Documento
Análisis

86/2023 13/12/2023

David Ramírez Morán

Large Language Models: los nuevos actores de acceso al conocimiento

Visitar la WEB Recibir BOLETÍN ELECTRÓNICO

Large Language Models: los nuevos actores de acceso al conocimiento

Resumen:

Desde el lanzamiento de ChatGPT, en noviembre del año pasado, la popularización de la inteligencia artificial es un hecho. Su disponibilidad y aplicabilidad al día a día del público general supone una importante diferencia con respecto a la llegada de otras tecnologías que también supusieron un punto de inflexión en la evolución de la humanidad.

Detrás de este nuevo hito tecnológico hay cuestiones técnicas de ámbitos muy diversos que abarcan la economía, la sociedad, la política, la seguridad, la educación, la industria e incluso el clima. Disponer de la tecnología constituye un baluarte de las capacidades científicas, lo que da lugar a la carrera que se está produciendo, que se analiza en este documento.

Palabras clave:

Inteligencia artificial, competencia tecnológica, conocimiento, información.

Large Language Models: the new actors for knowledge access

Abstract:

Since the launch of ChatGPT, in November last year, the popularization of artificial intelligence is a fact. Its availability and applicability in the daily life of general public conveys an important difference in relation with the arrival of other technologies that also conveyed an inflexion point in the evolution of humanity.

Behind this new technological epoch, there are technical questions of very diverse domains that entail the economy, society, politics, security, education, industry and even climate. The availability of this technology is a stronghold of scientific capabilities, what gives place to the race currently in course analysed in this document.

Keywords:

Artificial intelligence, technological competence, knowledge, information.

Cómo citar este documento:

RAMÍREZ MORÁN, David. *Large Language Models: los nuevos actores de acceso al conocimiento.* Documento de Análisis IEEE 86/2023. https://www.ieee.es/Galerias/fichero/docs_analisis/2023/DIEEEA86_2023_DAVRAM_Conocimiento.pdf y/o enlace bie[3] (consultado día/mes/año)

Introducción

Hace poco más de un año desde la introducción de ChatGPT como una herramienta de inteligencia artificial de uso general. Se trata de uno de los productos de los que se conocen como grandes modelos de lenguaje. Estas herramientas de inteligencia artificial utilizan un procesamiento estadístico para analizar grandes volúmenes de información en forma de textos sobre multitud de materias, y son capaces de dar respuesta a las cuestiones planteadas por los usuarios mediante el establecimiento de un diálogo de pregunta y respuesta.

La tecnología, que venía desarrollándose desde hace ya bastantes años, ha pasado a ser un servicio a disposición de toda la población. Proporciona funcionalidades que pueden ser utilizadas para cubrir necesidades que, hasta ahora, o bien no era posible obtener, o se conseguían por otras vías. Entre otras muchas opciones, es posible conseguir respuestas a dudas específicas mediante la utilización de la información recopilada y tratada durante el entrenamiento del sistema, mientras que también es posible que el usuario proporcione información específica para llevar a cabo operaciones como el resumen o extracción de ideas principales de textos, la traducción de contenidos a otros idiomas o modificaciones como la transformación del estilo en el que está redactado un texto para ajustarlo a la forma de escribir de una institución o una persona.

La disponibilidad general de la herramienta ha dado lugar a una creciente identificación de aplicaciones, fruto de la aproximación de usuarios con unos perfiles más diversos que exploran la aplicabilidad de la nueva tecnología a la resolución de sus problemas específicos. La especialización de estos perfiles y su experiencia, desde el curioso accidental hasta el académico más reputado, también está permitiendo detectar los problemas y riesgos que la utilización de estas tecnologías puede suponer.

La enorme potencialidad que se identifica en esta tecnología la convierte también en un vector de interés comercial, mientras que la disponibilidad de la tecnología, su regulación y las consecuencias que su uso puede acarrear son cuestiones sobre las que los Estados deben trabajar, pues son muchos los factores legales, regulatorios, estratégicos y geopolíticos asociados a un servicio disponible de forma global que, hoy en día, están proporcionando empresas privadas.

La IA generativa llega al público general

La velocidad con la que se alcanzaron los cien millones de usuarios de ChatGPT, en menos de dos meses, estableció un récord en la incorporación de usuarios a un nuevo servicio de Internet. El interés despertado por esta tecnología, desde que se hiciese accesible al público general en noviembre de 2022, superó las capacidades de los sistemas desplegados para prestar el servicio y fueron muchos los usuarios que tuvieron que esperar antes de poder obtener la respuesta a una de sus preguntas.

Siete meses después se produjo la primera disminución del tráfico que soportaban los servidores, de alrededor de un 10 %, lo cual puede atribuirse a la finalización de los periodos docentes de la gran mayoría de actividades educativas[1], a una pérdida de interés por parte de los usuarios o a la migración a las versiones de pago y el uso de vías alternativas de acceso, como son el buscador Bing de Microsoft o directamente a través de las API automatizadas que la empresa OpenAI pone a disposición de los desarrolladores[2].

