Influencia del formato y la complejidad de prompts en calidad de respuestas de modelos generativos

Wilmer Orley Zambrano Vera
Jessica Johanna Morales Carrillo
Escuela Superior Politécnica Agropecuaria de Manabí Manuel Félix López, **ESPAM**
wilmer_zambrano_mdw@espam.edu.ec
jmorales@espam.edu.ec
Calceta, Ecuador

DOI: https://doi.org/10.56124/encriptar.v8i16.011

Resumen

El objetivo de la investigación fue analizar el impacto y la complejidad del los prompts en calidad de respuestas generadas por modelo de lenguaje generativo (MLG), también conocido como Large Language Model (LLM), empleando a DeepSeek-R1 como caso de prueba. Para ello se diseñaron 90 prompts, los mismos que se distribuyeron en tres formatos: declarativo, interrogativo y estructurado, y tres niveles de complejidad: sencillo, moderado y complejo, aplicados a preguntas de cultura general y objetos comunes respectivamente. Las respuestas fueron evaluadas a través de una rúbrica en escala Likert (1–5), por tres expertos en tecnologías del lenguaje, considerando aspectos como: precisión, coherencia y relevancia, complementada con métricas automáticas de: ROUGE, BLEU y BERTScore. Los resultados evidenciaron que los prompts estructurados generaron respuestas mucho más precisas, coherentes y relevantes que los formatos declarativo e interrogativo. En cambio, en la complejidad, los prompts moderados y complejos mostraron mejores resultados que aquellos que eran sencillos en coherencia y contenido semántico, pero no en precisión léxica. Se realizaron pruebas estadísticas de ANOVA y post hoc de Tukey que revelaron contrastes relevantes en gran parte de los criterios. Entre los errores hubo problemas de respuestas con sobreajuste provocado por prompts muy detallados y otros por prompts declarativos, por otro lado, los prompts estructurados mostraron mejores respuestas. Estos resultados ponen de manifiesto la importancia que tienen la optimización de los prompts como una variable critica que determina la calidad obtenida en las respuestas.

Palabras clave: Inteligencia artificial, Tecnologías de la información, Aprendizaje en línea, Cognición.

Influence of prompt format and complexity on the quality of responses from generative models

ABSTRACT

The objective of the research was to analyze the impact and complexity of prompts on the quality of responses generated by a generative language model (GLM), also known as a Large Language Model (LLM), using DeepSeek-R1 as a case study. For this purpose, 90 prompts were designed and distributed across three formats—declarative, interrogative, and structured—and three levels of complexity: simple, moderate, and complex, applied respectively to general knowledge questions and common objects. The responses were evaluated using a Likert scale (1–5) rubric by three experts in language technologies, considering aspects such as accuracy, coherence, and relevance, complemented with automatic metrics including ROUGE, BLEU, and BERTScore. The results showed that structured prompts generated responses that were significantly more accurate, coherent, and relevant than declarative and interrogative formats. In terms of complexity, moderate and complex prompts yielded better results than simple ones in coherence and semantic content, but not in lexical precision. Statistical tests including ANOVA and Tukey's post hoc analysis revealed significant contrasts in many of the criteria. Among the errors, some responses presented overfitting issues caused by overly detailed prompts, while others were linked to declarative prompts; on the other hand, structured prompts produced more consistent answers. These findings highlight the importance of prompt optimization as a critical variable determining the quality of generated responses.

Keywords: Artificial intelligence, Information technology, Online learning, Cognition.

1. Introducción

Los modelos de lenguaje generativo (MLG), o Large Language Models (LLM), son herramientas cada vez más indispensables en diversas áreas, mostrando una gran eficacia en la obtención de resultados. Brown et al. (2020) enfatizaron su importancia en tareas que necesitan generar texto relevante y coherente a través de instrucciones humanas. La eficiencia de estos modelos está ligada a la forma en que se diseñan estas instrucciones, lo que ha

generado un creciente interés por comprender la influencia del prompt en la calidad de las respuestas.

Un aspecto clave en la interacción con modelos de lenguaje es el formato del prompt. Arora et al. (2022) destacaron la importancia de formatos como el declarativo, interrogativo y estructurado para mejorar la coherencia y relevancia de las respuestas. Lee et al. (2024) y Wei et al. (2022) afirman que una correcta estructuración, como el encadenamiento de razonamiento en ingles *Chain-of-Thought*, puede optimizar la generación de datos y mejorar las respuestas, incluso en tareas complejas de inferencia.

