presentan una clara preferencia por el futuro perifrástico, especialmente Gemma-3.4, GPT-4.1-mini, GPT-4.1-nano y Claude-3.5-Haiku, este último con un uso prácticamente exclusivo del FP en contextos prospectivos. Sin embargo, hay modelos que tienden a un uso más equilibrado entre ambas formas, como DeepSeek o Gemma-3.12b, con alrededor de un 50% de uso del perifrástico. Por el contrario, la mayoría de los modelos (Mistral-nemo, Gemma-3.27b, Qwen-3.13b, Claude-Sonnet-4, Llama 3.3-70b, Gemini-2.5, GPT-4o, GPT-4.1, Claude-Opus-4, Qwen-3.32b y Qwen-3.8b, en orden de frecuencia de FM) reflejan un mayor uso de formas morfológicas que de perifrásticas. Esto no parece reflejar el habla oral espontánea de los hablantes de español, sino más bien el uso del futuro morfológico como recurso formal en el plano escrito.

Resulta especialmente llamativo, por otro lado, que modelos de una misma empresa muestren resultados tan dispares, como ocurre, por ejemplo, con los modelos de OpenAI, donde GPT-4.1-mini y GPT-4.1-nano registran alrededor de un 75% de usos del perifrástico, mientras que GPT-4o y GPT-4.1 solo un ~30%; o con los de Google, Gemma-3.4b y Gemini-2.5, cuyos usos del futuro perifrástico se diferencian en 39 puntos. Este patrón de respuesta que se observa en los modelos refuerza una de las hipótesis planteadas: la distancia entre las respuestas de modelos como Claude-3.5-Haiku o GPT-4.1-nano, por un lado, y modelos como Qwen-3.32b y Qwen-3.8b, por el otro, sugieren la existencia de un notable desequilibrio entre los resultados proporcionados por cada modelo y, por tanto, entre sus procedimientos y datos de entrenamiento entre diferentes empresas, pero también entre modelos de una misma empresa.