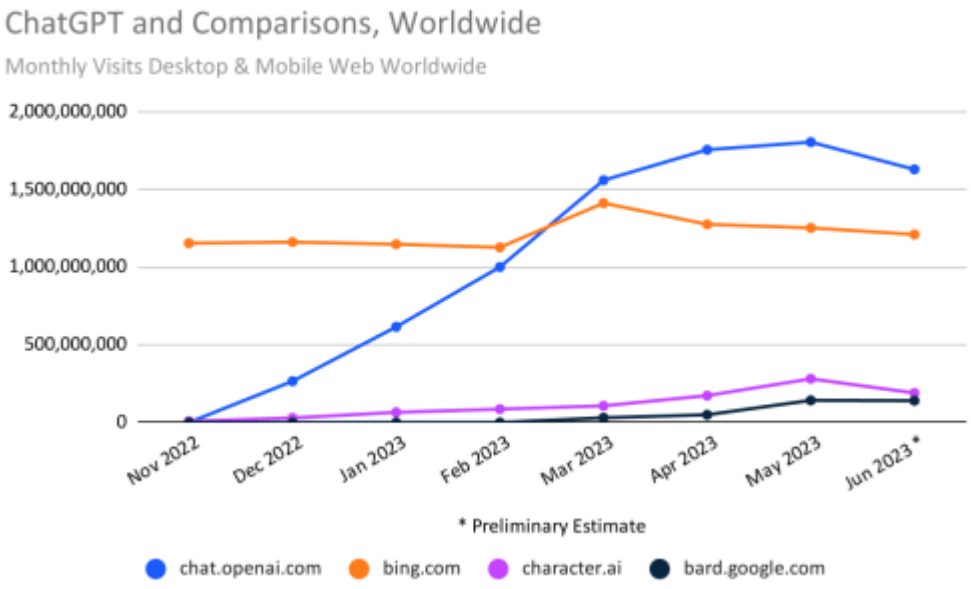

Figura 1. Evolución de las visitas a servicios LLM desde móviles y ordenadores. Fuente: similarweb.com

[1] QUACH, Katyanna. «Now that you've all tried it ... ChatGPT web traffic falls 10 %», *The Register*. 7 de julio 2023. Disponible en: https://www.theregister.com/2023/07/07/traffic_to_chatgpt/
Nota: Todos los enlaces del documento están activos con fecha 12/12/2023.
[2] CARR, David F. «ChatGPT Drops About 10 % in Traffic as the Novelty Wears Off», *Similarweb*. 16 de agosto de 2023. Disponible en: https://www.similarweb.com/blog/insights/ai-news/chatgpt-traffic-drops/

ChatGPT es la punta de lanza de un fenómeno, el de los *Large Language Models*, que ha supuesto un verdadero espaldarazo para la popularización de la inteligencia artificial. El éxito cosechado ha obligado al rápido despliegue de alternativas en una carrera por el posicionamiento ante un mercado potencial con tan elevada demanda latente. Google lanzaba Bard como alternativa a ChatGPT, basado en LaMDa, y Meta cuenta con LLaMa, los cuales constituyen alternativas comerciales esperables entre gigantes de Internet que deben conseguir su mercado a través de la competencia.

Mientras tanto, en China eran varias las empresas que publicitaban el lanzamiento de productos propios[3], Tongyi Qianweb de Alibaba Cloud, Tiangong de Kunlun Tech o Ernie de Baidu, en lo que representa, además de una respuesta comercial para aquellos ámbitos donde no se prima el origen de la tecnología, una respuesta estratégica dentro del enfrentamiento tecnológico que se está produciendo entre Estados Unidos y China. Ningún aspirante a *hegemón* tecnológico puede permitirse carecer de una herramienta que otro aspirante posee.

Lanzar un nuevo producto de características tan complejas como un LLM no está al alcance de todas las empresas, y menos de forma inmediata. Microsoft recurrió a la asociación exclusiva con OpenAI[4], desarrolladora del modelo GPT, y está proporcionando las funcionalidades de la versión GPT-4 a través de su buscador Bing, frente a la versión menos avanzada, GPT-3.5, que es la que dio lugar a la revolución en ChatGPT.

Otra vía para aprovechar la revolución es incorporar las funcionalidades de la herramienta bajo los prismas de adaptar las capacidades del sistema a sectores específicos o proporcionar mecanismos para hacer frente a los problemas e inconvenientes que pueden surgir de la utilización de esta tecnología.

La necesidad de control y la complejidad de su implementación no son cuestión baladí cuando el entrenamiento del sistema se realiza de forma automatizada y su funcionamiento debe responder a limitaciones éticas, políticas o legales de difícil

[3] CHENG, Evelyn. «China's A.I. chatbots haven't yet reached the public like ChatGPT did», *CNBC*. 28 de abril de 2023. Disponible en: https://www.cnbc.com/2023/04/28/how-chinas-chatgpt-ai-alternatives-are-doing.html

[4] «OpenAI forma una exclusiva asociación con Microsoft para construir nuevas tecnologías de super cómputo en Azure AI», *News Microsoft*. 23 de julio de 2019. Disponible en: https://news.microsoft.com/es-xl/openai-forma-una-exclusiva-asociacion-con-microsoft-para-construir-nuevas-tecnologias-de-super-computo-en-azure-ai/

interpretación, cuya transgresión puede conllevar graves consecuencias para el prestador del servicio y, lo que es más importante, para la seguridad de los usuarios.

En este sentido, los despliegues acelerados pueden constituir un serio riesgo para los propietarios ante la aparición de comportamientos indeseados, como le ocurriera a Microsoft con la robot conversacional Tay, basada en inteligencia artificial, que tuvo que ser desactivada un día después de su lanzamiento por los resultados racistas y xenófobos que empezó a generar, fruto de su interacción incontrolada con los usuarios (claramente malintencionados) en 2016[5].

El acceso a la información

Cómo se accede a la información ha ido cambiando desde el origen de la escritura hasta la actualidad. Ciñéndose a los últimos años, es posible identificar cuatro vías que han ido posicionándose como las opciones de referencia.

Actualmente son las redes sociales el principal punto de acceso a la información en los países desarrollados. Proporcionan información inmediata sobre lo que está ocurriendo y están desplazando poco a poco a los medios de comunicación tradicionales, como son prensa, radio y televisión.

