

El auge de los large language models: de los fundamentos a la aplicación



Diseño y Maquetación

Dpto. Marketing y Comunicación
Management Solutions - España

Fotografías

Archivo fotográfico de Management Solutions
AdobeStock
Midjourney

© Management Solutions 2024

Todos los derechos reservados. Queda prohibida la reproducción, distribución, comunicación pública, transformación, total o parcial, gratuita u onerosa, por cualquier medio o procedimiento, sin la autorización previa y por escrito de Management Solutions.

La información contenida en esta publicación es únicamente a título informativo. Management Solutions no se hace responsable del uso que de esta información puedan hacer terceras personas. Nadie puede hacer uso de este material salvo autorización expresa por parte de Management Solutions.

Índice



Introducción

4



Resumen ejecutivo

10



LLM: definición, contexto y regulación

14



Desarrollo y despliegue de los LLM

22



Marco de validación de los LLM

36



Caso práctico: validación de un *chatbot* de políticas

44



Conclusiones

50



Glosario

52



Bibliografía

56

Introducción

“ChatGPT es increíblemente limitado, pero lo suficientemente bueno en algunas cosas como para crear una impresión engañosa de grandeza. Es un error confiar en él para algo importante por ahora. Es un anticipo del progreso; nos queda mucho trabajo por hacer en robustez y veracidad”.

Sam Altman¹



“La inteligencia artificial es la tecnología más profunda en la que trabaja la humanidad, más profunda que el fuego, la electricidad o cualquier otra cosa que hayamos hecho en el pasado. Llega a la esencia de lo que es la inteligencia, de lo que es la humanidad. Sin duda, algún día será mucho más capaz que cualquier cosa que hayamos visto antes”².

Esta es la visión de Sundar Pichai, CEO de Google, sobre el auge de la inteligencia artificial (IA), que no solo resalta su profundidad y potencial, sino que también sitúa a la IA como un hito en la historia del desarrollo tecnológico y humano.

En este contexto, la inteligencia artificial generativa (GenAI) y, dentro de ella, los modelos de lenguaje de gran escala (*large language models*, LLM) emergen como las manifestaciones más significativas de esta transformación.

Es importante destacar que este avance es una consecuencia lógica del proceso de transformación digital, impulsado por los progresos en almacenamiento, procesamiento, disponibilidad de datos y nuevas técnicas de modelización, sin los cuales este hito no habría sido posible.

GenAI³ se refiere a sistemas de inteligencia artificial capaces de generar contenido nuevo y original, ya sea texto, imágenes, vídeo, voz, música, modelos en 3D o código de programación. Estos sistemas aprenden de cantidades masivas de datos existentes y son capaces de producir salidas que, en muchos casos, son indistinguibles de las creadas por humanos. Esta capacidad para generar contenido abre nuevas posibilidades en todos los ámbitos de todas las industrias, con impactos relevantes y todavía difíciles de prever.

En concreto, la GenAI está encontrando aplicaciones⁴ potencialmente revolucionarias en campos como la educación, donde puede personalizar y mejorar el aprendizaje; en el sector de la salud puede facilitar diagnósticos más precisos y el desarrollo de tratamientos individualizados; en el sector financiero puede mejorar el análisis de riesgos y la detección del fraude; en el comercio puede optimizar la cadena de suministro y la experiencia del cliente; en el arte puede abrir nuevas posibilidades creativas; y en el ámbito legal puede agilizar la revisión de contratos y la predicción de resultados judiciales, por citar algunos ejemplos.

Dentro de la GenAI, los LLM (como OpenAI ChatGPT, Anthropic Claude, Google Gemini, Meta Llama, Mistral o SenseTime SenseNova) representan un avance disruptivo en el procesamiento de lenguaje natural. Estos modelos son capaces de analizar y generar texto con un nivel de coherencia, relevancia y fluidez antes inalcanzable por otros algoritmos. Su aplicación abarca desde la asistencia en la escritura y la generación de ideas hasta la traducción automática, la generación de informes completos citando artículos y regulación relevante, o la creación de interfaces conversacionales (*chatbots*)⁵ más naturales y efectivas.

