DOI: 10.69117/TCD.2025.3.11

TermTestQA: evaluación del contenido y de la calidad de la información generada por ChatGPT en el campo de la terminología

Cristina Varga

1. Introducción

El desarrollo y la diversificación de las API basadas en modelos lingüísticos de gran tamaño (LLM) ha abierto un abanico de posibilidades del uso de dichas aplicaciones en varios campos como la traducción, la medicina y las relaciones públicas. Dado el interés suscitado por el uso de la IA en el campo de la traducción y de la traducción especializada, recientemente han empezado a surgir planteamientos sobre su uso en la investigación terminológica (Massion, 2021). Sin embargo, de momento no se conocen datos concretos sobre el alcance y los límites del uso de las API basadas en LLM en el campo de la terminología.

Para responder a la pregunta si los LLM se pueden utilizar de forma efectiva en los aspectos prácticos y teóricos de la investigación terminológica que planteamos en el presente artículo, creemos pertinente emprender un análisis basado en datos empíricos. Consideramos que este enfoque es necesario porque los LLM adquieren cada vez más fluidez en la comunicación lingüística y se les presenta con mayor frecuencia como posibles fuentes de información y conocimiento. Es por esto que los LLM están sometidos a ensayos en distintos ámbitos, en especial en medicina (Goddard, 2023; Siontis *et al.*, 2024), aunque también en informática (Wang *et al.*, 2018), donde los desarrolladores de software quieren optimizar su rendimiento.

Hasta el presente se han publicado distintos modelos basados en cuestionarios que miden el rendimiento de los LLM. Entre ellos mencionamos *SuperGLUE* (Wang *et al.*, 2019) un test que mide la comprensión del lenguaje natural y *MMLU – Massive Multitask Language Understanding* (Hendrycks *et al.*, 2020), que evalúa las LLM en varios campos especializados como las matemáticas, la literatura entre otros. Existen también diversas bases de datos utilizadas para la evaluación del conocimiento del cual disponen los LLM como *Web Questions* (Berant *et al.*, 2013), *TriviaQA* (Joshi *et al.*, 2017), *EfficientQA* (Min *et al.* 2021) y *TruthfulQA* (Lin *et al.*, 2022).

En el campo de la traducción automática se han publicado unos cuantos modelos de evaluación (Chowdhery *et al.*, 2022). Los que más nos han llamado la atención son los modelos que han sido utilizados para la evaluación de la traducción automática realizada por los LLM en varios idiomas entre los cuales el rumano (Bojar *et al.*, 2016; Laskar *et*

al., 2023), ya que la primera fase de la evaluación de los LLM en terminología pretendemos realizarla también en este idioma.

La mayoría de los modelos mencionados tienen como objetivo detectar los puntos débiles de las respuestas generadas por los LLM y analizar las respuestas erróneas de las mismas. Dichas respuestas erróneas se clasifican en varias categorías a partir de meras inadvertencias hasta alucinaciones importantes (Shuster *et al.*, 2021; Zhou *et al.*, 2021; Krishna *et al.*, 2021).

Para evaluar el uso de los LLM en el campo de la terminología opinamos que sería aconsejable seguir el ejemplo de los modelos de evaluación desarrollados previamente en otros campos. Ejemplos que convendría adaptar a las particularidades y a las exigencias del ámbito de la terminología para obtener resultados óptimos. Por lo tanto, en el presente artículo se describirá la elaboración y la configuración de una base de datos construida a partir de preguntas destinadas a evaluar la fiabilidad de los LLM en el ámbito de la terminología en rumano. También se presentarán resultados preliminares del análisis de las respuestas de ChatGPT 3.5 en el ámbito de la terminología y futuras vías de desarrollo.

2. *TermTestQA* – descripción y estructura de la base de datos

En el ámbito de la terminología, la base de datos *TermTestQA* tiene por objeto medir el contenido y la calidad de las respuestas generadas por ChatGPT 3.5, comprobando la conformidad de las mismas con los datos científicos publicados en la literatura. Teniendo en cuenta este hecho, para la elaboración de las preguntas que constituyen la base de datos se han utilizado trabajos de referencia en este campo.