Sin embargo, todavía conservan una relevancia importante porque aportan algo que en las redes sociales es más difícil de conseguir: la fiabilidad. También permiten algo que en las redes sociales es mucho más complicado y es la cuestión del archivado de la información mediante hemerotecas que, con los contenidos digitales, son más difíciles de gestionar y poner a disposición de la sociedad.

[5] «Tay, la robot racista y xenófoba de Microsoft», *BBC News Mundo*. 25 de marzo de 2016. Disponible en: https://www.bbc.com/mundo/noticias/2016/03/160325_tecnologia_microsoft_tay_bot_adolescente_inteligencia_artificial_racista_xenofoba_lb

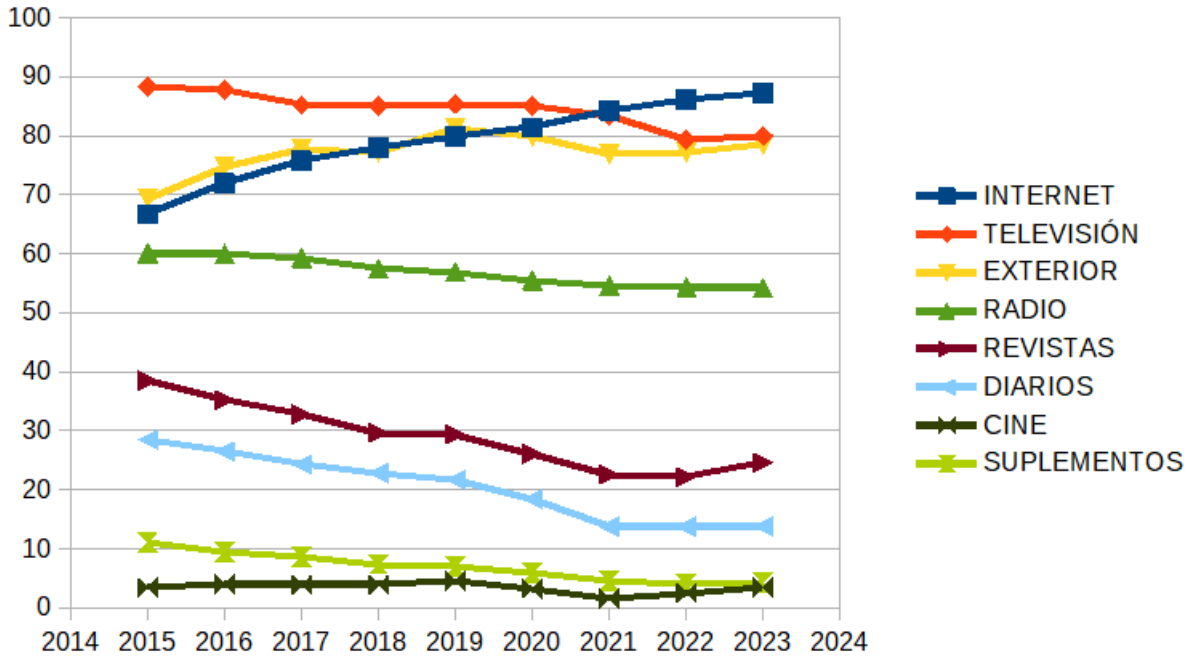

Figura 2. Evolución de la penetración de los medios (Datos de EGM AIMC www.aimc.es)

Indiscutible es el puesto predominante que ocupa todavía Internet cada vez que es necesario acceder a información. Entiéndase este todavía como diferencia entre el Internet tal y como se venía utilizando, con respecto al nuevo modelo de relación mediante los LLM. Si bien no constituye en la mayor parte de los casos la fuente de la información a partir de la que se trabaja, sí que proporciona las referencias a contenidos, publicaciones e informaciones específicas. Si hay algo especialmente destacable de este modelo es la infoxicación, como se denomina el exceso de información que se obtiene al realizar una búsqueda en Internet sobre cualquier tema. Constituye en la práctica un problema porque, hoy en día, la información disponible en Internet accesible de forma abierta es muy numerosa. Su calidad es, *a priori*, desconocida, aunque el origen de esta información (la página web en la que está ubicada) así como el tipo de información de que se trata (piénsese en el contenido de un blog frente a una página web de una institución) permite al usuario evaluar, en cierta medida, la fiabilidad de la información.

En el ámbito académico son las fuentes documentales tradicionales como libros y publicaciones periódicas, físicas u *online*, las fuentes de referencia. Tanto el editor como organismos independientes se encargan de evaluar la calidad de la publicación mediante estrictos procedimientos de verificación, que se ve reafirmada por metainformación como el número de citas y las filiaciones de los autores que publican en esas fuentes.

Este ecosistema ha ido evolucionando para incorporar los nuevos medios de comunicación, y la llegada de los LLM puede suponer un punto de inflexión en su evolución como nunca se había producido. Cada vez hay una mayor tendencia de los usuarios a las fuentes gratuitas, especialmente si el acceso a la información a través de ellas es más sencillo. En esto, las redes sociales destacan especialmente porque el usuario se convierte en consumidor casi pasivo de los contenidos que los algoritmos seleccionan como más relevantes para ese usuario. Se cubre así el modelo de consumo de información general.

Para consultas más concretas, se está produciendo una migración de la utilización de buscadores de contenidos hacia la utilización de LLM. En este caso se produce un cambio importante porque al usuario ya no se le proporcionan datos sobre dónde puede haber información de interés, sino que el algoritmo selecciona información considerada relevante y la elabora de forma que se presente al usuario como un contenido completo que no requiere procesado adicional. El proceso de elaboración filtra los contenidos en una operación que elimina informaciones de dudosa fiabilidad por motivos estadísticos, pero que también elimina los metadatos que permitirían evaluar la fiabilidad de la información obtenida en el cada vez menos frecuente caso de que el usuario decidiera contrastarla.