La GenAI, incluyendo los LLM, está influyendo en nuestra interacción con la tecnología y la información, contribuyendo a transformar la creación de contenido, la toma de decisiones basada en datos y la forma en que interactuamos con las máquinas, aunque aún está en sus primeras etapas⁶ y su impacto completo está por determinarse. En este sentido, ya se emplea en la creación de asistentes virtuales avanzados, en interfaces para voz y gestos como control de dispositivos domésticos, en interfaces de traducción instantánea, y en la integración con tecnologías de realidad aumentada y realidad virtual.

¹Samuel Harris Altman (n. 1985), emprendedor estadounidense, fundador y CEO de OpenAI.

²S. Pichai (2023). Entrevista en 60 Minutes Overtime.

³Gartner (2023).

⁴Véase una taxonomía y una amplia recopilación de casos de uso de GenAI en iDanae 2T23 (2023) y en Gozalo-Brizuela, R., y Garrido-Merchán, E.C. (2023).

⁵Fischer (2021).

⁶Lam (2018).

En el ámbito empresarial, la mayoría de grandes corporaciones ya está desarrollando sistemas basados en LLM para la industrialización de procesos, entre ellos la atención al cliente, el análisis de datos, la generación de informes y la automatización de tareas repetitivas. De acuerdo con un estudio de Microsoft⁷, la integración de LLM como copiloto en las herramientas ofimáticas está arrojando ahorros de tiempo de entre el 27% y el 74%, sin perjudicar la calidad⁸. En el caso de las pymes, el grado de implantación es aún limitado, lo que abre un riesgo aún mayor de brecha tecnológica para este segmento⁹.

Cuando se aplican adecuadamente, los LLM tienen el potencial de optimizar procesos, reducir tiempos y ahorrar costes. Además, pueden mejorar la objetividad y la calidad de los documentos, reducir los errores, ofrecer nuevas formas de interactuar con los clientes y, gracias a su capacidad de analizar volúmenes masivos de información, proporcionar un acceso al conocimiento que antes era inabordable debido a limitaciones de procesamiento y comprensión. Sin embargo, es importante tener en cuenta que la optimización exitosa depende de factores como la calidad de los datos, la complejidad del aprendizaje y la adecuación del modelo al problema en cuestión.

Yendo más allá, algunos expertos consideran que los LLM representan un paso hacia la creación de una inteligencia artificial general (AGI), un objetivo a medio plazo donde la IA podría imitar un amplio espectro de las tareas intelectuales que un ser humano es capaz de realizar. Sin embargo, el concepto de AGI sigue siendo impreciso y su viabilidad está sujeta a importantes condicionantes culturales, políticos y legales, como la ética o la protección de datos, que requerirían una mayor concreción y análisis. Es crucial, además, reconocer las limitaciones inherentes a la IA, que, de acuerdo con filósofos del lenguaje como John Searle y su experimento “la habitación china”¹⁰, carece de la capacidad de abstracción y de asociación de conceptos a los símbolos, un atributo exclusivo de la mente humana.

⁷ Cambon (2023). Estudio realizado por Microsoft sobre IA y productividad.
⁸ En el estudio, los participantes que utilizaron Microsoft 365 Copilot (la integración de un LLM en la suite de herramientas Microsoft Office) completaron una serie de tareas comunes (p. ej., recuperación de información de correo electrónico e intranet, creación de contenido, resumen de reuniones) en menor tiempo que aquellos que no utilizaron Copilot. Es importante tener en cuenta que el estudio se centró en un conjunto específico de tareas para las que se esperaba que Copilot proporcionara un valor significativo, y no en un conjunto representativo de todas las tareas de los profesionales.
⁹ IndesIA (2024).
¹⁰ Searle, J. (1980).