La base de datos *TermTestQA* incluye, en la presente fase de desarrollo, 300 preguntas relativas al ámbito de la terminología, que se ampliarán a un conjunto de 500, un volumen que se acerca más al de los otros modelos de evaluación de los LLM mencionados anteriormente. La elaboración de las preguntas está a cargo de una persona, a saber, la autora del presente artículo, y ninguna de ellas se ha generado automáticamente.

La mayoría de las preguntas consisten en una sola frase. La pregunta la más corta está formada por 14 caracteres mientras que la pregunta más compleja incluye 205 caracteres. En escasas situaciones, con el fin de especificar mejor la pregunta, se han utilizado 2 frases interrogativas como, por ejemplo: *Câte teorii ale terminologiei există în prezent? Care sunt acestea? (¿Cuántas teorías de la terminología existen hoy en día? ¿Cuáles son?)*.

En la elaboración de las preguntas se han empleado varias estructuras como frases asertivas, exclamativas y frases interrogativas. Asimismo, se han utilizado distintos estilos y registros lingüísticos empezando por el registro académico y llegando hasta el registro familiar con el fin de observar si se produce alguna variación expresiva a la hora de generar las respuestas. En general, las preguntas formuladas se han expresado de una forma muy parecida a las preguntas que se encuentran en los exámenes de los estudiantes en terminología.

Desde un punto de vista temático, las preguntas pertenecen a varias categorías. Sin atenerse a una distribución uniforme bajo el punto de vista temático, se pueden observar 8 categorías de preguntas con diferentes grados de representatividad en el conjunto de la base de datos. Dichas categorías constituyen aspectos importantes de la disciplina terminológica, tales como: "historia y evolución", "conceptos específicos", "perspectivas teoréticas y escuelas", "métodos de investigación", "razonamientos y estimaciones", "herramientas informáticas", "autores y figuras de autoridad" y "bibliografía". Para el mejor entendimiento del contenido de la base de datos utilizada para la evaluación de ChatGPT 3.5, a continuación, se describirán con ejemplos las categorías de preguntas mencionadas.

Asimismo, en la primera categoría temática "historia y evolución" que han incluido preguntas sobre los precursores de la terminología moderna, sobre autores de libros fundamentales para este campo de conocimiento, sobre los representantes de las distintas escuelas de terminología. La base de datos incluye también preguntas sobre la periodización del desarrollo de la terminología. La categoría más bien representada se refiere a "conceptos específicos" del ámbito de la terminología y contiene preguntas sobre conceptos como: termino, terminologización, árbol conceptual, ficha terminológica, definición terminológica, glosario, etc. Las preguntas que conciernen las "perspectivas teoréticas y escuelas terminológicas" se refieren a cuestiones sobre las primeras escuelas terminológicas, sobre las contribuciones teóricas fundamentales de las distintas teorías de la terminología, sobre distinciones teóricas entre dichas teorías etc. Las preguntas enfocadas hacia la "metodología de trabajo" incluyen cuestiones sobre como realizar operaciones concretas en el ámbito de la terminología. Por ejemplo *¿cómo estructurar una entrada en un glosario? ¿Cómo se representa un árbol conceptual? ¿Cómo se realiza un análisis conceptual?* etc. Preguntas cuya respuesta supone la descripción de un procedimiento específico.

Una categoría que evalúa ChatGPT 3.5 y los LLM de manera diferente es la que se refiere a "razonamientos y estimaciones" ya que se le pregunta al programa sobre la motivación/finalidad/importancia del uso de cierto concepto en el ámbito terminológico. Asimismo, una pregunta como *¿Con que fin se utilizan los sistemas conceptuales en terminología?* Necesita una respuesta elaborada a partir de un razonamiento y no se trata de una mera búsqueda en un corpus de textos.