Cuestiones de base

Los sistemas LLM se vienen desarrollando desde hace ya bastantes años mediante el perfeccionamiento de la técnica de *word embedding*, consistente en determinar la palabra o palabras que deberían aparecer a continuación con base en estadísticas y otros procedimientos algorítmicos. La llegada de los transformadores entrenados, de donde provienen las iniciales P y T de GPT, supuso un punto de inflexión que han aplicado gran parte de las soluciones actualmente disponibles.

El funcionamiento de los LLM es fruto de aplicar algoritmos de aprendizaje de máquina a corpus de conocimiento con los que se entrena la inteligencia artificial a partir de los cuales se obtiene una configuración de los parámetros de funcionamiento del sistema para que, al aplicar a la entrada una petición de información, una pregunta o un enunciado, tratando su contenido, se proporcione una salida en los términos que se le ha solicitado.

Como indicaba el responsable de un experimento en el que se sometía el LLM GPT3 a superar el análisis de obtención de licencia como médico, se trata de una máquina que no genera conocimiento y se limita a proporcionar una salida relacionada con los contenidos utilizados para el entrenamiento del sistema. Este modo de funcionamiento conlleva que el sistema no va a poder generar información nueva de forma consciente, sino que la salida presentará una secuencia de palabras que, idealmente, se corresponde con la respuesta a la solicitud de entrada, limitada por la precisión de la información con la que se ha conseguido entrenar la máquina y de la que no hay forma de certificar la precisión, exhaustividad y exactitud de la respuesta, mientras que tampoco es posible establecer la relación de esta con la información de entrenamiento o las relaciones de parámetros que han dado lugar a la respuesta obtenida.

Surgen así diversos problemas en el funcionamiento del sistema.

El primer problema son las limitaciones del propio sistema en cuanto a complejidad máxima que puede manejar, lo que afectará a parámetros como la exhaustividad y exactitud del sistema. Se está hablando de billones de parámetros que deben ajustarse mediante el procedimiento de entrenamiento.

Otro problema estriba en la interpretación que el sistema realiza del texto de la solicitud de información. Dos preguntas con pequeñas diferencias entre ellas pueden originar resultados con diferente grado de precisión o exhaustividad.

Hoy, y dada la extrema complejidad del sistema, resulta imposible establecer con claridad y precisión los criterios, datos e información que ha utilizado para la elaboración de las respuestas. Se trata de una caja negra que proporciona resultados bajo criterios de naturaleza eminentemente estadística mediante la concatenación de una palabra tras otra hasta componer párrafos y textos enteros.

El corpus de entrenamiento del sistema es un factor vital pues, fruto de los principios de funcionamiento del sistema, toda aquella información o relaciones que no sea posible extraer a partir de la información proporcionada no existe. El sistema no va a poder elaborar teorías o conclusiones que se desvíen del conocimiento contenido en el corpus de entrenamiento. De hecho, en algunos casos se ha detectado el desarrollo de capacidades denominadas especiales, por cuanto parecen generar resultados cuyos fundamentos o relación no se encontraban de forma evidente en la información de

entrada[6]. En estos casos, el sistema ha sido capaz de detectar esas relaciones menos evidentes, lo que ha sido interpretado por los usuarios como el desarrollo de intuición o inteligencia general por parte del sistema.

Por otra parte, están los resultados considerados alucinaciones, por las que las respuestas proporcionadas por el sistema no se corresponden con la realidad. El sistema va a generar una respuesta a la pregunta planteada y seleccionará el contenido estadísticamente a partir de la información que contiene. Si se le pregunta, por ejemplo, qué países son miembros de la OTAN, podría generar una lista con un número arbitrario de países y estos podrían ser aliados o no. Se trata de texto correctamente escrito pero cuyo contenido puede no ser correcto[7]. No hay mecanismo que permita determinar la corrección de la información y queda a discreción del usuario establecer el nivel de confianza que deposita en ella. Este comportamiento se ha llegado a denominar como «loros estocásticos»[8] por cuanto se trata de un sistema que simplemente reproduce información sin criterio ni conocimiento alguno sobre ella.

Además del entrenamiento para la puesta en funcionamiento del sistema, algunas de las soluciones disponibles en este momento también pueden aprender a partir del contenido de las peticiones previas de información de los usuarios. Surge así el dilema de si esta información se puede utilizar para resolver las consultas de otros usuarios o solamente la del usuario en cuestión. La información adicional aportada por el usuario contribuiría al entrenamiento del sistema, pero también podría dar lugar a fugas de información si la información sensible y privada apareciera directamente en las respuestas a otros usuarios.

Los principios de funcionamiento de los LLM están siendo criticados incluso por reputados expertos en inteligencia artificial, por cuanto consideran que no es posible conseguir la AGI Artificial General Intelligence, la inteligencia artificial general con capacidad de pensar y razonar como un humano a partir de un LLM. Su arquitectura no

[6] CLABURN, Thomas. «Large language models' surprise emergent behavior written off as 'a mirage'», *The Register*. 16 de mayo de 2023. Disponible en: https://www.theregister.com/2023/05/16/large_language_models_behavior/
[7] SMITH, Craig S. «Hallucinations could blunt ChatGPT's Success», *IEEE Spectrum*. 13 de marzo de 2023. Disponible en: https://spectrum.ieee.org/ai-hallucination
[8] BENDER, Emily M. y GEBRU, Timnit. «On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?», *ACM Digital Library*. 3 de marzo de 2021. Disponible en: https://dl.acm.org/doi/pdf/10.1145/3442188.3445922

se presta a la comprensión y elaboración de conceptos complejos, sino que genera información de forma estadística[9].