La interacción entre el usuario y el MLG, así como la calidad de las respuestas, depende directamente del diseño de los prompts. Según Cheng et al. (2024), los prompts actúan como puente que traduce intenciones cognitivas en instrucciones comprensibles; si están mal estructurados pueden generar ambigüedades en la producción textual. Zhou et al. (2022) sostienen que en tareas de razonamiento multietapa son necesarios factores como claridad, especificidad y secuencialidad. Así, el diseño del prompt se infiere como un mecanismo estratégico de comunicación en el ecosistema humano-Inteligencia Artificial (IA) y no solo como entrada técnica.

Asimismo, el diseño de prompts influye en tareas especializadas como resolución de problemas matemáticos, preguntas de dominio o resúmenes. L. Wang et al. (2023) demostraron que prompts estructurados logran mayor precisión semántica que los genéricos en resúmenes legales. Según B. Wang et al. (2022), una mejor formulación posibilita respuestas adaptadas, con adecuada interpretación. Así, las variaciones en formato y complejidad del prompt pueden alterar significativamente los procesos del MLG, incluso en arquitecturas robustas.

Cuando se evalúan estos MLG en lenguajes diferentes al anglosajón son susceptibles a inconsistencias en su resolución de respuestas. Investigaciones como las de Moraes et al. (2024) afirman que las

ambigüedades morfológicas tienden a presentarse con más frecuencia en el lenguaje español que en el inglés, lo cual influye directamente en su coherencia contextual. Tepe et al. (2024) expresan una crítica por el poco grado de atención que se brinda a modelos multilingües no anglocéntricos como lo es DeepSeek-R1, que, si bien es cierto está entrenado en lenguaje asiático, con una evaluación adecuada puede aportar ventajas significativas en tareas multilingües.

El uso de métricas como ROUGE y BLEU es una limitación frecuente, ya que no bastan para evaluar respuestas complejas. Newman et al. (2020) señalan que estas métricas tienden a favorecer coincidencias superficiales sobre el contenido semántico. Sattele et al. (2023) advierten que pueden dar puntuaciones bajas a respuestas creativas pero válidas. Para una mejor evaluación automática, se propone incorporar BERTScore, que permite una valoración más sensible al significado mediante comparaciones semánticas con representaciones contextuales (Zhang et al., 2020).

Esta investigación aborda ese vacío y, a diferencia de estudios previos con GPT-3 y GPT-4, toma como objeto de análisis al modelo generativo DeepSeek-R1, desarrollado en 2025. Se eligió por tres razones: (1) su arquitectura Transformer multilingüe que incluye español, favoreciendo su desempeño en contextos hispanohablantes; (2) su apertura y accesibilidad técnica, que facilita la reproductibilidad científica y la auditoría; (3) la ausencia de estudios prácticos sobre el efecto de variaciones en prompts sobre la calidad de respuestas. Este trabajo contrasta hallazgos previos en un entorno distinto con un modelo poco documentado, aportando una contribución original al estudio de modelos abiertos.

Con este objetivo, se aplicó una metodología mixta que combinó una evaluación manual, mediante una rúbrica semántica en escala de Likert, con una evaluación automática que usó las métricas de: ROUGE, BLEU y BERTScore. Se diseñó un conjunto balanceado de prompts, aplicados a 20

preguntas de cultura general y 10 objetos comunes, seleccionados por criterios de neutralidad cultural. Esto permitió analizar el desempeño general del MLG e identificar casos de éxito, errores frecuentes y sesgos lingüísticos.

El propósito principal del presente estudio es definir directrices mucho más sólidas para el diseño de prompts en español, encaminados a mejorar la precisión, coherencia y relevancia de las respuestas generadas por MLG. Con esta investigación se espera contribuir a futuras investigaciones sobre prompting multilingüe y sirvan como guía práctica en contextos educativos, comunicativos y tecnológicos hispanohablantes.

2. Metodología

En este estudio se empleó un diseño experimentar mixto el cual combinó un análisis cuantitativo y cualitativo que permitieron evaluar el impacto que tenía el formato y la complejidad del prompt en la calidad de respuestas generadas por el MLG DeepSeek-R1. Se elaboró un conjunto de 90 prompts distribuidos en dos dimensiones independientes las cuales eran, por un lado, él (1) formato prompts y por el otro él (2) nivel de complejidad. En ambos casos, se utilizaron contenidos neutros para minimizar sesgos temáticos, culturales o ideológicos.