Los aspectos técnicos de la investigación terminológica están cubiertos por preguntas sobre varias herramientas informáticas como extractores de terminología, concordancias, listas de frecuencia e infografías. La categoría de preguntas "autores y figuras de autoridad" contiene preguntas cuya respuesta debe mencionar el nombre de un autor o de una personalidad del ámbito de la terminología. Las preguntas se refieren a citas de varios autores, a acontecimientos y fechas importantes y a contribuciones de varios terminólogos. Un ejemplo de pregunta perteneciendo a esta categoría es *¿Cuáles son los más importantes terminólogos rumanos?*

La última categoría de preguntas concierne las "referencias bibliográficas" y se refiere a preguntas que deberían mencionar el título de una publicación, artículo o libro en el campo de la terminología. Como, por ejemplo: *¿Qué documento marca la pauta en el trabajo terminológico?*

Para una mejor comprensión de la distribución de las preguntas dentro de las diferentes categorías temáticas, proponemos la siguiente tabla:

Tabla 1 - *Distribución temática de las preguntas*

Categoría	*Núm. de preguntas*
historia y evolución	13
conceptos específicos	125
perspectivas teoréticas y escuelas	24
métodos de investigación	7
razonamientos y estimaciones	83
herramientas informáticas	11
autores y figuras de autoridad	23
Bibliografía	14

El grado de dificultad y de especialización de las preguntas incluidas en la base de datos equivale al de un terminólogo en formación (estudiante de master o de doctorado) y la bibliografía utilizada para la elaboración de las mismas es de referencia y está disponible en acceso abierto en Internet. Concretamente, se ha recurrido a las siguientes obras: ISO 704 (2009; 2022), Cabré (1998), Pavel y Nolet (2001) y Wüster (2010). Ellas constituyen también el punto de referencia para la evaluación del contenido y de la calidad de las respuestas generadas automáticamente por ChatGPT 3.5.

En una primera fase, las preguntas se han elaborado en rumano y posteriormente se han traducido al inglés. Pretendemos seguir desarrollando la base de datos *TermTestQA*, añadiendo más preguntas hasta lograr un conjunto de datos que contenga un mínimo de 500 preguntas. Asimismo, nos proponemos traducir las preguntas hacia otros idiomas romances (FR, ES, CAT) para poder evaluar el rendimiento de ChatGPT en un mayor número de lenguas ya que los LLM están entrenados de manera diferente en función del idioma.

Todas las 300 preguntas se han sometido a un ensayo con ChatGPT 3.5, lo que nos ha permitido observar que este primer conjunto de preguntas se centra principalmente en aspectos teóricos y que *TermTestQA* se puede desarrollar también hacia la evaluación de los LLM sobre los aspectos prácticos de la terminología.

3. Metodología de evaluación de las respuestas generadas automáticamente

Para evaluar la calidad y el contenido de las respuestas de ChatGPT 3.5 entorno a la terminología el programa ha sido sometido a prueba con las 300 preguntas de las que dispone *TermTestQA* en esta fase de desarrollo. Como resultado se ha obtenido un corpus de respuestas generadas automaticamente por el programa. Dichos resultados se han guardado en un fichero Ms Word de 224 de paginas en formato A4 de extención y 85.783 palabras que han sido analizadas. Dado el hecho de que la base de datos *TermTestQA* se

sigue desarrollando, consideramos los presentes resultados como un primer acercamiento hacia la evaluación de los LLM en el ámbito de la terminología, evaluación que será afinada más adelante a partir de un conjunto más complejo de preguntas.

Para la evaluación de las respuestas se plantea el siguiente protocolo de análisis:

Respuesta ChatGPT		La extensión de la respuesta está adaptada a la complejidad de la cuestión	Si/No
	Aspectos lingüísticos formales	Registro lingüístico	culto/estándar/familiar/jerga
		Problemas de ortografía	Si/No
		Problemas de puntuación	Si/No
		Problemas gramaticales	Si/No
	Contenido discursivo	Respuesta informativa	completa/parcial
		Respuesta redundante	completa/parcial
		Respuesta vacía de contenido	completa/parcial
	Calidad de la información generada	Información correcta	Si/No
		Información incorrecta	Imprecisión
			Omisión
			Alucinación

Las respuestas generadas por ChatGPT 3.5 han sido examinadas y analizadas por un investigador humano, la autora del presente artículo, comprobándose 3 aspectos principales: a) aspectos lingüísticos formales, b) contenido discursivo y c) calidad de la información.