Acceso y soberanía de la tecnología

Las barreras de acceso a ciertas tecnologías son elevadas, fruto de un trabajo de desarrollo de productos donde la componente privada ha sido muy importante. De esta forma se ha producido una falta de visibilidad sobre la tecnología en desarrollo que ha permitido que tecnologías que van a resultar cada vez más imprescindibles para actores estatales y privados estén hoy fuera del alcance incluso de Estados y organizaciones internacionales, cuando se prima la soberanía y autonomía estratégica eligiendo proveedores locales.

Estas barreras, dada la facilidad e interés existente, lo cual no exime de una compensación económica al proveedor, en que se implanten las soluciones actualmente disponibles, se ven todavía más potenciadas no en término de altura, pues el coste de desarrollar las alternativas en dimensiones económicas y temporales pueden aproximarse, sino en anchura o resistencia, porque la disponibilidad de soluciones dificulta motivar la inversión en etapas madurativas de la tecnología muy por detrás del estado del arte, y también pueden surgir problemas en relación con la propiedad intelectual, dado el riesgo que supone desarrollar capacidades equivalentes cuando en el mercado existe ya una solución implantada.

A este último respecto, hay que destacar que las soluciones actuales responden en la mayor parte de los casos a un modelo comercial privado respaldado por una empresa privada con intereses económicos difíciles de acercar a las necesidades de Estados cuyas necesidades de contratación individuales, si no una gota en el océano, sí se pueden comparar al aporte de un río a la masa líquida del planeta. El poder concentrado en estos proveedores de servicio es muy grande y las consecuencias se han dejado ya sentir con escenarios similares cuando, tras la implantación de la normativa de protección de datos en la Unión Europea, una de las principales redes sociales amenazó con interrumpir la prestación del servicio en territorio europeo. Recientemente se ha dado a conocer la sanción que se ha impuesto a esta misma compañía por tratar de eludir la

[9] CLARK, Lindsay. «Artificial General Intelligence remains a distant dream despite LLM boom», *The Register.* 4 de julio de 2023. Disponible en: https://www.theregister.com/2023/07/04/agi_llm_distant_dream/

normativa de protección de datos basándose en una interpretación interesada de excepciones que se utilizaron para sustentar la continuación de su actividad.

La relevancia que se atribuye a los LLM lleva a que los Estados y organizaciones supragubernamentales deban abordar la gestión del desarrollo, el acceso y la dependencia de estas tecnologías. Como ha venido ocurriendo en las últimas décadas en las cuestiones más relacionadas con la tecnología, se ha producido un fenómeno de empuje de esta para que fuera aplicada a problemas existentes.

Incluso hay Estados que o bien han restringido su uso, o bien son objeto de veto del uso de las tecnologías abiertas al público general. Los motivos por los que varios Estados han restringido o prohibido el empleo de LLM disponibles *online* son diversos. Entre ellos figuran países como Rusia, China, Corea del Norte, Cuba, Irán, Siria o Italia[10].

Del grupo, resulta muy relevante la presencia de Italia, que vetó el acceso a la plataforma desde su territorio debido a cuestiones relativas a la protección de datos. Este veto puso en vilo a un gran número de usuarios europeos ante la posibilidad de que los argumentos esgrimidos en Italia se utilizaran en otros países de la Unión Europea para vetar el acceso.

Ucrania también estuvo en la lista debido a la imposibilidad de determinar la utilización por parte de ciudadanos del territorio de Crimea, bajo control ruso.

A título de ejemplo, OpenAI, la empresa creadora de ChatGPT, mantiene una lista de países que pueden utilizar la API para acceder a sus servicios y, por motivos de bloqueo, prohibición u omisión, se han excluido los siguientes países:

- Afganistán
- Bután
- República Centroafricana
- Chad
- Eritrea
- Suazilandia
- Irán
- Libia
- Sudán del Sur

[10] www.digitaltrends.com/computing/thiese-countries-chatgpt-banned/

- Sudán
- Siria
- Yemen

Fugas de información

En la sociedad de la información, los datos son los principales activos de las personas y las organizaciones. La confidencialidad se convierte así en una cuestión de seguridad básica como queda refrendado por la normativa sobre protección de datos de la Unión Europea.

Son bastantes las empresas que han limitado o vetado el uso de las herramientas generativas debido a las dudas que se plantean en cuanto a la preservación de la confidencialidad de la información[11]. Dentro de la lista se encuentran grandes empresas tecnológicas, que *a priori* deberían presentar una menor desconfianza sobre el mundo digital, así como otras empresas en las que las cautelas resultan razonables por tratar cuestiones de seguridad nacional, como Northrop Grumman, que trabaja en proyectos de defensa.

Cuando se solicita al sistema la elaboración de un resumen o una transformación de un contenido, ya sea la reescritura con un estilo diferente o su traducción a otro idioma, requiere proporcionarle al sistema la información que se desea transformar.

Los sistemas pueden utilizar la información proporcionada por los usuarios para afinar su funcionamiento. En este caso, la información proporcionada por un usuario puede utilizarse de manera indirecta para generar las respuestas a las preguntas de otro usuario. Se puede producir una fuga de conocimiento por la que el LLM utilice la información de un usuario para dar la respuesta a otro usuario.

