Cuando el Modelo se lo Imagina Todo: resolución de crímenes con LLMs

Javier Fontes Basabe JAVIERFONTBAS@GMAIL.COM
Grupo C121

Tutor(es):
Lic. Daniel Alejandro Valdés Pérez *Facultad de Matemática y Computación*

Resumen

Los grandes modelos de lenguaje han sido utilizados en tareas de comprensión del lenguaje natural y tareas de preguntas y respuestas. Sin embargo, los mismos pueden presentar un desempeño limitado cuando la calidad del prompt proporcionado es deficiente. El objetivo de este estudio es demostrar que la información ampliada automaticamente conduce a mejores respuestas en este tipo de modelos. Para el estudio se utilizó el dataset Murder Mysteries y para evaluar el impacto del enriquecimiento automático se construyó un nuevo dataset derivado del original con una estrategia de tres etapas: generación de preguntas, generación de respuestas y fusión de contenido. Para la evaluación de la respuesta se utilizó el modelo Llama 3.3 70B. El modelo obtuvo una tasa de aciertos del 62 %con las historias originales y del 74 % con las historias ampliadas, lo que representa una mejora absoluta de 12 puntos porcentuales.

Abstract

Large language models have been used in natural language understanding tasks and question answering. However, their performance can be limited when the quality of the provided prompt is poor. The aim of this study is to demonstrate that automatically enriched information leads to better responses in such models. The Murder Mysteries dataset was used for the study, and to evaluate the impact of automatic enrichment, a new dataset derived from the original was built using a three-stage strategy: question generation, answer generation, and content fusion. For response evaluation, the LLaMA 3.3 70B model was used. The model achieved an accuracy rate of 62 % with the original stories and 74 % with the enriched stories, representing a 12-percentage point absolute improvement.

Palabras Clave: preguntas, respuestas, dataset, razonamiento, contexto
Tema: Grandes Modelos de Lenguaje, Evaluación de razonamiento, Generación automática de contexto

1. Introducción

Los grandes modelos de lenguaje (*LLMs*, por sus siglas en inglés), como LLaMA 3 y DeepSeek, han alcanzado niveles sobresalientes de rendimiento en tareas de comprensión del lenguaje natural y tareas de preguntas y respuestas [3]. Sin embargo, su capacidad para responder preguntas sobre textos narrativos aún se ve influida significativamente por la calidad y extensión del contexto proporcionado. En contextos donde la información no es explícita y las historias presentan ambigüedades, omisiones o una construcción narrativa limitada, estos modelos pueden tener un desempeño limitado significativamente para inferir respuestas correctas; incluso cuando cuentan con arquitecturas avanzadas y altos niveles de rendimiento [4]. El trabajo y todos los recursos utilizados se encuentran disponibles públicamente en el repositorio LLMs-ExpandedContext

Motivación

EL desarrollo y entrenamiento de los *LLMs* puede ser muy costoso. Por esta razón se hace necesario desarrollar métodos eficientes para mejorar la calidad de las respuestas de estos modelos sin tener la necesidad de entrenarlos nuevamente [1]. En casos donde el contexto es deficiente o ambigüo la ampliación de manera automática de los mismos podría ser una solución. Este enfoque disminuye la necesidad de la intervención de especialistas durante el proceso de ampliación del contexto.

Problema tratado

Diversos estudios han demostrado que el uso de contexto adicional puede mejorar la calidad de las respuestas generadas por los *LLMs*. Esto abre la posibilidad de automatizar la mejora del input mediante técnicas de ampliación contextual, como el uso de los propios LLMs para enriquecer el contexto o completar omisiones. La automatización de estas técnicas reduce el costo de entranamiento y la necesidad de supervición humana. Sin embargo, aún no está claro en qué medida estas técnicas realmente mejoran el rendimiento de los modelos y en qué tipos de contextos resultan más eficaces.

Antecedentes

El Grupo de Investigación de Inteligencia Artificial de la Facultad de Matemática y Computación de la Universidad de la Habana ha investigado en los últimos años temas relacionados con el razonamiento y el descubrimiento de conocimiento. Con el surgimiento de los *LLMs* se ha reforzado esta línea de estudio. Entre las investigaciones del grupo encontramos la construcción de un dataset para evaluar la capacidad de razonamiento de estos modelos. Así mismo encontramos otros trabajos relacionados con el uso de *LLMs* y razonamiento para la resolución de diferentes tareas.