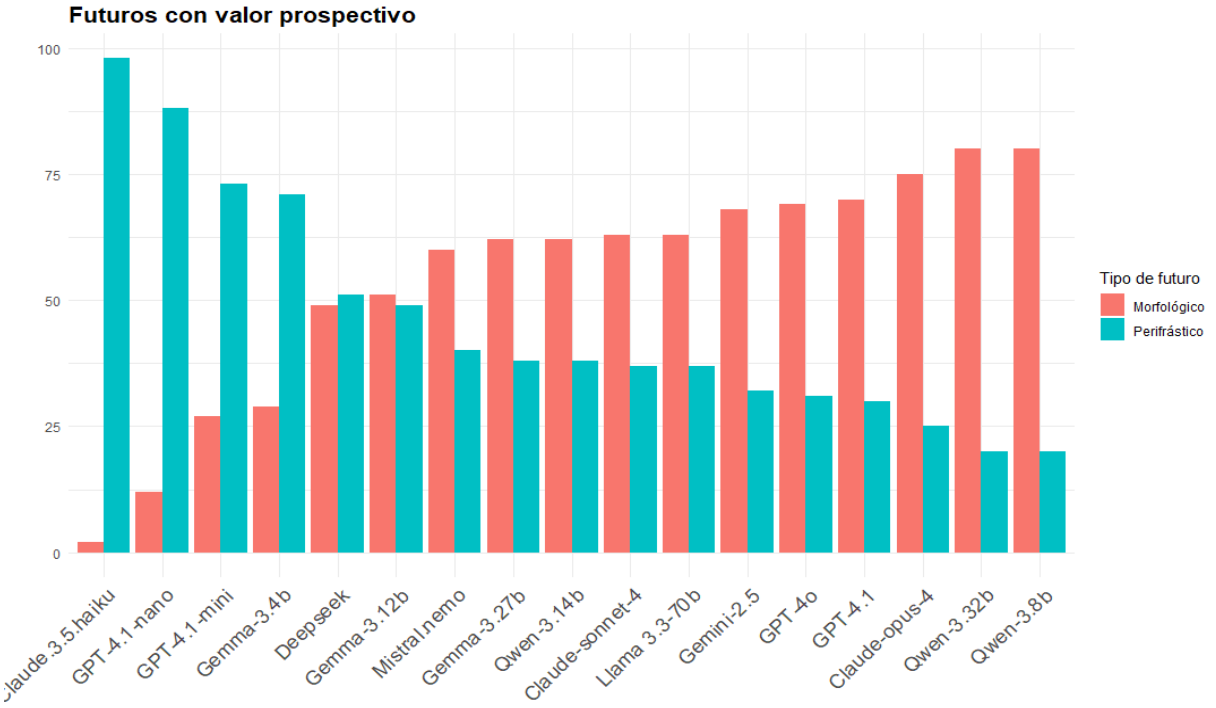

Figura 1: Proporción de selección de futuros morfológico y perifrástico por cada modelo en contextos prospectivos.

Esta proporción de usos resulta interesante como punto de partida para comprender las tendencias generales en la elección de formas, pero no refleja las razones de elección de una u otra forma. A continuación, se analizan dos variables que podrían influir en estas decisiones: la existencia o no de especificación adverbial y la distancia temporal entre la enunciación y el evento enunciado.

4.1.1 Especificación adverbial

Se ha descrito que el futuro perifrástico se suele emplear más que el morfológico cuando el primero va acompañado por complementos, por ejemplo, adverbiales, que indiquen conexión temporal con el momento de la enunciación (Blas Arroyo, 2000). Es decir, podemos estudiar

si cuando se introduce algún tipo de especificación temporal, como en *No te preocupes ahora. Ya pronto lo* ____ (saber), el modelo tiende a elegir el perifrástico, frente a casos donde no se da esta especificación, como en *Trabajando y con la mente clara* ____ (lograr) *esos objetivos y revertir este mal comienzo.* Se clasificaron manualmente 100 ejemplos según la presencia (1) o ausencia (0) de especificación adverbial temporal, obteniendo una distribución equilibrada: 51 con marcador temporal explícito y 49 sin él. Con base en esta clasificación se elaboraron tablas de contingencia por modelo que cruzan el tipo de futuro (morfológico o perifrástico) con la presencia o ausencia de especificación temporal. Debido a que varias celdas presentaban frecuencias esperadas inferiores a cinco, se aplicó la prueba exacta de Fisher, más adecuada que la chi-cuadrado en estas condiciones.

Sin embargo, la figura 2 muestra que no existe una asociación estadísticamente significativa entre la presencia de especificación temporal y la elección de forma verbal en ninguno de los diecisiete modelos analizados ($p > 0.05$). Por tanto, los modelos no parecen ajustar sistemáticamente su elección entre futuro morfológico y perifrástico en función de la presencia de marcadores temporales, lo que indica que este factor no influye de forma consistente en sus decisiones.

Figura 2: Relación entre especificación temporal y tipo de futuro para cada modelo.

4.1.2 Distancia temporal del evento

Otro parámetro analizado es la distancia temporal entre el momento de la enunciación y el evento mencionado. En español, el futuro perifrástico se usa con mayor frecuencia cuando la acción es cercana al momento de habla (Blas Arroyo, 2000), una distancia que puede estimarse a partir del tipo de adverbio o expresión temporal presente en la cláusula (Sedano, 1994). Así, el futuro morfológico tiende a aparecer en contextos de posterioridad imprecisa o lejana, mientras que el perifrástico se emplea en casos de posterioridad inmediata (Sedano, 2006). Partiendo de esa base, las oraciones del primer dataset se clasificaron manualmente según la distancia temporal del evento, utilizando una adaptación de la tipología de cinco categorías de Blas Arroyo (2008) reducida aquí a tres, como se resume en la tabla 3:

Propuesta Blas Arroyo (2008)	Propuesta adaptada para el estudio de la futuridad en LLM.
Cercana o inmediata: límite un día.	Cercana o atenuada: límite de un día, días posteriores al momento del habla y

	distancia psicológicamente cercana.
Intermedia: días posteriores al momento del habla.	
Indefinida: eventos para los que el hablante no tiene datos de cuándo ocurrirán.	Indefinida: eventos para los que el hablante no tiene datos de cuándo ocurrirán.
Atenuada: distancia objetivamente lejana, psicológicamente cercana para el hablante.	
Máxima: evento lejano física y psicológicamente.	Máxima: evento lejano física y psicológicamente.