En la categoría "aspectos lingüísticos formales (a)", se han analizado 5 aspectos de las respuestas generadas por ChatGPT 3.5:

1. la adaptación de la respuesta a la complejidad de la pregunta;
2. el registro empleado en la generación de las respuestas (culto, estándar, familiar, jerga);
3. ortografía correcta;
4. puntuación correcta;
5. morfosintaxis correcta.

Con el fin de evaluar el "contenido discursivo de las respuestas (b)", se ha examinado la calidad de las mismas y se han identificado 3 categorías diferentes: *respuesta informativa*, *respuesta redundante* y *respuesta vacía de contenido*. En el marco de esta investigación, entendemos por respuesta informativa cualquier respuesta generada por ChatGPT 3.5 que aporte datos científicos relevantes, completos y veraces a una pregunta. En el mismo marco, se consideran como respuestas redundantes las respuestas que presentan un contenido informativo repetitivo, formulado de distintas maneras en diferentes partes de la respuesta generada automáticamente. Asimismo, asumimos que las respuestas vacías de contenido son aquellas respuestas generadas por ChatGPT 3.5, o partes de las mismas, en las que, aún tratándose de un texto articulado de forma completa y correcta, al usuario no se le transmite ninguna información.

En cuanto a la calidad de la información generada, se distinguen dos categorías de respuestas: 1) respuestas correctas y 2) respuestas incorrectas. Se consideran como *respuestas correctas* aquellas que presentan información completa y adecuada sobre el tópico de la pregunta, teniendo como punto de referencia la bibliografía mencionada anteriormente.

En cuanto a las respuestas incorrectas, en función del tipo de error que presentan, se pueden clasificar en 3 categorías distintas: 1) respuestas imprecisas, 2) omisiones y 3) alucinaciones.

En la presente investigación se entiende por *respuestas imprecisas* a aquellas respuestas que, a pesar de presentar información correcta hasta cierto punto, contienen errores. Otra categoría de errores detectada en el corpus de respuestas generadas por ChatGPT 3.5 la constituyen las *omisiones*. Se trata de respuestas en las que la información se presenta correctamente, pero de forma incompleta. Una categoría distinta la constituyen las *alucinaciones de la IA*, término con el que se designan las situaciones en las que los LLM generan respuestas equivocadas que no se basan en datos/informaciones auténticas. Consideramos las alucinaciones como el más grave tipo de respuestas erróneas, ya que pueden manipular datos erróneos a los usuarios que utilizan ChatGPT como medio de información.

A continuación, se presentarán los resultados obtenidos en esta primera fase de evaluación de ChatGPT 3.5 a partir del análisis cuantitativo y cualitativo de las respuestas generadas por el programa en rumano, en el ámbito de la terminología.

4. Resultados de la evaluación de las respuestas de ChatGPT 3.5

4.1 Análisis cuantitativo

Analizando las respuestas de ChatGPT 3.5 en esta primera fase de desarrollo de *TermTestQA* se ha observado que, desde un punto de vista cuantitativo, 100% de las 300 respuestas generadas por el programa no han sido adaptadas a la complejidad de la pregunta. Incluso si la respuesta correcta tenía que constar solo del nombre de un autor, de la denominación de una teoría o del titulo de un libro, ChatGPT 3.5 ha generado en rumano textos complejos, en general estructurados en tres párrafos. Para obtener una respuesta más reducida, el usuario tiene que pedir explícitamente que dicha respuesta sea expresada en un número exacto de palabras.