Objetivos

El objetivo de este estudio es aportar evidencias a través de un caso de estudio de que la información ampliada automáticamente conduce a mejores respuestas. Además, esto proporcionaría evidencia de que el enriquecimiento contextual puede ser una estrategia clave para optimizar el rendimiento de los *LLMs* sin requerir un nuevo proceso de entrenamiento.

Como objetivos específicos se propone analizar la calidad de las respuestas del modelo Llama 3.3 70B ante el dataset Murder Mysteries, generar un dataset de manera automática con contexto adicional y analizar la mejora del desempeño del modelo ante la misma tarea con el contexto ampliado.

Organización del documento

En primer lugar se presenta un breve resumen del estado del arte de las investigaciones sobre la calidad de la respuesta de los *LLMs* utilizando contextos enriquecidos. Se presenta la propuesta del estudio, el dataset Murder Mysteries, utilizado para el desarrollo de la investigación. Se desarrolla la estructura del mismo y las motivaciones para seleccionarlo. En un segundo momento se explica la estructura del dataset generado, el cual contiene tanto las historias originales como las correspondientes historias ampliadas. Se explica luego el modelo de generación automática de historias ampliadas utilizando la generación de preguntas y respuestas para ampliar el contexto. Se evalúa el desempeño de la propuesta y se realiza un análisis de los resultados obtenidos.

2. Resumen del Estado del Arte

Un Gran Modelo de Lenguaje consiste en una red neuronal con muchos parámetros, entrenado sobre grandes cantidades de texto sin etiquetar mediante aprendizaje autosupervisado. En general los *LLMs* que utilizan inteligencia artificial generativa aprenden el mundo a partir de textos, para luego producir nuevas respuestas similares a las humanas e incluso participar en conversaciones.

Uno de los enfoques más investigados para optimizar el rendimiento de los LLMs en tareas complejas ha sido la incorporación de contexto adicional para mejorar la calidad de las respuestas. Estudios como el de Peters et al. introdujeron representaciones contextualizadas de palabras a través de Embedings from Language Models(ELMo), demostrando que la integración del contexto semántico mejora significativamente el rendimiento en tareas de comprensión de texto [5]. Asimismo, Petroni et al. mostraron que enriquecer los modelos de lenguaje mediante la recuperación de información contextual externa puede igualar o incluso superar el rendimiento de sistemas supervisados en tareas de respuesta a preguntas, sin necesidad de entrenamiento adicional [6].

En 2025, un equipo de investigadores de la Universidad de Tsinghua, en China, presentó Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents, donde introdujeron un nuevo paradigma: el aprendizaje de agentes evolutivos basado en simulaciones (SEAL, por sus siglas en inglés). Este enfoque se diferencia de los modelos de lenguaje tradicionales al construir entornos simulados donde los agentes autónomos pueden evolucionar sin la necesidad de datos etiquetados manualmente. Los resultados demostraron que agentes múltiples entrenados bajo SEAL obtuvieron un mejor desempeño en tareas como Respuesta a Preguntas Médicas (MedQA) [2].

3. Propuesta

Para el desarrollo de la investigación se propone un sistema de ampliación automático de contexto. El uso del mismo sobre un dataset especializado para comparar la capacidad de mejorar los resultados de un *LLMs* ante la tarea de responder preguntas sobre el dataset y el mismo luego de pasar por el sistema propuesto.

Dataset Murder Mysteries

El dataset utilizado en este estudio es Murder Mysteries (Figura 1), una colección estructurada para tareas QA centradas en narrativas de misterio. Este conjunto contiene 250 relatos cortos de asesinatos, cada uno de los cuales incluye: una narrativa principal, una pregunta que apunta a descubrir al culpable, un conjunto de opciones de respuesta, el índice de la respuesta correcta y la respuesta correcta.

narrative string · lengths	question string · classes	choices string · lengths	answer_index int64	answer_choice string · lengths
6.63k–6.86k 26.2%	Who is the… 100%	20–21 21.6%	0 50%	9–10 4.4%
In an adrenaline inducing bungee jumping site, Mack's thrill-seeking adventure…	Who is the most likely murderer?	['Mackenzie', 'Ana']	0	Mackenzie
In an adrenaline inducing bungee jumping site, Mack's thrill-seeking adventure…	Who is the most likely murderer?	['Mackenzie', 'Ana']	1	Ana
In the haze of neon lights and the serving of a silent hand of fate,…	Who is the most likely murderer?	['Harry', 'Rosemary']	0	Harry
In the haze of neon lights and the serving of a silent hand of fate,…	Who is the most likely murderer?	['Harry', 'Rosemary']	1	Rosemary