Tabla 3: Adaptación de las categorías propuestas por Blas Arroyo (2008) para la distancia temporal en el estudio de la futuridad en LLM.

Los datos del primer *dataset* se reclasificaron en tres categorías de distancia temporal: (1) distancia cercana o atenuada (29/100), que incluye eventos física o psicológicamente próximos; (2) distancia indefinida (49/100), para casos en que el hablante no especifica cuándo ocurrirá el evento; y (3) distancia máxima (22/100), correspondiente a eventos lejanos tanto física como psicológicamente. Dado que algunas celdas seguían presentando recuentos bajos, se aplicó una prueba chi-cuadrado con corrección mediante simulación de Montecarlo (10.000 réplicas) para obtener estimaciones más robustas del valor p y evaluar la relación entre la distancia temporal y el tipo de futuro elegido por los modelos. Como puede observarse en la figura 3, la mayoría de los modelos no muestra diferencias estadísticamente significativas ($p > 0.05$), lo cual indica que, en general, la distancia temporal no parece influir de manera sistemática en la elección entre futuro morfológico y perifrástico. No obstante, se observan seis excepciones con valores *p* cercanos al umbral de significación: Claude-Sonnet-4 ($p = 0.004$), DeepSeek ($p = 0.022$), Gemma-3.12b ($p = 0.038$), Gemma-3.4b ($p = 0.037$), GPT-4.1 ($p = 0.009$) y Qwen-3.14b ($p = 0.018$). En particular, estos modelos muestran un menor uso del futuro perifrástico en contextos indefinidos, es decir, en aquellos en los que el evento referido carece de un anclaje temporal claro o el hablante no dispone de información precisa sobre cuándo ocurrirá, de la misma forma que lo identifica Sedano (2006) en hablantes de español. Este patrón sugiere un primer indicio de sensibilidad a las restricciones que rigen la selección entre formas de futuro en español. Aunque puntuales, estos resultados apuntan a que ciertos modelos podrían estar comenzando a representar esta variable de manera incipiente.