2.1 Dimensión 1: Formato del prompt

Esta primera dimensión comprendía 20 preguntas de cultura general, elegidas por su presencia habitual en contextos educativos y por su neutralidad temática, las mismas que fueron desarrolladas en base a contenidos presentes en instructivos y simulacros vinculados al Examen Ser Bachiller del INEVAL (Ministerio de Educación del Ecuador/INEVAL, 2024), así como en evaluaciones estandarizadas aplicadas en la región andina (ICFES, 2020), con la finalidad de garantizar su relevancia y representatividad.

A cada pregunta se le aplicaron tres formatos de prompt: declarativo,

interrogativo y estructurado, lo que dio lugar a un total de 60 prompts, además cada prompt fue evaluado por medio de tres iteraciones, esto quiere decir, que se generaron tres diferentes respuestas para cada uno de los prompts, lo que dio un total de 180 respuestas en esta dimensión. Para poder captar posibles fluctuaciones del MLG estas iteraciones se ejecutaron en diferentes momentos del día.

2.2 Dimensión 2: Complejidad del prompt

Aquí se abordó el nivel de complejidad del prompt partiendo de las descripciones de 10 objetos comunes: silla, reloj, bolígrafo, espejo, mesa, lámpara, zapatos, gafas, llave y camisa, las cuales fueron recompiladas por su alta frecuencia de aparición en tareas de procesamiento de lenguaje natural, particularmente aquellas que se dan en el razonamiento basado en objetos y en la clasificación semántica (Gonen et al., 2023). Para estos objetos se priorizó su familiaridad en diferentes contextos lingüísticos y su baja carga cultural.

A cada objeto se le diseñaron tres niveles de complejidad: sencillo, moderado y complejo, dando un total de 30 prompts, a los cuales solo se le generó una respuesta por prompt, esta única iteración se realizó para evitar redundancias semánticas y controlar el volumen total de datos. En conjunto, el corpus de análisis final estuvo compuesto por 210 respuestas.

2.3 Evaluación de las respuestas

Cada una de las respuestas fueron evaluadas mediante dos criterios complementarios:

- Evaluación automática: Para esta evaluación se emplearon las métricas de: ROUGE que mide la coherencia, BLEU la precisión léxica y BERTScore-F1 su similaridad semántica contextual, las mismas que fueron calculadas con relación a una respuesta de referencia redactada

manualmente para cada prompt.

- Evaluación manual: En esta etapa se aplicó una rúbrica semántica con escala Likert (1–5) para tres criterios: precisión, coherencia y relevancia, empleada por tres evaluadores independiente, expertos en lingüística computacional y tecnologías del lenguaje. Se consideró el cálculo de tres indicadores para validar la calidad de la evaluación: en primer lugar, el coeficiente de alfa de Cronbach que obtuvo $\alpha = 0.85$, lo cual demostró una alta consistencia de la rúbrica, en segundo lugar, la confiabilidad interevaluador, determinada mediante el coeficiente de correlación intraclase que dio ICC = 0.84 y por último el índice de concordancia de Kendall con W = 0.74. Estos valores fueron un indicativo de la alta fiabilidad entre evaluadores y garantizaron la estabilidad en los juicios humanos a lo largo del corpus evaluado.
-

2.4 Control de sesgos y análisis cualitativo

Para controlar la aparición de sesgos temáticos o de dominio se verificó que las preguntas de cultura general no estuvieran relacionadas con los objetos empleados en la dimensión de complejidad. Asimismo, se mantuvo una distribución balanceada de prompts en ambos criterios. Este control fue complementado con análisis cualitativos y estadísticos que permitieron identificar errores frecuentes y patrones atípicos en las respuestas.

En cuanto al control de sesgos culturales y lingüísticos, se observó que las preguntas de cultura general mantuvieron un equilibrio temático al no favorecer contenidos propios de un país específico, lo que evitó sesgos de localización geográfica. No obstante, se identificó que en prompts declarativos formulados en español surgieron ambigüedades morfológicas que afectaron la coherencia de las respuestas, mientras que en inglés el modelo mostró mayor estabilidad léxica.

Estas diferencias se evidencian en los análisis cualitativos (sección 3.4),

donde se reportan respuestas más vagas en español ante instrucciones poco estructuradas. En contraste, el uso de prompts estructurados redujo considerablemente dichas inconsistencias, lo que sugiere que un adecuado diseño de instrucciones puede mitigar los sesgos lingüísticos y culturales en MLG multilingües.