En cuanto al registro lingüístico, al formular las respuestas en rumano, en 100% de los casos, sin importar el registro lingüístico utilizado en la pregunta, ChatGPT 3.5 contesta en un registro estándar. Para un cambio de registro, el usuario tiene que mencionar el registro exacto que se tiene que utilizar para una respuesta en concreto. También el cambio de registro afecta generalmente la calidad del contenido de la respuesta generada.

Las respuestas generadas por ChatGPT en rumano presentan problemas de morfosintaxis y de ortografía que, en el estado presente del análisis apreciamos en un 2% del texto generado. Apreciamos que dichos problemas se pueden explicar por el hecho de que el rumano es un idioma con el que los LLM no son muy bien entrenados, hecho mencionado también por otros estudiosos (véase 1. Introducción). No se han observado problemas de puntuación en las respuestas generadas en rumano por ChatGPT 3.5.

En cuanto al contenido de las respuestas, de momento se han obtenido 47,28% de respuestas correctas y 52,72% de respuestas erróneas a las preguntas del *TermTestQA*. Entre las respuestas erróneas, se han registrado 15,45% respuestas imprecisas, 9,09% omisiones y 28,18% alucinaciones de la IA. Lo que nos llama más la atención en este primer análisis no es tanto el porcentaje elevado de respuestas correctas generadas por el programa sino el porcentaje muy elevado de las alucinaciones de la IA. Casi un tercio de las preguntas son formadas completamente o parcialmente por alucinaciones, o sea datos/hechos inventados por el LLM, lo que no recomienda las API basados en dichos modelos lingüísticos como fuente fiable de información especializada.

Tal como se ha mencionado, los presentes datos cuantitativos son provisionales e ilustran el estado presente de desarrollo de la base de datos *TermTestQA.* De esta manera, solo se han analizado las respuestas de ChatGPT 3.5 en rumano. En el futuro, contamos con ampliar la base de datos *TermTestQA* en un 40% para una evaluación más detallada de los LLM. Igualmente, pretendemos trasformar la base de datos en una herramienta de evaluación de los LLM en un entorno multilingüe enfocado hacía las lenguas románicas.

4.2 Análisis cualitativo

Desde un punto de vista cualitativo, los aspectos lingüísticos formales, de las respuestas generadas por ChatGPT 3.5 presentan varios tipos de errores ortográficos, uno de los más importantes siendo el uso de los caracteres especiales en rumano. Se ha observado en las respuestas del asistente conversacional el uso alternativo de palabras con y sin caracteres especiales como en el ejemplo: (1) Termenii din nomenclatură sunt selectați și *definiti* în conformitate cu criterii standardizate.

En otras situaciones, ChatGPT no respeta la normativa del rumano utilizando una forma de sustantivo equivocada "tehnicii" (de genitivo/dativo) en lugar de un adjetivo: (2) Densitatea terminologică se referă la proporția de termeni specializați sau *tehnicii* care apar într-un text sau într-un corpus de texte în raport cu numărul total de cuvinte.

También se ha observado el uso de palabras erróneas en rumano como (3) "laţo-vegetarian" en lugar de "lacto-vegetarian", no resulta claro si se trata de una alucinación de la IA, o sea una palabra inventada por el LLM o si es una unidad léxica que está mal escrita en las fuentes a los que acude ChatGPT para la minería de datos.

Dadas los limites impuestos al presente artículo no citaremos ejemplos de respuestas que no han sido correctamente adaptadas a la complejidad de la respuesta, solo mencionaremos el hecho de que a una pregunta del tipo: (4) *¿Cómo se denomina la herramienta informática que sirve para visualizar todos los contextos en los que aparece un término dado en un texto?* y cuya respuesta más sencilla seria herramienta de concordancias, ChatGPT 3.5 genera en rumano una respuesta de 201 palabras, estructurada en 3 párrafos, explicando que se trata de un "motor de búsqueda".