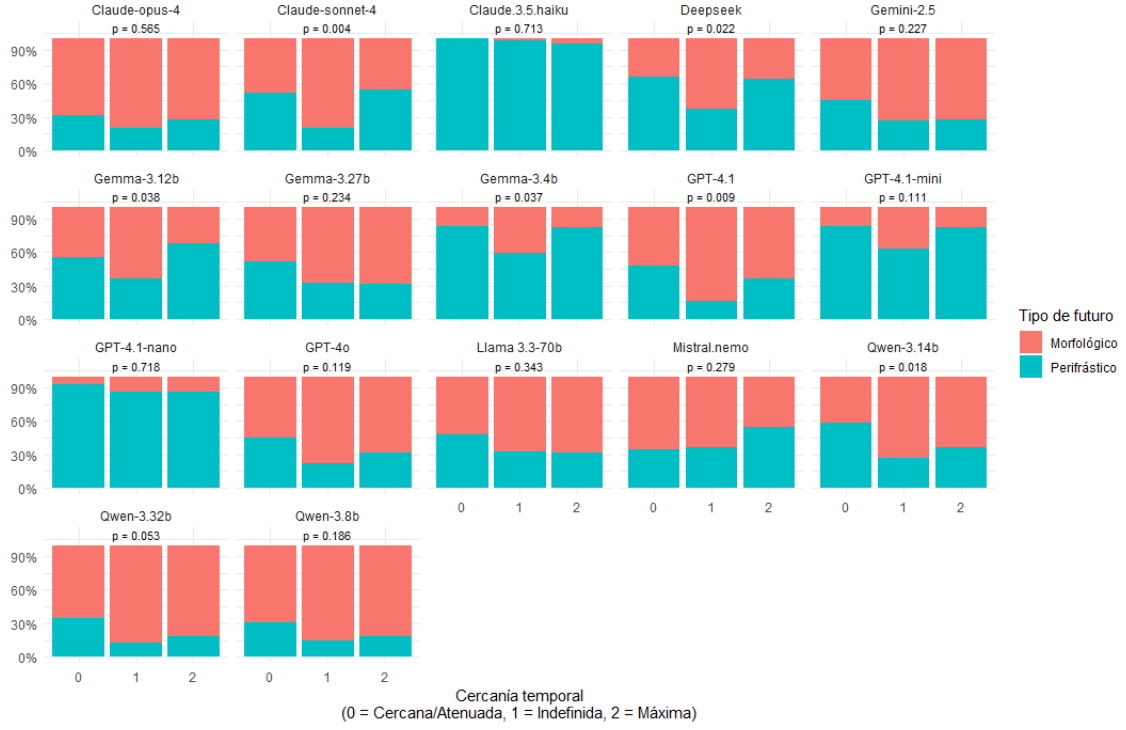

Figura 3: Relación entre distancia temporal y tipo de futuro para cada modelo.

4.2 Futuros con valor epistémico

A diferencia del primer bloque, centrado en usos prospectivos, este segundo conjunto de datos evalúa si los modelos identifican la forma verbal adecuada en contextos con valor epistémico, donde el futuro expresa inferencia o suposición del hablante (p. ej., *Para el tamaño que tiene, digo yo que la tarta ____ unas tres tazas de harina*). En cada oración se ofrecen dos opciones: futuro morfológico y futuro perifrástico, aunque solo el morfológico resulta gramaticalmente apropiado para este valor conjetural. El objetivo no es observar preferencias formales, sino medir la precisión lingüística de los modelos para comprobar si emplean el futuro morfológico cuando el matiz semántico lo exige, evaluando así tanto su competencia gramatical como su sensibilidad a los matices semántico-pragmáticos del español. Como se observa en la figura 4, la mayoría de los modelos (12/17) muestra una tendencia clara a seleccionar la forma correcta, la morfológica, en los contextos epistémicos. Modelos como Claude-Opus-4, DeepSeek, GPT-4.1-mini, GPT-4.1, Claude-Sonnet-4, Llama 3.3-70b, Qwen-3.32b, GPT-4o, Gemini-2.5, Mistral-nemo, Gemma-3.27b o Qwen-3.14b registran más de un 75% de respuestas con futuro morfológico, lo que indica un buen manejo de este matiz semántico y una adecuada correspondencia con los usos normativos del español peninsular. Sin embargo, no todos los modelos alcanzan este nivel de precisión. Claude-3.5-Haiku o Gemma-3.4b destacan por su baja tasa de respuestas correctas, con una clara preferencia por el futuro perifrástico (alrededor del 77%), lo que lo sitúa como el modelo con mayor índice de errores en esta prueba. Resulta igualmente llamativo el comportamiento de modelos como Gemma-3.12b, GPT-4.1-nano o Qwen-3.8b, cuyos resultados se reparten de forma más equilibrada entre FM y FP, con un porcentaje de respuestas incorrectas cercano al 55%. Esta distribución sugiere que estos modelos no logran discriminar con claridad el valor epistémico del futuro morfológico, lo que podría reflejar una interpretación ambigua o inestable de este uso específico.

De nuevo, se aprecian contrastes significativos entre modelos de una misma empresa. El caso más claro se encuentra en los modelos de Anthropic: Claude-Opus-4 es el mejor modelo contestando FM en casi el 100% de las preguntas, mientras que Claude-3.5-Haiku es el peor modelo de toda la muestra. Dentro de OpenAI, GPT-4.1-mini, GPT-4.1 y GPT-4o muestran un desempeño sólido, mientras que GPT-4.1-nano resulta menos consistente.