Figura 1: Dataset: Murder Mysteries

La elección de este dataset se debe principalmente a su carácter narrativo. Cada historia está protagonizada por el detective Winston, quien debe resolver un asesinato entre un grupo limitado de sospechosos. El objetivo es identificar al asesino más probable a partir de pistas presentadas explícita o implícitamente en el texto. Esta estructura hace que el dataset sea ideal pa-

ra estudiar la comprensión profunda del lenguaje y los efectos del contexto narrativo en el rendimiento de los modelos.

Dataset generado automáticamente

Para evaluar el impacto del enriquecimiento automático, se construyó un nuevo dataset derivado a partir del original. Este nuevo conjunto contiene:

1. Las historias originales tal como aparecen en Murder Mysteries
2. Un conjunto de historias ampliadas mediante la estrategia automática propuesta en este estudio: generación de preguntas y respuestas sobre el texto.

Cada historia ampliada conserva la pregunta original, sus opciones de respuesta y el índice correcto, de modo que puedan evaluarse directamente bajo las mismas condiciones.

Modelo de generación de historias ampliadas

Para la ampliación de cada historia se desarrolló un sistema automatizado de generación de preguntas y respuestas utilizando el modelo LLaMA 3.3 70B. Este sistema sigue una estrategia de tres etapas:

1. **Generación de preguntas:** El modelo genera un conjunto de preguntas inferenciales útiles para resolver el caso, pero cuyas respuestas no se encuentran directamente en el texto original
2. **Generación de respuestas narrativas:** Para cada pregunta generada, el mismo modelo produce un párrafo narrativo que contiene su respuesta de forma coherente con la historia.
3. **Fusión de contenidos:** Las respuestas generadas se integran en la narrativa original, dando como resultado una historia ampliada.

Las historias originales tienen una longitud promedio de entre 5000 y 6000 caracteres. Luego de pasar por el sistema y convertirse en historias ampliadas, su extensión asciende entre 10000 y 11000 caracteres, aproximadamente el doble como se muestra en la Figura 2.

Esta expansión no es meramente cuantitativa: aporta una cantidad sustancial de información adicional que complementa y profundiza la narrativa base.

La selección de este enfoque se fundamenta en el modo en que los misterios narrativos son resueltos. Las preguntas generadas no dependen directamente de eventos específicos del texto, sino que buscan explorar aspectos latentes o ausentes. Así, el sistema es capaz de adaptarse a omisiones o ambigüedades, completando espacios vacíos en la narración. Además, este método permite explorar dimensiones ocultas del relato: motivaciones de los personajes, relaciones interpersonales, contradicciones, emociones, intenciones o consecuencias indirectas. En conjunto, la estrategia ofrece una expansión más rica, variada y coherente del texto original, incrementando el potencial del modelo para realizar inferencias más precisas.

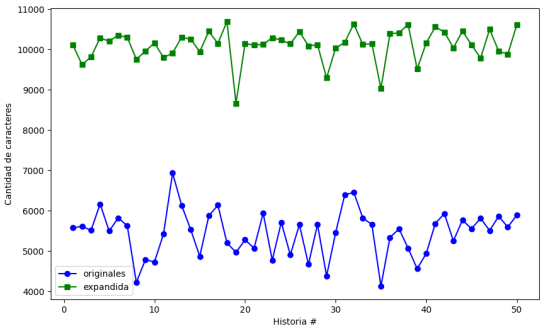

Figura 2: Comparación de la longitud de las historias

3.1 Evaluación de respuestas

Para la evaluación del dataset se utilizó el modelo LLaMA 3.3 70B. En una primera instancia, se le presentaron las historias originales junto con la pregunta y las opciones de respuesta. Posteriormente, se repitió la evaluación utilizando las historias ampliadas, manteniendo las mismas preguntas y opciones. Se evaluaron un total de 50 historias.

4. Resultados

La comparación de los resultados de la evaluación de las historias originales y ampliadas se muestran en el Cuadro 1

Cuadro 1: Comparación de las respuestas

Historia	Aciertos	Porciento
Original	31	62 %
Ampliada	37	74 %

El modelo obtuvo una tasa de aciertos del 62 % con las historias originales y del 74 % con las historias ampliadas, lo que representa una mejora absoluta de 12 puntos porcentuales. En 11 casos, el modelo corrigió su respuesta errónea al usar la versión ampliada. Mientras que en 8 historias, el modelo falló tanto con la narrativa original como con la ampliada como se muestra en el Cuadro 2.