Desde el punto de vista del contenido discursivo, la mayoría de las respuestas de ChatGPT 3.5 presentan una estructura heterogénea. Parte de las respuestas son informativas, pero también presentan fragmentos redundantes y textos vacíos de contenido.

Como, por ejemplo, (5) *¿En qué época y dónde se originó la terminología?* La respuesta generada presenta en 156 palabras una parte informativa:

> Terminologia ca disciplină distinctă și sistematică a apărut în secolul al XX-lea, odată cu publicarea lucrării lui Eugen Wüster intitulate "Einführung in die allgemeine Terminologielehre". (La terminología como disciplina independiente y sistemática surgió en el siglo XX con la publicación de la obra de Eugen Wüster "Einführung in die allgemeine Terminologielehre")

Una parte redundante:

> Terminologia ca disciplină distinctă și sistematică [...], terminologia ca disciplină academică distinctă, a început să fie dezvoltată abia în secolul al XX-lea, [...]. Astfel, terminologia a apărut ca disciplină distinctă în secolul al XX-lea [...]. (La terminología como disciplina distinta y sistemática [...], la terminología como disciplina académica distinta, empezó a desarrollarse sólo en el siglo XX, [...]. Así pues, la terminología surgió como disciplina distinta en el siglo XX [...])

Y una parte de contenido vacío ya que se puede utilizar casi en cualquier contexto:

> [...] odată cu evoluția științelor lingvistice și a altor domenii de cunoaștere. [...] iar lucrarea sa a reprezentat un punct de cotitură în dezvoltarea acestei discipline, [...]. ([...] con la evolución de las ciencias lingüísticas y otros campos del conocimiento. [...] y su obra supuso un punto de inflexión en el desarrollo de esta disciplina, [...])

En cuanto a la calidad de la información proporcionada por ChatGPT 3.5 una gran parte de las respuestas son acertadas. Como en el ejemplo a la pregunta *¿Las palabras del vocabulario general pueden convertirse en términos? ¿Mediante qué proceso?* Se ha generado una respuesta formada por 124 palabras, completamente correcta de la que presentamos solo una parte:

> (6) Da, cuvintele din vocabularul general pot deveni termeni într-un anumit domeniu de cunoaștere prin procesul de specializare terminologică sau terminologizare. (Sí, las palabras del vocabulario general pueden convertirse en términos de un campo de conocimiento concreto mediante el proceso de especialización terminológica o terminologización)

Sin embargo, la mayoría de las respuestas al cuestionario *TermTestQA* pertenecen a la categoría de respuestas erróneas, desde diferentes puntos de vista. Un tipo de error recurrente son las respuestas con un contenido informativo poco preciso como en el caso de la respuesta a la pregunta *¿Cuál es el primer texto que atestigua la actividad terminológica?*:

> (7) Atestarea activității terminologice este dificil de stabilit cu precizie, deoarece termenii au fost utilizati și definești încă din cele mai vechi timpuri. Însă, unul dintre primele texte care sugerează o preocupare pentru definirea și standardizarea

> termenilor într-un domeniu specific este "De Architectura" ("Despre arhitectură") scris de Marcus Vitruvius Pollio, cunoscut simplu sub numele de Vitruvius, în secolul I î.Hr. (Resulta difícil determinar la naturaleza exacta de la actividad terminológica, ya que los términos se han utilizado y definido desde la antigüedad. Sin embargo, uno de los primeros textos que sugieren una preocupación por definir y normalizar términos en un campo específico es "De Architectura", escrito por Marco Vitruvio Polio, conocido simplemente como Vitruvio, en el siglo I a.C.)