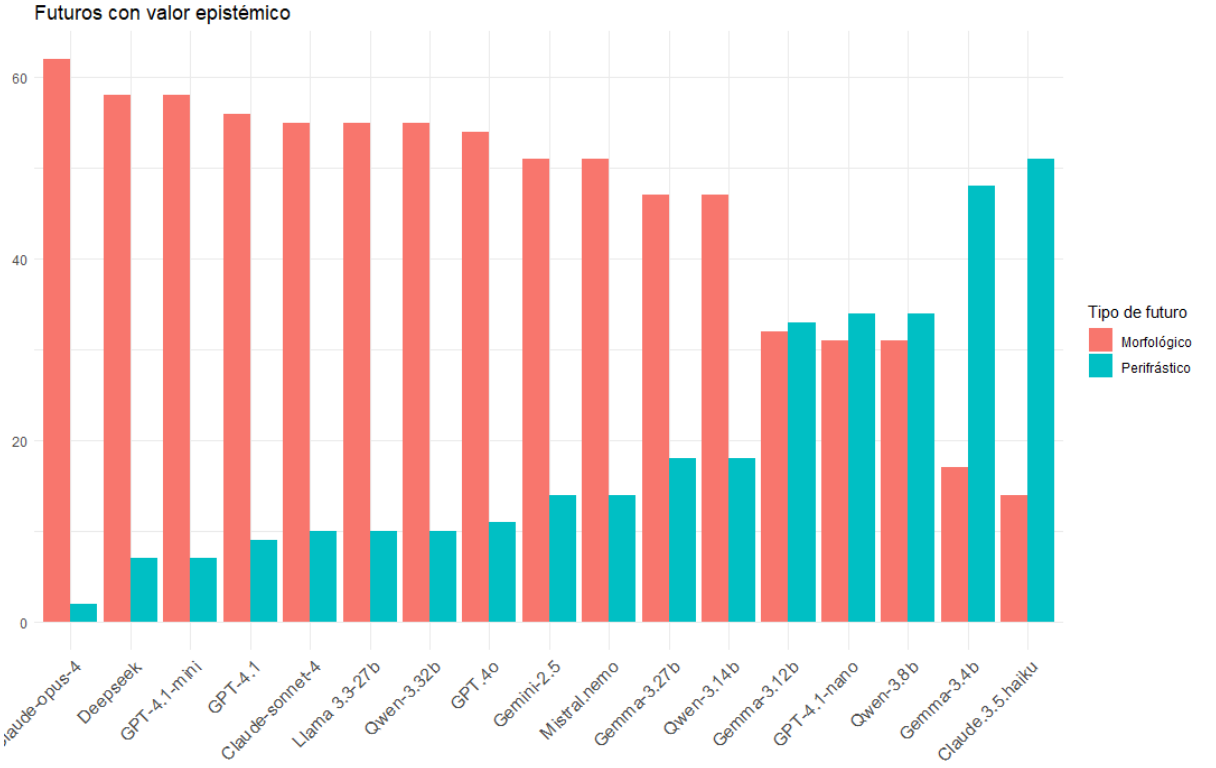

Figura 4: Proporción de selección de futuros morfológicos por cada modelo en contextos epistémicos.

En conjunto, los resultados muestran que la mayoría de los modelos identifican correctamente el futuro morfológico como la forma adecuada en contextos epistémicos. No obstante, algunos presentan dificultades para distinguir entre el uso correcto de esta forma y el uso incorrecto del futuro perifrástico, lo que sugiere limitaciones para captar ciertos matices semánticos y discursivos del español. En estos casos, el futuro morfológico no implica solo una elección formal, sino también una interpretación epistémica que requiere una comprensión más profunda del significado semántico de la oración. Modelos como Claude-3.5-Haiku o Gemma-3.4b, con una mayor proporción de respuestas incorrectas, parecen reflejar una representación insuficiente de este valor en sus datos de entrenamiento, mientras que otros, como Claude-Opus-4, muestran mayor sensibilidad a estos matices. Estas diferencias ponen de relieve el posible impacto de los datos y del entrenamiento en la competencia lingüística de los modelos.