3. Resultados

3.1 Diseño de prompts experimentales

Para este diseño experimental se elaboraron un conjunto de 90 prompts realizados de forma sistemática, los cuales fueron distribuidos de acuerdo a dos dimensiones: formato y nivel de complejidad. El diseño de los prompts se realizó en función de principios de control semántico, variación estructural y claridad. Con lo cual se busca provocar distintos niveles de carga cognitiva que permitan obtener respuestas variadas.

Se inició con la elaboración de los 60 prompts partiendo de 20 preguntas de cultura general los cuales se aplicaron a tres formatos: declarativo, interrogativo y estructurado. Cada formato tiene un nivel diferente de guía, desde preguntas abiertas hasta instrucciones paso a paso. La Tabla 1 muestra un ejemplo aplicado a una pregunta representativa:

Tabla 1. Ejemplo de prompts por formato aplicado a una pregunta de cultura general.

Pregunta de Cultura General	Tipo de Formato	Descripción Técnica del Prompt	Prompts Aplicado
¿Cuál es la capital de Ecuador?	Declarativo	Solicita una descripción o explicación directa sobre un tema.	Describe la capital de Ecuador.
	Interrogativo	Plantea una	¿Qué ciudad es la

		pregunta directa para obtener una respuesta específica.	capital de Ecuador?
	Estructurado	Divide la solicitud en componentes específicos para guiar la respuesta.	"Identifica [Ecuador] | Menciona su capital | Proporciona información adicional relevante."

Fuente: Los autores (2025).

Seguidamente, se elaboraron 30 prompts los mismos que se aplicaron a 10 objetos comunes, con la finalidad de determinar el efecto de los niveles de complejidad de los prompts en la calidad de las respuestas. Para cada uno de estos objetos se emplearon tres criterios de complejidad: sencillo, moderado y complejo. La Tabla 2 muestra un ejemplo con el objeto “silla”:

Tabla 2. Ejemplo de prompts por nivel de complejidad aplicado al objeto “silla”.

Tipo de Formato	Descripción Técnica del Prompt	Prompts Aplicado
Sencillo	Solicita una descripción básica del objeto sin incluir detalles adicionales.	Describe qué es una silla.
Moderado	Requiere una explicación más detallada, incluyendo usos comunes y características generales.	Explica los usos comunes de una silla y los materiales de los que suele estar hecha.
Complejo	Propone un análisis profundo, abarcando aspectos históricos, técnicos o de diseño.	Analiza la evolución del diseño de sillas a lo largo del tiempo, centrándote en cómo la ergonomía y los materiales han influido en las sillas modernas.

Fuente: Los autores (2025).

Gracias a este enfoque experimental, se pudo observar como el formato y el nivel de complejidad afectan directamente la calidad de los resultados generados por el MLG. Fue posible asegurar una apropiada diferenciación de las categorías de los prompts dado que se contaba con expertos en evaluación de lenguaje natural, lo cual contribuyó a la validez interna del estudio.

3.2 Impacto del formato del prompt

Para analizar el efecto del formato del prompt en la calidad de las respuestas, se evaluaron tres formatos (declarativo, interrogativo y estructurado) en función de tres criterios: precisión, coherencia y relevancia. La Tabla 3 resume los promedios obtenidos a partir de las evaluaciones manuales de los tres expertos.

Tabla 3. Promedios de evaluación manual por formato de prompt.

Formato de Prompt	Precisión	Coherencia	Relevancia
Declarativo	4.40	3.72	4.05
Interrogativo	4.52	3.80	4.85
Estructurado	5.00	4.82	4.95

Fuente: Los autores (2025).

Se analizo las diferencias entre formatos para determinar si eran estadísticamente significativas mediante un ANOVA de una vía para cada variable. Los resultados mostraron diferencias significativas entre los tres formatos en los criterios de precisión, coherencia y relevancia (ANOVA, $p < 0.001$). El formato estructurado obtuvo consistentemente los puntajes más altos, seguido del interrogativo, mientras que el declarativo alcanzó los valores más bajos. En precisión y relevancia no se encontraron diferencias entre el formato interrogativo y el estructurado, pero ambos superaron al declarativo.

3.3 Impacto de la complejidad del prompt

De igual manera para medir el impacto en la dificultad del prompt en la precisión de las respuestas generadas del MLG, se analizaron tres niveles: sencillo, moderado y complejo, aplicando métricas automáticas de evaluación: ROUGE (coherencia), BLEU (precisión léxica) y BERTScore-F1 (similaridad semántica). La Tabla 4 resume los valores promedio obtenidos.