Otra categoría de respuestas erróneas la constituyen las omisiones. Un ejemplo de omisión se puede observar en la respuesta a la pregunta *¿Cuál es la relación fundamental en la que se basa la terminología?* La respuesta menciona solo la relación concepto-termino omitiendo el tercer elemento de la relación, el objeto:

> (8) Relația fundamentală pe care se bazează terminologia este legătura dintre termeni și concepte. (La relación fundamental en la que se basa la terminología es el nexo entre términos y conceptos)

Y, por último, la categoría mejor representada y al mismo tiempo la más problemática desde el punto de vista informativo está constituida por las *alucinaciones de la IA*. Se trata de respuestas cuyo contenido informativo ha sido totalmente inventado por el LLM y que no se basan en ningún hecho real. Para ilustrar esta categoría se ha escogido la respuesta insólita y difícil de elaborar a la pregunta *¿Cuál es el primer glosario terminológico que se conoce? ¿Cómo se titula?*

> (9) Primul glosar terminologic cunoscut este "Glosarul lui Esarhaddon", care datează de la sfârșitul secolului al VII-lea î.Hr. Esarhaddon a fost un rege asirian care a domnit între 681 și 669 î.Hr. (El primer glosario terminológico que se conoce es el "Glosario de Esarhaddon", que data de finales del siglo VII a.C. Esarhaddon fue un rey asirio que gobernó entre los años 681 y 669 a.C.)

Se puede observar que la respuesta contiene información inventada sobre un glosario inexistente que se pone en relación con una figura histórica real de la antigüedad, el rey asirio Esarhaddon. Si se le pide al ChatGPT de mencionar una fuente bibliográfica que mencione dicho glosario, el LLM inventa una fuente bibliográfica:

> (10) Un titlu care menționează "Glosarul lui Esarhaddon" este: "Mesopotamian Glossaries from Esarhaddon's Library: First-Person Voice and Semantic Agency in Early Second Millennium BCE Akkadian Lexicography". *(Una de las fuentes que menciona el "Glosario de Esarhaddon" es: "Glosarios mesopotámicos de la biblioteca de Esarhaddon: voz en primera persona y agencia semántica en la lexicografía acadia de principios del segundo milenio a.C.")*

Dada la facilidad con la cual se presentan fuentes y datos inventados, consideramos que este tipo de contenidos es el más peligroso por el usuario que piensa utilizar ChatGPT 3.5 como fuente de información especializada.

5. Conclusiones

Con esta contribución hemos querido destacar la importancia de la evaluación de los LLM en ámbitos especializados como la terminología en el contexto del desarrollo y del uso generalizado de los LLM. Consideramos que la base de datos *TermTestQA* es un barómetro que le puede indicar al terminólogo y a los profesionales del campo cual es de nivel de fiabilidad de herramientas informáticas como ChatGPT, cuáles son los puntos fuertes y las debilidades de los LLM.

Con estos primeros resultados procedentes de la aplicación del protocolo de análisis, nos motivamos a seguir desarrollando la base de datos para obtener información más detallada sobre la forma y el contenido de las respuestas generadas por ChatGPT. Opinamos que una vez hayamos obtenido resultados cualitativos y cuantitativos detallados en rumano, estaremos en condiciones de ampliar el protocolo de análisis a otros idiomas como el inglés, el francés, el español y el catalán. De este modo obtendremos conjuntos de datos que podrán compararse posteriormente para observar respuestas contrastadas en distintos idiomas.

Una vez finalizado el desarrollo de *TermTestQA* y demostrada su utilidad, la base de datos de preguntas en los cinco idiomas (EN, FR, ES, CAT, RO) se publicará en acceso abierto para su posterior uso por parte de terminólogos, profesores y estudiantes que deseen evaluar el rendimiento de los LLM en el campo de la terminología.

Actualmente, después de haber puesto a prueba la base de datos, podemos afirmar que, en rumano, ChatGPT 3.5 presenta dificultades a distintos niveles, como por ejemplo la macroestructura discursiva de las respuestas generadas, la ortografía y el léxico utilizado. Respecto al contenido informativo de las respuestas proporcionadas, se puede decir que no sólo es informativo, como cabría esperar de un diálogo especializado, sino que también presenta redundancia y fragmentos textuales vacíos de contenido.