4.3 Correlación entre resultados y costes del modelo

Una de las hipótesis de partida de este estudio planteaba la existencia de sesgos lingüísticos derivados de los procesos y datos de entrenamiento de los modelos, lo que podría reflejarse también en su coste computacional. En este apartado se explora la posible correlación entre el comportamiento lingüístico de los modelos de la muestra y sus costes de uso, medidos en términos del precio por millón de *tokens* de entrada y salida (v. tabla 2). Es decir, se trata de determinar si existe una relación entre el coste por *token* y la tendencia de los modelos a seleccionar un futuro u otro, tanto en contextos prospectivos, donde ambos son posibles, como en contextos epistémicos, donde solo el FM es el adecuado.

En general, se observa que los modelos de mayor coste tienden a seleccionar con mayor frecuencia el futuro morfológico, tanto en contextos epistémicos donde esta forma es la natural, como en aquellos prospectivos donde compite con el perifrástico. Este patrón sugiere que los modelos más costosos han incorporado con mayor precisión los usos formales y normativos del español escrito, reflejados en la preferencia por el FM. Como se observa en la fila superior de la figura 5, las grandes compañías como OpenAI, Google o Anthropic siguen la misma tendencia: cuanto mayor es el coste, mayor es el uso del futuro morfológico en los contextos prospectivos. Este patrón no es tan claro en Qwen, cuyos modelos muestran frecuencias de respuesta similares independientemente del coste del modelo en los contextos prospectivos. Por otro lado, en "Other" se agrupan aquellos modelos únicos por empresa. De izquierda a derecha: Meta, Mistral y DeepSeek.

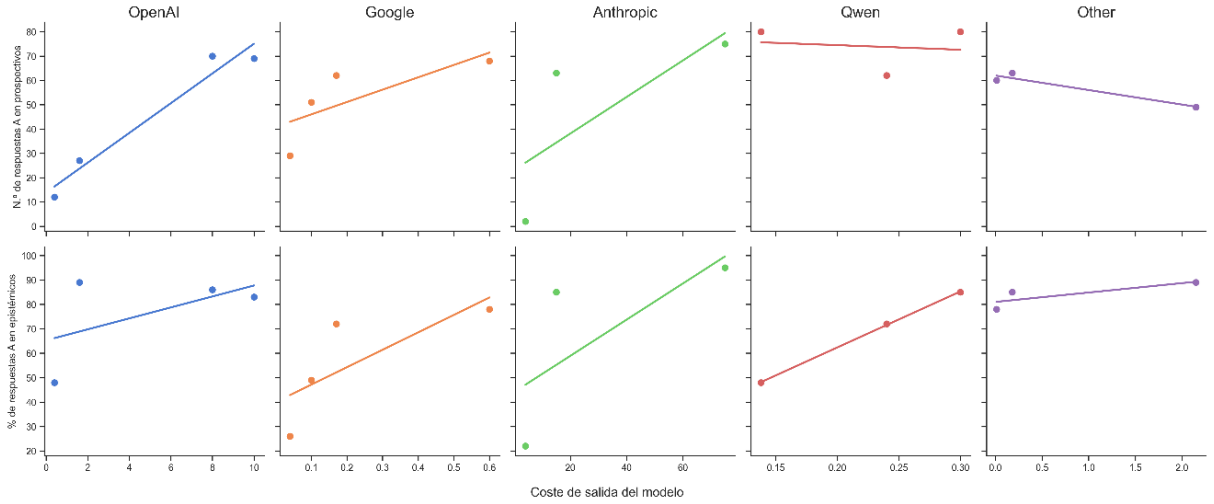

Figura 5: Correlación entre costes y uso del futuro morfológico en contextos prospectivos y epistémicos.

Esta tendencia es más clara en los contextos epistémicos y conjeturales, donde solo el futuro morfológico es la opción gramaticalmente correcta. Así, para todas las compañías, cuanto mayor es el coste del modelo, mayor precisión obtiene este a la hora de identificar el futuro morfológico en los contextos epistémicos y conjeturales, como se observa en la fila inferior de la figura 5.