Cuadro 2: Comparación de respuestas correctas

		Ampliada	
		Correcta	**Incorrecta**
Original	**Correcta**	26	5
	Incorrecta	11	8

5. Discusión

La mejora en los aciertos de los casos utilizando historias ampliadas indica que la información adicional proporcionada fue determinante para resolver ambigüedades presentes en la narrativa original.

Los resultados obtenidos en este estudio respaldan la hipótesis de que el enriquecimiento de textos narrativos puede mejorar significativamente el rendimiento de los LLMs en tareas de comprensión lectora. La diferencia en la tasa de aciertos sugiere que el modelo logra realizar inferencias más precisas cuando dispone de un contexto más rico y explícito. Esta observación está alineada con trabajos previos como los de Petroni et al. [6], quienes destacaron el impacto positivo de la recuperación de información adicional para tareas de QA.

Uno de los aspectos más relevantes es que la estrategia utilizada no requiere supervisión humana ni ajustes manuales. A diferencia de enfoques que dependen de anotaciones expertas o reentrenamiento con datos sintéticos, nuestro sistema amplía las historias originales de manera autónoma, permitiendo que cualquier texto pueda beneficiarse del proceso.

También se destaca que los beneficios observados no dependen exclusivamente de un tipo de historia o de pregunta.

Finalmente, cabe destacar el potencial del enfoque para aplicaciones en dominios más allá de la ficción, como la educación, la medicina o el análisis jurídico, donde el razonamiento sobre texto narrativo juega un papel clave. Para aplicar el sistema a estos enfoques se deben realizar ajustes para adaptarlo a contextos reales. La adaptación de esta técnica a otros géneros y formatos puede abrir nuevas vías para mejorar la capacidad explicativa y analítica de los modelos de lenguaje.

No obstante, se identificaron límites claros en el rendimiento del modelo ampliado. En los casos donde la historia original es particularmente pobre en información relevante o está construida de forma incoherente, el proceso de ampliación no logra corregir del todo las deficiencias. Esto sugiere que, si bien la estrategia mejora el contexto, no puede sustituir por completo una base narrativa sólida. Investigaciones futuras podrían explorar mecanismos de evaluación automática de la calidad narrativa como paso previo a la ampliación.

6. Conclusiones

El sistema desarrollado logró una mejora significativa del rendimiento y redujo el esfuerzo al generar contextos ampliados mediante un proceso automatizado. Se cumplió con el objetivo general al aportar evidencia positiva de la utilidad de proporcionar contextos ampliados para la solución de tareas del tipo preguntas y respuestas. Así mismo se logró construir un sistema efectivo para ampliar el contexto de manera automática generando un dataset que ofrece mejores resultados.

7. Recomendaciones

Se recomienda optimizar el proceso de ampliación automática del contexto narrativo e integrar feedback humano en el proceso y evaluar modelos más especializados en tareas narrativas.

8. Futuras líneas de investigación

1. Evaluar la propuesta con otros modelos de lenguaje.
2. Construir otros métodos de generación de contexto automático.
3. Evaluar la calidad del contexto generado de manera automática.
4. Extender la propuesta a otros datasets

Referencias

[1] Ben Cottier, Robi Rahman, Loredana Fattorini, Nestor Maslej, Tamay Besiroglu, and David Owen. The rising costs of training frontier ai models. *arXiv:2405.21015*, 2025.

[2] Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren ans Meng Zhang, and Xinhui Kang te al. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv:2405.02957*, 2025.

[3] Alan B McMillan. Performance of large language models in technical mri question answering: A comparative study. *arXiv:2411.12238*, 2024.

[4] Maya Patel and Aditi Anand. Factuality or fiction? benchmarking modern llms on ambiguous qa with citations. *arXiv:2412.18051*, 2024.

[5] Matthew E. Peters, Mark Neumann, Mohit Iyyer Matt Gardner, Christopher Clark, and Kenton Lee et al. Deep contextualized word representations. *arXiv:1802.05365*, 2018.

[6] Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, and Alexander H. Miller et al. How context affects language models' factual predictions. *arXiv:2005.04611*, 2020.