Tabla 4. Promedios de métricas automáticas por nivel de complejidad del prompt.

Nivel de Complejidad	ROUGE %	BLEU %	BERTScore-F1 %
Sencillo	32	22	76
Moderado	44	29	78
Complejo	40	28	76

Fuente: Los autores (2025).

En cuanto al nivel de complejidad, el ANOVA evidenció diferencias significativas en coherencia (ROUGE) y similaridad semántica (BERTScore, $p < 0.05$). Los prompts moderados y complejos presentaron un desempeño superior a los sencillos, aunque no se observaron diferencias relevantes entre moderados y complejos. En precisión léxica (BLEU), las diferencias no fueron significativas, lo que indica que la complejidad no influyó de manera notable en este criterio.

3.4 Análisis de errores y respuestas atípicas

Durante la revisión cualitativa de las respuestas generadas por el modelo se pudo apreciar ciertos patrones repetitivos de errores. Para empezar, en los prompts de formato declarativo especialmente las preguntas generales o abiertas, muchas respuestas adolecían de ambigüedad semántica, contenían repeticiones u ofrecían definiciones poco claras, influyendo de forma negativa en las puntuaciones de coherencia y relevancia,

también se observaron respuestas que eludían el foco principal de la pregunta, sugiriendo limitaciones en la capacidad del modelo para inferir intención cuando el prompt carece de estructura.

Además de lo anterior mencionado, también se pudo apreciar que los prompts estructurados con complejidad moderada produjeron respuestas más precisas y coherentes, aunque en algunos casos resultaron demasiado técnicas o detalladas para usuarios no especializados.

En este punto se identificaron 4 respuestas atípicas con rendimientos relevantemente inferiores a los demás, todas generadas por prompts declarativos aplicados a preguntas de cultura general de carácter abstracto o general, en estos casos, el modelo tendió a ofrecer definiciones ambiguas, poco estructuradas o con un tono excesivamente vago que escapaba al propósito educativo del prompt.

Como contrapunto, los prompts estructurados sobre objetos comunes produjeron 6 respuestas sobresalientes (puntuación perfecta en todos los aspectos), distinguidas por su claridad meridiana, pertinencia contextual y ajuste perfecto a la intención original.

En estos casos de polaridad, tanto si eran negativos como si eran positivos el modelo se mostró sensible al diseño del prompt. Con lo cual se puede sugerir que cuando se equilibra una estructura clara y complejidad moderada, el sistema alcanza su máximo potencial generativo.

3.5 Discusión

Los hallazgos obtenidos permitieron evidenciar que tanto los prompts de formato como los de complejidad del prompt, influyen significativamente en la calidad de las respuestas generadas por DeepSeek-R1. En lo que respecta al formato los prompts estructurados mostraron tener respuestas más precisas, coherentes y relevantes, que coincide con los estudios de Wei et al. (2022), quienes expresan que las instrucciones que están segmentadas

facilitan la comprensión del MLG en tareas complejas. Este efecto también fue reportado por White et al. (2023), que señalaron que los prompts con delimitadores explícitos mejoran la alineación semántica y reducen la ambigüedad.

El formato declarativo fue inferior en coherencia y precisión al interrogativo, mientras que el formato estructurado demostró ser significativamente más efectivo que ambos. Esto respalda lo expuesto por Arora et al. (2022) y D. Lee y Palmer (2025) que destacan que, para una mejor interpretación del modelo se deben suministrar preguntas directas en vez de preguntas formuladas de forma poco claras. La diferencia más relevante se obtuvo en el criterio de relevancia ($p < 0.0001$, $\eta 2 = 0.152$), sugiriendo que el tipo de instrucción no solo afecta la exactitud sino también la pertinencia de la información proporcionada.

En cuanto a la complejidad del prompt, el nivel moderado y complejo obtuvieron los mejores resultados en lo que respecta a coherencia y similaridad semántica (ROUGE y BERTScore-F1); sin embargo, no fueron buenos en precisión léxica (BLEU). Estas pautas se correlacionan con estudios realizados por de J. Mu et al., 2023) y Gonen et al. (2023), quienes demostraron que los prompts con mayor complejidad estimulan procesos de inferencia y razonamiento contextualizado. En contraste, BLEU no captó variaciones significativas, tal como advierten Sattele et al. (2023), que expresan que existen obstáculos para medir bien la calidad semántica en procesos generativos.