Por lo que respecta a la precisión de las respuestas proporcionadas, ChatGPT no alcanza un nivel suficientemente elevado para constituir una fuente de información en el ámbito de la terminología en rumano, ya que más de la mitad de las respuestas analizadas resultaron ser erróneas. Entre los tipos de errores que se producen al interrogar el asistente conversacional en rumano, los más graves son las alucinaciones de la IA.

Referencias bibliográficas

Berant J *et al.*, *Semantic parsing on Freebase from question-answer pairs*, in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013. URL: <https://aclanthology.org/D13-1160.pdf>, último accesso el 28-11-2025.

Bojar O. *et al.*, *Findings of the 2016 conference on machine translation,* in *Proceedings of the First Conference on Machine Translation*, 2016. URL: <https://aclanthology.org/W16-2301/>, último accesso el 28-11-2025.

Cabré M.T., *Terminology: theory, methods and applications*, Philadelphia, John Benjamins, 1998.

Chowdhery A. *et al.*, *PaLM: Scaling language modeling with pathways*, 2021. URL: <https://arxiv.org/abs/2204.02311>, último accesso el 28-11-2025.

Goddard J., *Hallucinations in ChatGPT: a Cautionary Tale for Biomedical Researchers*, in «The American Journal of Medicine», 136/11, 2023. URL: <https://doi.org/10.1016/j.amjmed.2023.06.012>, último accesso el 28-11-2025.

Hendrycks D. *et al.*, *Measuring Massive Multitask Language Understanding*, 2020. URL: <https://arxiv.org/abs/2009.03300>, último accesso el 28-11-2025.

ISO 704: 2009, *International Standard, Terminology work – Principles and methods*. URL: <https://ddialliance.org/sites/default/files/shared/2012/DDI%20Moving%20Forward/ISO/ISO704_2009.pdf>, último accesso el 28-11-2025.

Joshi M. *et al.*, *Triviaqa: a large scale distantly supervised challenge dataset for reading comprehension*, in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017. URL: <https://aclanthology.org/P17-1147/>, último accesso el 28-11-2025.

Krishna K. *et al.*, *Hurdles to progress in long-form question answering,* in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021. URL: <https://aclanthology.org/2021.naacl-main.393/>, último accesso el 28-11-2025.

Laskar M. *et al.*, *A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets*, in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023. URL: <https://aclanthology.org/2023.findings-acl.29/>, último accesso el 28-11-2025.

Lin S. *et al.*, *TruthfulQA: Measuring How Models Mimic Human Falsehoods*, in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022. URL: <https://aclanthology.org/2022.acl-long.229/>, último accesso el 28-11-2025.

Massion F., *Terminology in the Age of Artificial Intelligence*, Berlin, Peter Lang, 2021.

Min S. *et al.*, *NeurIPS 2020 EfficientQA Competition: Systems, Analyses and Lessons Learned*, in «Neural Information Processing Systems», 2021. URL: <https://arxiv.org/abs/2101.00133>, último accesso el 28-11-2025.

Pavel S., Nolet D., *Handbook of Terminology*, Quebec, Minister of Public Works and Government Services Canada, 2001.

Shuster K. *et al.*, *Retrieval augmentation reduces hallucination in conversation*, in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021. URL: <https://aclanthology.org/2021.findings-emnlp.320/>, último accesso el 28-11-2025.

Siontis K.C. *et al.*, *ChatGPT hallucinating: can it get any more humanlike?*, in «European Heart Journal», 45/5, 2024, pp. 321-323.

Wang A. *et al., GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*, in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018. URL: <https://aclanthology.org/W18-5446/>, último accesso el 28-11-2025.

Wang A. *et al., SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems, 2019. URL: <https://arxiv.org/abs/1905.00537>,* último accesso el 28-11-2025.

Wüster E., *Introducción a la teoría general de la terminología y a la lexicografía terminológica*, Barcelona, Publicacions de l'Institut Universitari de Lingüística Aplicada, 2010.

Zhou C. *et al.*, *Detecting hallucinated content in conditional neural sequence generation*, in *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, 2021. URL: <https://arxiv.org/abs/2011.02593>, último accesso el 28-11-2025.