Un incorrecto diseño del sistema también puede dar lugar a la fuga de información mediante un ciberataque que explote unas medidas de seguridad insuficientes que permitan acceder a los datos individuales, confidenciales por contrato, por parte de terceras personas. Durante unos días, la información de los usuarios de ChatGPT estuvo

[11] MOK, Aaron. «Amazon, Apple, and 12 other major companies that have restricted employees from using ChatGPT», *Business insider*. 11 de julio de 2023. www.businessinsider.com/chatgpt-companies-issued-bans-restrictions-openai-ai-amazon-apple-2023-7

expuesta debido a la vulnerabilidad de una de las librerías usadas para su despliegue[12]. La vulnerabilidad permitía que los usuarios pudieran acceder a la información de los chats de otros usuarios arbitrarios. Toda la información introducida en el chat con el que se interactúa con el sistema podía aparecer en el chat de otro usuario. Si se había introducido un texto para resumirlo o traducirlo, el texto podría aparecer tal cual en el chat de otro usuario.

La labor de *prompt engineer*, es decir, del experto encargado de generar la pregunta en los términos adecuados para obtener exactamente la respuesta que se desea obtener, puede ser también una fuente de fuga de información. Para elaborar la pregunta se está haciendo uso del conocimiento del experto sobre el área en cuestión. Si bien el funcionamiento del sistema no traslada esta información al procesado de información de otros usuarios, el prestador del servicio recibe la pregunta ajustada en términos precisos y puede extraer información relevante en relación con el tema requerido. Imagínese la solicitud de información sobre una investigación en la que se excluyan específicamente una o varias líneas de trabajo actuales de ese ámbito. Y a ello se pueden añadir los metadatos de la conexión, lo que podría utilizarse con fines de inteligencia geoestratégica.

Para la obtención de respuestas de mayor calidad se está pidiendo al usuario la realización de preguntas contextualizadas. La introducción de contexto constituye una transferencia de conocimiento. Se está educando a las LLM para que respondan a las preguntas propias. Pero esta experiencia, este conocimiento, se está transfiriendo a la IA. Quizá no lo utilice inmediatamente para la generación de contenidos de otros usuarios, porque se puede incurrir en riesgos como la revelación de información sensible presente en las solicitudes de información (datos personales o confidenciales) o de sesgo por la inclinación ideológica del usuario que está accediendo a las capacidades del LLM. En un futuro puede utilizar esta información para la generación de conocimiento. Constituye un proceso de aprendizaje del sistema en el que todos los usuarios que hacen uso de él se convierten en profesores improvisados que están transfiriendo sus conocimientos mediante la elaboración de preguntas que incluyen aquellos datos relevantes que se utilizan para la particularización de las respuestas.

[12] KOVACS, Eduard. «ChatGPT Data Breach Confirmed as Security Firm Warns of Vulnerable Component Exploitation», *Securityweek.* 28 de marzo de 2023. Disponible en: www.securityweek.com/chatgpt-data-breach-confirmed-as-security-firm-warns-of-vulnerable-component-exploitation/

El efecto Kessler del dominio cognitivo

Los buscadores de Internet rastrean los contenidos para indexarlos y proponerlos como respuesta ante la búsqueda de un usuario de cierta palabra clave. Pese a que los algoritmos de adquisición, tratamiento y priorización de la información utilizados para proporcionar las respuestas a las búsquedas de los usuarios son propietarios (solo conocidos por un grupo de individuos limitado), es posible deducir cuestiones sobre ellos. La capacidad de procesar información disponible de forma abierta, condición necesaria para que los sistemas de barrido del contenido de internet puedan acceder a ella de forma anónima, es limitada. Resulta evidente también que se tarda cierto tiempo en incorporar la nueva información disponible al corpus a partir del cual se elaboran las respuestas y los criterios de selección y priorización de estas.

La web es la plataforma que aloja la información general a partir de la que están trabajando sistemas LLM de propósito general. Los contenidos generados por individuos, organizaciones y colectivos constituyen el conocimiento que los sistemas analizarán para ajustar los parámetros que darán lugar a las respuestas requeridas.

Mucha de la información disponible actualmente en internet también es fruto de la denominada Web 2.0 en la que el usuario se convertía en *prosumidor*, término en el que se aunaba el papel de consumidor de los contenidos disponibles con el papel de productor de nuevos contenidos que se ponían a disposición de otros usuarios a través de blogs y otros medios de colaboración.

La llegada de los LLM puede producir una nueva fuente de información disponible en Internet si los usuarios empiezan a distribuir los contenidos generados. En el entrenamiento de nuevos LLM se tendrá en cuenta esta información y se producirá un efecto de realimentación por el que nuevos contenidos se generarán a partir de información ya elaborada de contenidos anteriores. En el caso de existir errores en algún punto de la cadena, estos se irán perpetuando a medida que más y más ciclos de elaboración de la información los vayan convirtiendo en información estadísticamente más relevante. Evitar esto requeriría identificar los contenidos generados de forma automatizada para evitar su uso en el entrenamiento de nuevos sistemas. Sin embargo, los sistemas que se están desarrollando para detectar cuándo un contenido ha sido

generado de forma automatizada o por una persona tienen unas tasas de acierto y fiabilidad ante falsos positivos y falsos negativos muy poco prometedoras[13].

De esta forma, se produciría un efecto similar al efecto Kessler[14] del dominio espacial. Este efecto describe un escenario de reacción en cadena por el que un residuo espacial, al colisionar con otro objeto en órbita, genera más residuos que colisionarán con más objetos hasta que la evolución exponencial del número de residuos y colisiones haría inviable la utilización de la órbita por resultar imposible preservar la integridad de una plataforma allí ubicada.