Estos hallazgos fueron confirmados por el análisis cualitativo, los cuales mostraron que de los prompts declarativos se obtenían respuestas vagas, no obstante, los prompts estructurados daban una mayor claridad y especificidad en las respuestas. Sin embargo, se identificaron casos de sobreajuste semántico en prompts excesivamente detallados, tal como lo advierten Zhao et al. (2021). Finalmente, los hallazgos de Lin (2024) respaldan que los

prompts con instrucciones claras y componentes segmentados activan de manera más eficiente las rutas de inferencia del modelo.

Desde un punto de vista metodológico, los resultados dejan clara la necesidad de un análisis en varias dimensiones de calidad textual al mismo tiempo. Estudios recientes como los de Patel et al. (2023) revelan que al usar estrategias de prompting unidimensionales se suele pasar por alto elementos clave como la coherencia discursiva y la adecuación al contexto. Al realizar la integración de métricas automáticas con evaluaciones humanas, este estudio logró una triangulación robusta, tal como sugieren Lu et al. (2024).

Además, la forma en que DeepSeek-R1 responde a instrucciones detalladas indica que, al igual que otros modelos avanzados como Claude o GPT-4, se ve favorecido por una preparación cognitiva previa. Esto se ve respaldado en investigaciones como la de Brown et al. (2020), que demostraron que prompts con instrucciones claras y bien estructuradas mejoran de forma relevante la precisión semántica y la coherencia en modelos generativos de lenguaje.

En términos cognitivos, los prompts bien diseñados pueden funcionar como apoyos instruccionales, reduciendo de esta manera la ambigüedad y dirigiendo la atención del MLG a aspectos más relevante. Esta hipótesis guarda relación con lo que se expone en investigaciones como las de Mischler et al. (2024), los cuales estudian como las representaciones semánticas en el MLG evolucionan a lo largo de sus capas jerárquicas y se alinean con los patrones de activación del cerebro humano.

En la parte teórica, estos hallazgos contribuyen a una comprensión del prompting como un mecanismo de interacción cognitiva entre los usuarios humanos y el MLG. Esto coincide con lo que dicen autores como Beurer-Kellner et al. (2023), que consideran el prompt como una especie de codificación funcional que dirige el comportamiento del modelo de manera análoga a una programación blanda.

Otro aspecto a tomar en consideración de los resultados obtenidos en este estudio, es que pueden ser aplicados al diseño de sistemas conversacionales que se adapten de manera dinámica a diferentes usuarios. Según lo que afirman autores como Swamy et al. (2023), el incorporar estrategias de prompting dinámico hacen que los MLG rindan mejor en tareas de resolución de ambigüedades y seguimiento de contexto. Esto pone de relieve el valor del diseño instruccional no solo para mejorar la precisión de las respuestas, sino también para fomentar interacciones más sostenidas en sistemas de IA conversacional.

Una limitación del presente estudio radica en que se evaluó exclusivamente al modelo DeepSeek-R1, lo que restringe la generalización de los hallazgos. Sin embargo, la elección de este modelo se fundamentó en su carácter multilingüe y su accesibilidad técnica, que lo convierten en un objeto de análisis relevante en contextos hispanohablantes.

Es importante señalar que investigaciones previas han reportado efectos similares del formato y la complejidad de los prompts en otros modelos de referencia como GPT-3, GPT-4 o Claude (Wei et al., 2022; White et al., 2023; Lin, 2024), lo que sugiere que las tendencias observadas no son exclusivas de DeepSeek-R1. No obstante, futuros trabajos deberían ampliar esta línea de investigación mediante comparaciones experimentales directas entre múltiples LLM, a fin de consolidar y contrastar los patrones identificados en el presente estudio.

Las implicaciones éticas son otro punto relevante a considerar en cuanto a lo que estas prácticas pueden suscitar. Estudios realizados por Feyza et al. (2022) dan alertas sobre los riesgos que pueden generar respuestas sesgadas al modificar el comportamiento del modelo generativo mediante del diseño de prompts, sobre todo cuando no se cuenta con mecanismos de monitoreo y control adecuados. Por todo lo expresado se vuelve indispensable complementar las buenas prácticas de prompting con lineamientos éticos y

transparencia sobre el propósito y el tipo de interacción que se busca lograr.

En su conjunto la investigación realizada respalda la idea de que un adecuado diseño de los prompts influye de una forma decisiva en la calidad de las respuestas generadas. Así como hace una contribución para la definición de prácticas recomendadas para diseñar instrucciones dirigida a MLG, aplicables en entornos educativos, comunicativos y científicos.