Estos resultados permiten establecer una tendencia general entre mayor coste y mayor sensibilidad a la forma morfológica, lo que podría entenderse como una aproximación más cercana al español formal, característico del plano escrito. De esta manera, los modelos más costosos, probablemente entrenados con corpus más amplios, cuidados o especializados, captan con mayor precisión las distinciones funcionales del futuro en español. Por tanto, si bien el coste no garantiza siempre una mayor competencia lingüística, en este estudio sí parece correlacionar con una mayor frecuencia de uso del futuro morfológico, tanto en contextos donde es gramaticalmente obligatorio, como los epistémicos, como en contextos prospectivos donde su empleo puede asociarse a registros más cultos o formales. En este sentido, el mayor uso del futuro morfológico por parte de estos modelos sugiere no solo una mayor sensibilidad a la norma escrita, sino también un posible sesgo hacia registros formales del español. Este hecho resulta especialmente relevante en el marco del Sesgo Lingüístico Digital, ya que apunta a una reproducción desigual de las formas lingüísticas según el modelo utilizado, lo que podría contribuir a una estandarización artificial de ciertas variantes. Analizar la relación entre coste —a menudo asociado al tamaño del modelo— y el rendimiento lingüístico se presenta así como una vía clave para comprender qué representación de la lengua están consolidando estas tecnologías.

5 Conclusiones y líneas futuras

A lo largo de este trabajo se ha afrontado el desafío de analizar una variable lingüística compleja como es el uso del futuro en el español generado por los *Large Language Models* (LLM), con el fin de comprender no solo cómo reproducen esta construcción gramatical, sino también qué tipo de datos y patrones lingüísticos reflejan. A partir de los resultados presentados y de las hipótesis formuladas en la introducción —una preferencia general por el futuro perifrástico, el reconocimiento del futuro morfológico en contextos epistémicos, y la presencia de sesgos derivados del entrenamiento—, podemos plantear las siguientes conclusiones en torno al comportamiento lingüístico de los LLM en la expresión de futuridad en español.

En primer lugar, los modelos de lenguaje muestran una preferencia general por el uso del futuro morfológico frente al perifrástico para expresar futuridad en contextos prospectivos, lo que no se ajusta al habla oral de los hablantes de español. No obstante, esta preferencia sí refleja una mayor frecuencia de uso del futuro morfológico en el plano formal y escrito, que es el plano de entrenamiento de los modelos.

En segundo lugar, en los contextos epistémicos, donde el futuro morfológico es la única forma gramatical, la mayoría de los modelos aciertan en la selección de esta forma. Sin embargo, se observan también importantes diferencias entre modelos, lo que evidencia la desigual sensibilidad de los LLM a los matices semánticos del español. Algunos modelos como Claude-3.5-Haiku o Gemma-3.4b presentan dificultades para identificar correctamente los usos epistémicos, lo que indica que estos siguen resultando opacos para algunos modelos.

Por último, existe una correlación entre el coste de los modelos y su sensibilidad al uso de los futuros. Los modelos más costosos, probablemente entrenados con datos más amplios y formales, tienden a seleccionar el futuro morfológico tanto en contextos prospectivos como epistémicos, lo que apunta a un sesgo hacia los registros cultos y formales del plano escrito.

El análisis del uso de las formas de futuro en el español de los modelos de lenguaje constituye una aportación al incipiente campo de evaluación lingüística de estas tecnologías. En un contexto en el que los LLM se integran progresivamente en la producción lingüística cotidiana, resulta cada vez más necesario examinar cómo reproducen las variaciones internas de la lengua española y qué tipo de sesgos emergen de sus datos de entrenamiento. De esta manera, evaluar variables en proceso de cambio en español, como es el caso de la expresión de futuridad, permite no solo detectar patrones de comportamiento gramatical, sino también reflexionar sobre la representación desigual de variedades y registros en estos sistemas. A pesar de las limitaciones metodológicas señaladas, los datos obtenidos resultan relevantes para avanzar en la comprensión del "dialecto digital" que estos modelos generan; con el deseo de que este estudio pueda ampliarse en futuras investigaciones, por ejemplo, en la cobertura de otras variedades del español o en el análisis de nuevas variables lingüísticas que permitan seguir profundizando en los sesgos, tendencias y posibilidades del lenguaje artificial.