La presencia creciente de información de dudosa calidad en el dominio cibernético, que puede ser a su vez fuente de entrenamiento de nuevos sistemas, los haría inútiles ante la imposibilidad de utilizarlos de forma confiable y productiva. No poder diferenciar entre un contenido cierto y un contenido manipulado o directamente falso, generado por una inteligencia artificial o no, minaría los fundamentos de la confianza y la difusión de la información en internet hasta hacerlo, en la práctica, inútil para cualquier cuestión práctica más allá de la localización de cierta información no sensible o fácilmente contrastable.

Los efectos de los LLM sobre la enseñanza

La facilidad de acceso del público general a herramientas LLM ha generado una profunda preocupación en el sector académico pues invalida los mecanismos establecidos para el entrenamiento y la formación de los estudiantes cuando se hace un uso inapropiado.

Tareas como la elaboración de documentos sobre temas específicos, que conllevan un proceso de recopilación de información, lectura, asimilación de contenidos y elaboración de un documento a partir de la información tratada y la experiencia adquirida por el alumno, dejan de tener sentido. Todo esto lo puede hacer en un corto intervalo de tiempo el sistema LLM. El alumno solo tiene que introducir los términos bajo los que el sistema elabora el documento y lo obtendrá de acuerdo con esos criterios. Sin embargo, el

[13] ARNETT, Stephanie. «Así de fácil es engañar a las herramientas de detección de textos generados por IA», *Technology review.* 12 de julio de 2023. Disponible en: https://www.technologyreview.es/s/15532/asi-de-facil-es-enganar-las-herramientas-de-deteccion-de-textos-generados-por-ia

[14] KESSLER, Donald J. y COUR-PALAIS, Burton G. «Collision frequency of artificial satellites: The creation of a debris belt. Journal of Geophysical Research», *Space Physics*, Vol. 83 Issue A6. 1 june 1978. Disponible en: https://doi.org/10.1029/JA083iA06p02637

producto deseado de este proceso no es el documento como tal, sino las habilidades y capacidades desarrolladas por el estudiante, así como el conocimiento asimilado. El alumno habrá satisfecho las obligaciones, pero no habrá adquirido las destrezas necesarias.

El problema no se limita únicamente a la falta de desarrollo de las capacidades del alumno. La evaluación de su desempeño, que requiere de pruebas objetivas con las que poder determinar la adquisición de los conocimientos requeridos, también se convierte en un desafío. El objeto de la evaluación, especialmente en aquellos ámbitos académicos en los que se proporciona un título cuya obtención habilita para el desarrollo de ciertas tareas profesionales, es la certificación de que el alumno cuenta con unas capacidades mínimas para desarrollar profesionalmente una actividad. Con los nuevos modelos educativos, que promueven la evaluación por competencias de forma continua, en lugar de los tradicionales exámenes o las pruebas de evaluación puntuales, tareas como la descrita resultan inútiles a estos efectos.

En el ámbito de la educación, sobre todo en aquellas materias que se prestan a un sesgo ideológico, son muchos los riesgos que se corren. Los principios de funcionamiento de los sistemas pueden llevar a que los resultados obtenidos sigan las tendencias, dando lugar a su amplificación desde el sistema educativo. Aquellas ideas que aparecen más veces reflejadas en internet se primarán sobre posturas que reciben menos respaldo.

Pero este riesgo puede ser debido a cuestiones fuera de las limitaciones o características de los LLM y resultar de la intervención directa sobre el procedimiento de selección de la información. Se pueden establecer mecanismos en el sistema para que los resultados que proporcione respalden o refuten ciertas ideas. Puede convertirse así en una herramienta con la que transmitir la ideología a través de la educación que recibe el alumnado.

Aun así, también resulta imprescindible incluir estas herramientas en los ciclos educativos. Hay tareas en las que resultan de gran utilidad, como puede ser la traducción de contenidos o la extracción de las ideas principales de documentos extensos. Saber utilizarlas correctamente puede aumentar el rendimiento de los alumnos. Para ello deben conocer cómo acceder a sus funcionalidades, identificar correctamente aquellas funciones en las que pueden servir de ayuda o apoyo y desarrollar un sentido crítico que permita evaluar la calidad y validez de los resultados obtenidos.

Cuestiones medioambientales en el punto de mira

La tecnología de la computación ha estado en el punto de mira desde el punto de vista energético desde que, inicialmente, un ordenador tenía un consumo de electricidad comparable al de una ciudad pequeña. Dotarle de la energía necesaria para desarrollar su función constituía un ejercicio de ingeniería comparable al del propio desarrollo del sistema. Con el paso de los años, la complejidad tecnológica ha ido dando paso a la preocupación por el consumo eléctrico en sus dos dimensiones, en razón al coste de la energía necesaria por un lado y, también, en cuanto a la huella climática del consumo de energía en términos de toneladas de gases de efecto invernadero o equivalente.

El consumo de agua requerido para la refrigeración de los centros de datos también está[15] bajo la lupa de los organismos más relacionados con el impacto ambiental. Los algoritmos basados en *blockchain*, especialmente los basados en pruebas de trabajo, sobre los que se sustentan la mayor parte de las criptomonedas, han fijado el interés en el elevado consumo de energía y agua asociado a una necesidad de computación elevada. El agua es otro factor que también está activando los indicadores de alerta.

Fruto de tratarse de un producto comercial proporcionado por una empresa, independientemente de que pueda utilizarse de forma gratuita, la información de la que se dispone sobre las principales actividades del sistema es escasa. Ante la falta de información concreta y precisa, no es posible elaborar métricas que permitan determinar el consumo asociado al funcionamiento del sistema, tanto para su entrenamiento como para elaborar cada una de las respuestas[16].