4. Conclusiones

Los estudios realizados en el MLG DeepSeek-R1 revelaron que la calidad de las respuestas generadas está influenciada de manera decisiva por el diseño del prompt, especialmente en lo que respecta a su formato y nivel de complejidad. Las evaluaciones mostraron que los prompts estructurados y con un nivel de complejidad moderada favorecen respuestas más precisas, coherentes y relevantes. En contraste, los prompts declarativos, al no ofrecer una orientación clara, tendieron a generar respuestas ambiguas o poco enfocadas, particularmente en preguntas abstractas o generales.

En cuanto a la complejidad, se observó que los prompts intermedios o complejos activan de mejor manera las capacidades de razonamiento semántico del modelo. Sin embargo, los excesivamente detallados pueden producir sobreajuste, lo que resalta la necesidad de mantener un equilibrio entre claridad, accesibilidad y profundidad de la instrucción.

La integración de evaluaciones manuales, métricas automáticas y análisis estadístico permitió validar los resultados con robustez y detectar patrones de error frecuentes. Esto proporciona insumos valiosos tanto para la investigación académica como para el diseño de aplicaciones educativas, informativas y tecnológicas basadas en MLG.

Limitaciones del estudio: los hallazgos deben interpretarse con cautela por ciertas restricciones metodológicas. En primer lugar, se utilizó un único modelo (DeepSeek-R1), lo que limita la generalización a otros LLM. En

segundo lugar, aunque el corpus fue balanceado, su tamaño y carácter experimental reducen la escalabilidad hacia contextos más amplios. Finalmente, se identificaron sesgos lingüísticos asociados al español frente al inglés, que pueden afectar la coherencia y precisión de las respuestas. Estas limitaciones abren líneas de trabajo futuro orientadas a comparar múltiples modelos, distintos idiomas y entornos de aplicación.

5. Referencias

Arora, S., Narayan, A., Chen, M., Orr, L., Guha, N., Bhatia, K., Chami, I., & Ré, C. (2022). Ask Me Anything: A simple strategy for prompting language models. 11th International Conference on Learning Representations, ICLR 2023. https://doi.org/https://doi.org/10.48550/arXiv.2210.02441

Beurer-Kellner, L., Fischer, M., & Vechev, M. (2023). Prompting Is Programming: A Query Language for Large Language Models. Proceedings of the ACM on Programming Languages, 7. https://doi.org/10.1145/3591300;TAXONOMY:TAXONOMY:ACM-PUBTYPE;PAGEGROUP:STRING:PUBLICATION

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., 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, 2020-December. https://doi.org/https://doi.org/10.48550/arXiv.2005.14165

Cheng, K., Ahmed, N. K., Willke, T. L., & Sun, Y. (2024). Structure Guided Prompt: Instructing Large Language Model in Multi-Step Reasoning by Exploring Graph Structure of the Text. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 9407–9430. https://doi.org/10.18653/V1/2024.EMNLP-MAIN.528

Feyza, A., Muhammed, A., Kocyigit, Y., Paik, S., & Wijaya, D. (2022). Challenges in Measuring Bias via Open-Ended Language Generation. Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), Parte de Las Conferencias de La ACL, 76–76. https://doi.org/10.18653/v1/2022.gebnlp-1.9

Gonen, H., Iyer, S., Blevins, T., Smith, N. A., & Zettlemoyer, L. (2023). Demystifying Prompts in Language Models via Perplexity Estimation. Findings of the Association for Computational Linguistics: EMNLP 2023, 10136–10148.

https://doi.org/10.18653/V1/2023.FINDINGS-EMNLP.679

ICFES. (2020). Informe nacional de resultados Saber 11: Educación media en Colombia. Instituto Colombiano para la Evaluación de la Educación. https://www.icfes.gov.co

Lee, D., & Palmer, E. (2025). Prompt engineering in higher education: a systematic review to help inform curricula. International Journal of Educational Technology in Higher Education, 22(1), 1–22. https://doi.org/10.1186/S41239-025-00503-7/TABLES/6

Lee, S. Y. Te, Bahukhandi, A., Liu, D., & Ma, K. L. (2024). Towards Dataset-scale and Feature-oriented Evaluation of Text Summarization in Large Language Model Prompts. IEEE Transactions on Visualization and Computer Graphics. https://doi.org/10.1109/TVCG.2024.3456398