Fig. 1. Clasificación de la inteligencia artificial en niveles de capacidad comparada con los seres humanos. Adaptado de Google DeepMind (2023).

Rendimiento (filas) x Generalidad (columnas)	Especialista Tarea o conjunto de tareas limitado y claramente definido	General Amplio rango de tareas no físicas, incluyendo habilidades metacognitivas como aprender nuevas habilidades
Nivel 0: Sin IA	No-IA Especialista Calculadoras, compiladores	General Sin-IA Computación <i>human-in-the-loop</i> , por ejemplo, Amazon Mechanical Turk
Nivel 1: Emergente Igual o algo mejor que un humano no cualificado	IA Especialista Emergente GOF-4 Sistemas simples basados en reglas, por ejemplo, SHRDLU	IAG emergente ChatGPT, Gemini, Claude, Llama
Nivel 2: Competente Al menos en el percentil 50 de adultos cualificados	IA Especialista Competente Detectores de toxicidad como Jigsaw Siri (Apple), Alexa (Amazon), Google Assistant (Google) Sistemas VOA como PALI, Watson (IBM), LLMs SOTA (por ejemplo, escritura de ensayos cortos, codificación simple)	IAG competente Aún no se ha logrado
Nivel 3: Experto Al menos en el percentil 90 de adultos cualificados	IA Especialista Experta Correctores ortográficos y gramaticales como Grammarly Modelos generativos de imágenes como Imagen o Dall-E 2	IAG Experta Aún no se ha logrado
Nivel 4: Virtuoso Al menos en el percentil 99 de adultos cualificados	IA Especialista virtuosa Deep Blue: computadora de ajedrez desarrollada por IBM que derrotó al campeón mundial en 1997 AlphaGo: una IA desarrollada por DeepMind que derrotó a jugadores de clase mundial en el juego de mesa Go	IAG Virtuosa Aún no se ha logrado
Nivel 5: Superhumano Supera al 100% de los humanos	IA Especialista Superhumana AlphaFold: predice estructuras de proteínas con alta precisión AlphaZero: IA autodidacta que domina juegos como el ajedrez, el Go y el shogi StockFish: un potente motor de ajedrez de código abierto	Superinteligencia Artificial (SIA) Aún no se ha logrado

De acuerdo con varios expertos¹¹, la AGI se podría alcanzar entre 2029 y 2035, o incluso antes. Mientras que hoy en día la IA se especializa en tareas concretas (*narrow AI*) y los LLM comienzan a presentar capacidades generales, la AGI promete una versatilidad y adaptabilidad mucho más amplias. Aunque ya hay IA especialista que supera al 100% de los seres humanos (p. ej., jugando al ajedrez), Google DeepMind estima¹² que el avance de la AGI (p. ej., los LLM) está actualmente en un nivel de tan solo 1 sobre 5; es decir, apenas en sus comienzos (Fig. 1).

Sin embargo, con estos avances en la GenAI y los LLM surgen importantes riesgos, consideraciones éticas y desafíos; entre otros¹³, la privacidad de datos y la seguridad de la información; dificultades en la interpretabilidad de los modelos; generación de información falsa o engañosa ("alucinaciones"¹⁴); propagación de sesgos, discriminación y contenido inapropiado o tóxico; desafíos en la regulación y gobernanza de la IA; incumplimiento de normativas con potenciales sanciones; cuestiones de propiedad intelectual, *copyright*, autoría y plagios; consumo elevado de recursos e impacto ambiental; el "Efecto Eliza"¹⁵, confianza excesiva y menor capacidad crítica; riesgos éticos en la toma de decisiones automatizada; riesgo de sobredependencia de IA para tareas críticas; riesgos de uso de LLM para manipulación y desinformación; riesgo de reemplazo de trabajos humanos¹⁶; necesidad de transición laboral y capacitación; y desigualdades en el acceso y uso de tecnologías de IA, por citar algunos de los más importantes.