Agradecimientos

Pendiente para versión final no anónima.

Bibliografía

Blas Arroyo, J. L. 2000. Aspectos sobre la variación lingüística en la lengua escrita: la expresión de futuridad en el español literario. *Lingüística Española Actual*, 22, 181–200.

Blas Arroyo, J. L. 2008. The variable expression of future tense in Peninsular Spanish: The present (and future) of inflectional forms in the Spanish spoken in a bilingual region. *Language Variation and Change*, 20(1), 85–126.

Grieve, J., Bartl, S., Fuoli, M., Grafmiller, J., Huang, W., Jawerbaum, A., Murakami, A., Perlman, M., Roemling, D., & Winter, B. 2025. The sociolinguistic foundations of language modeling. *Frontiers in Artificial Intelligence*, 7, 1472411. https://doi.org/10.3389/frai.2024.1472411

Mayor-Rocher, M., Pozo Huertas, C., Melero, N., Martínez, G., Grandury, M., & Reviriego, P. 2025. Es igual pero no es lo mismo: ¿Distinguen los LLMs las variedades del español? *Revista de Procesamiento de Lenguaje Natural,* 75, 137-146.

Moreno Fernández, F. (2000). *Qué español enseñar*. Arco/Libros.

Muñoz-Basols, J., & Hernández Muñoz, N. 2019. El español en la era global: agentes y voces de la polifonía panhispánica. *Journal of Spanish Language Teaching*, 6(2), 79–95. https://doi.org/10.1080/23247797.2020.1752019

Muñoz-Basols, J., Palomares Marín, M., & Moreno Fernández, F. 2024. El sesgo lingüístico digital (SLD) en la inteligencia artificial: implicaciones para los modelos de lenguaje masivos en español. *Lengua y Sociedad. Revista de Lingüística Teórica y Aplicada*, 23(2), 623–647.

Plaza, I., Melero, N., del Pozo, C., Conde, J., Reviriego, P., Mayor-Rocher, M., & Grandury, M. (2024). Spanish and LLM benchmarks: Is MMLU lost in translation? *arXiv preprint*, arXiv:2406.17789.

Plaza-del-Arco, F. M., Montejo-Ráez, A., Ureña-López, L. A., & Martín-Valdivia, M. T. 2021. OffendES: A new corpus in Spanish for offensive language research. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 1096–1108.

Real Academia Española. 2009. *Nueva gramática de la lengua española*. Madrid: Espasa.

Real Academia Española. 2025. *Nueva gramática de la lengua española* (edición revisada y ampliada). Madrid: Espasa.

Real Academia Española. (s. f.). *Banco de datos (CREA Anotado): Corpus de referencia del español actual*. https://www.rae.es

Sedano, M. 1994. El futuro morfológico y la expresión *ir a + infinitivo* en el español hablado de Venezuela.

Sedano, M. 2005. Futuro simple y futuro perifrástico en el español hablado y escrito. Trabajo presentado en el XIV Congreso de ALFAL (Asociación de Lingüística y Filología de América Latina), Monterrey, México.

Sedano, M. 2006. Importancia de los datos cuantitativos en el estudio de las expresiones de futuro. *Revista Signos*, 39(61), 283–296.

Singh, S., Romanou, A., Fourrier, C., Adelani, D. I., Ngui, J. G., Vila-Suero, D., Limkonchotiwat, P., Marchisio, K., Leong, W. Q., Susanto, Y., Ng, R., Longpre, S., Ko, W.-Y., Smith, M., Bosselut, A., Oh, A., Martins, A. F. T., Choshen, L., Ippolito, D., Ferrante, E., Fadaee, M., Ermis, B., & Hooker, S. 2024. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint*, arXiv:2412.03304.