Lin, Z. (2024). Prompt Engineering for Applied Linguistics: Elements, Examples, Techniques, and Strategies. English Language Teaching, 17(9), p14. https://doi.org/10.5539/ELT.V17N9P14

Lu, Q., Qiu, B., Ding, L., Zhang ♢♠, K., Kocmi, T., & Tao, D. (2024). Error Analysis Prompting Enables Human-Like Translation Evaluation in Large Language Models. Findings of the Association for Computational Linguistics ACL 2024, 8801–8816. https://doi.org/10.18653/V1/2024.FINDINGS-ACL.520

Mischler, G., Li, Y. A., Bickel, S., Mehta, A. D., & Mesgarani, N. (2024). Contextual feature extraction hierarchies converge in large language models and the brain. Nature Machine Intelligence 2024 6:12, 6(12), 1467–1477. https://doi.org/10.1038/s42256-024-00925-4

Ministerio de Educación del Ecuador/INEVAL. (2024). Instructivos de pruebas y pruebas modelo para bachillerato. Quito: Ministerio de Educación del Ecuador. https://educacion.gob.ec/instructivos-de-pruebas-y-pruebas-modelo/

Moraes, L. de C., Silvério, I. C., Marques, R. A. S., Anaia, B. de C., de Paula, D. F., de Faria, M. C. S., Cleveston, I., Correia, A. de S., & Freitag, R. M. K. (2024). Análise de ambiguidade linguística em modelos de linguagem de grande escala (LLMs). https://doi.org/https://doi.org/10.48550/arXiv.2404.16653

Mu, J., Li, X. L., & Goodman, N. (2023). Learning to Compress Prompts with Gist Tokens. Advances in Neural Information Processing Systems, 36. https://doi.org/https://doi.org/10.48550/arXiv.2304.08467

Newman, B., Cohn-Gordon, R., & Potts, C. (2020). Communication-based Evaluation for Natural Language Generation (G. J. J. P. Allyson Ettinger, Ed.; pp. 116–126). Association for Computational Linguistics. https://aclanthology.org/2020.scil-1.16/

Patel, D., Kadbhane, S., Sameed, M., Chandorkar, A., & Rumale, A. S. (2023). Prompt Engineering Using Artificial Intelligence. IJARCCE, 12(10). https://doi.org/10.17148/IJARCCE.2023.121018

Sattele, V., Reyes, M., & Fonseca, A. (2023). La Inteligencia Artificial Generativa en el Proceso Creativo y en el Desarrollo de Conceptos de Diseño. UMÁTICA. Revista Sobre Creación y Análisis de La Imagen, 6, 53–73. https://doi.org/10.24310/UMATICA.2023.V5I6.17153

Swamy, S., Tabari, N., Chen, C., & Gangadharaiah, R. (2023). Contextual Dynamic Prompting for Response Generation in Task-oriented Dialog Systems. EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, 3102–3111. https://doi.org/10.18653/V1/2023.EACL-MAIN.226

Tepe, M., Emekli, E., Tepe, M., & Emekli, E. (2024). Assessing the Responses of Large Language Models (ChatGPT-4, Gemini, and Microsoft Copilot) to Frequently Asked Questions in Breast Imaging: A Study on Readability and Accuracy. Cureus, 16(5). https://doi.org/10.7759/CUREUS.59960

Wang, B., Min, S., Deng, X., Shen, J., Wu, Y., Zettlemoyer, L., & Sun, H. (2022). Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 1, 2717–2739. https://doi.org/10.18653/v1/2023.acl-long.153

Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K. W., & Lim, E. P. (2023). Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 1, 2609–2634. https://doi.org/10.18653/v1/2023.acl-long.147

Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, brian, Xia, F., Chi, E. H., Le, Q. V, & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems, 35. https://doi.org/https://doi.org/10.48550/arXiv.2201.11903

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. Proceedings of the 30th Conference on Pattern Languages of Programs. https://arxiv.org/pdf/2302.11382

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. 8th International Conference on Learning

Representations, ICLR 2020. https://doi.org/https://doi.org/10.48550/arXiv.1904.09675

Zhao, T. Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate Before Use: Improving Few-Shot Performance of Language Models. Proceedings of Machine Learning Research, 139, 12697–12706. https://doi.org/https://doi.org/10.48550/arXiv.2102.09690

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., & Chi, E. (2022). Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. 11th International Conference on Learning Representations, ICLR 2023. https://doi.org/https://doi.org/10.48550/arXiv.2205.10625