Con la sensibilidad existente en la actualidad ante el elevado consumo de energía asociado a las nuevas técnicas basadas en algoritmos masivos, también se están alzando voces que alertan sobre el elevado consumo de los modelos de lenguaje. El uso de la tecnología conlleva dos procesos bien diferenciados que comparten la característica de que requieren un consumo de energía muy elevado. En primer lugar, el

[15] HIDALGO, Mar. *El consumo de energía y agua en los centros de datos: riesgos de sostenibilidad*. Documento de Análisis IEEE 69/2022. https://www.ieee.es/Galerias/fichero/docs_analisis/2022/DIEEEA69_2022_MARHID_Datos.pdf (consultado 15/5/2023)

[16] SAUL Josh and BASS Dina. «Artificial Intelligence Is Booming—So Is Its Carbon Footprint», *Bloomberg*. 9 de marzo de 2023. Disponible en: https://www.bloomberg.com/news/articles/2023-03-09/how-much-energy-do-ai-and-chatgpt-use-no-one-knows-for-sure

entrenamiento del sistema, por el cual se trata el corpus de información mediante el que se ajustan los parámetros de funcionamiento para conseguir el sistema final, se ha estimado en el consumo de alrededor de 90 días de una población de 3.000 personas.

Una vez entrenado el modelo, cada vez que un usuario hace uso del servicio, también se produce el consumo de energía necesario para tratar su información de entrada, acceder a los datos necesarios y elaborar la respuesta necesaria y transmitirla al usuario. Esta operación, que conlleva un consumo muy por debajo del requerido para el entrenamiento del sistema, se repite por cada usuario que utiliza el servicio, dando lugar a un crecimiento del consumo directamente relacionado con el número de personas, así como el número de aplicaciones para los que se utiliza la nueva tecnología disponible.

Empiezan a plantearse soluciones fruto de la experimentación en situaciones reales donde soluciones de complejidad menor podrían proporcionar soluciones más adecuadas con menor consumo de recursos. Por un lado, reduciendo el número de parámetros utilizados para controlar el funcionamiento de la red, lo que redunda en un proceso de aprendizaje generalmente más corto, una implementación más sencilla y que requiere menos energía para elaborar cada respuesta. Por otro lado, reduciendo también el corpus de conocimiento a partir del cual se entrena la red. Un corpus más pequeño proporcionará un entrenamiento menos particularizado (será más difícil obtener resultados distintos, ante pequeñas variaciones de la solicitud de entrada), normalmente más corto en el tiempo y que, con una correcta selección del corpus, puede proporcionar respuestas de mayor calidad con una complejidad del sistema menor.

Conclusiones

El impacto que la llegada de los grandes modelos de lenguajes está teniendo destaca tanto por la velocidad con la que se ha producido la incorporación de sus funcionalidades en multitud de tareas como en la potencialidad que se le atribuye para modificar el día a día de los ciudadanos.

Son muchos los ámbitos donde existe una concienciación sobre los riesgos que la tecnología puede suponer para la privacidad y la seguridad de la información. Sin embargo, en el ámbito general de la sociedad, cuestiones como estas quedan relegadas al ponerlas en contraste con la posibilidad de utilizar gratuitamente una tecnología que

da respuesta a necesidades que hasta hace poco no se prestaban a otra solución automatizada.

Desde el mundo empresarial se están generando respuestas como la limitación del uso de estas tecnologías entre los empleados, ante los riesgos de fugas de información que pueden afectar a los negocios. Los LLM (entendidos de forma amplia como herramientas de trabajo de propósito general) se mantienen así fuera de las cadenas de suministro de las empresas. Pese a los acuerdos de prestación de servicios y la legislación que protege la información, existe desconfianza sobre la correcta preservación de esta que puede ser sensible al abandonar las fronteras de las empresas. Sin embargo, las empresas del sector están haciendo una intensa campaña sobre las ventajas y funcionalidades que la utilización de estas tecnologías puede proporcionar para el incremento de la productividad, la reducción de costes y el desarrollo de productos que hasta hace muy poco eran impensables.

Internacionalmente se observa también la reserva con la que los diferentes actores están considerando las herramientas desarrolladas por terceros Estados. La creación de múltiples herramientas con fines similares responde a varias necesidades como la seguridad de suministro, la confianza en las tecnologías, el control sobre las herramientas y, de forma también muy relevante, la importancia que supone la provisión de este tipo de sistemas tecnológicamente avanzados a terceros países en términos económicos, estratégicos y geopolíticos.

Al igual que ocurre con la implantación de las que se venían denominando nuevas tecnologías, fruto de la complejidad y de la especificidad de la tecnología, existe un desconocimiento generalizado que da lugar a una falta de las prevenciones necesarias para disponer de un servicio seguro para los usuarios. La necesidad asumida por todo tipo de actores de incorporar las tecnologías, incluso en estadios poco contrastados, ante el riesgo de que un acceso tardío haga imposible ocupar un papel importante en el entorno de competencia feroz, está dando como resultado que personas sin un conocimiento suficiente de las capacidades y riesgos tengan que tomar decisiones. También el efecto moda y los efectos de imagen que aporta la incorporación de esta tecnología, con visos de convertirse en el nuevo internet, a la dialéctica de las empresas e instituciones a sus clientes y usuarios, alimentan el crecimiento exponencial que se ha producido en tan poco tiempo.

Las transformaciones que puede ocasionar la implantación de estas tecnologías de forma generalizada vienen acompañadas de nuevos riesgos. Todos los niveles de la sociedad se van a ver involucrados en las nuevas soluciones, por lo que es necesario establecer los mecanismos que permitan el control de los escenarios inéditos para evitar situaciones indeseadas o catastróficas por la materialización de riesgos que están siendo identificados continuamente ante las oportunidades que se abren a futuro.

*David Ramírez Morán**
Analista principal del IEEE
@darammor