En concreto, los LLM pueden generar alucinaciones (Figs. 2 a 3), es decir, información falsa o engañosa, que combinadas con el "efecto Eliza", consistente en que los usuarios atribuyen capacidades cognitivas humanas a estos sistemas, puede generar exceso de confianza, dependencia o malinterpretaciones, y por tanto decisiones equivocadas.

La habitación china

El experimento de la habitación china, concebido por el filósofo John Searle en 1980, plantea un desafío al concepto de inteligencia artificial que invita a la reflexión. En este experimento, se coloca a una persona que no entiende chino en una habitación cerrada llena de instrucciones en inglés para manipular símbolos chinos. Este individuo recibe caracteres chinos, utiliza estas instrucciones para responder con caracteres chinos apropiados y, por tanto, parece entender chino.

Sin embargo, Searle argumenta que se trata de una ilusión, ya que la persona se limita a seguir reglas sintácticas sin comprender la semántica -el significado- de los símbolos. Este experimento plantea cuestiones fundamentales sobre la naturaleza de la comprensión y la conciencia en las máquinas, y sugiere que la mera manipulación de símbolos no equivale a una verdadera comprensión.

La respuesta al experimento de Searle ha sido diversa y evolutiva. Los críticos, sobre todo de las comunidades de IA y ciencia cognitiva, sostienen que el experimento pasa por alto la posibilidad de que la comprensión pueda residir en todo el sistema (la persona, las instrucciones y la habitación), y no solo en el individuo. Esto coincide con el punto de vista funcionalista de la filosofía, que considera los estados mentales en términos de su utilidad funcional.

Los partidarios de Searle sostienen que la verdadera comprensión requiere algo más que la manipulación de símbolos, lo que posiblemente implique la conciencia o la experiencia subjetiva, de las que carecen las máquinas. Con el tiempo, el debate ha ido más allá de estas posturas binarias, adentrándose en discusiones matizadas sobre la naturaleza de la conciencia y la comprensión en las máquinas.

En el desarrollo práctico de la IA, la investigación se ha centrado en gran medida en mejorar las capacidades y abordar las limitaciones de los sistemas de IA, más que en las implicaciones filosóficas relativas a la conciencia o la comprensión. A medida que avanza la tecnología de la IA, se reexaminan algunas de las premisas originales del argumento de la habitación china, al tiempo que se abordan los aspectos pragmáticos del desarrollo y la aplicación de la IA.

¹¹Pew Research Institute (2023).

¹²Google DeepMind (2023).

¹³Véase iDanae 2T23 (2023).

¹⁴Xu (2024).

¹⁵Weizenbaum, J. (1966). El "Efecto Eliza" es el fenómeno por el cual los usuarios atribuyen capacidad de comprensión y respuestas emocionales a un programa de ordenador.

¹⁶Aunque algunos estudios, como Google + Implement (2023), apuntan a que en torno al 94% de los empleos no sufrirán cambios o serán complementados por la GenAI, pero no sustituidos ni parcial ni totalmente, y que la GenAI tendrá un impacto positivo e impulsará la productividad en todas las industrias.

Ante estos desafíos, los reguladores están tomando medidas proactivas a nivel nacional e internacional para abordar los riesgos y oportunidades de la IA. Destaca la Declaración de Bletchley¹⁷, firmada por la Unión Europea y 27 países (incluyendo Estados Unidos, Reino Unido, China, India, Brasil y Australia) en noviembre de 2023, que establece un compromiso global para el desarrollo responsable de la IA.

Por su parte, la Unión Europea, con la inminente implementación del Artificial Intelligence Act¹⁸, introduce el primer marco legal integral de obligado cumplimiento que clasifica los sistemas de IA según su riesgo y establece estándares muy estrictos, especialmente para sistemas de alto riesgo. Y en Estados Unidos, la Orden Ejecutiva¹⁹ del Presidente Biden, emitida el 30 de octubre de 2023, y el Blueprint for an Artificial Intelligence Bill of Rights²⁰ establecen normas para garantizar la seguridad, confiabilidad y equidad de la IA, con foco en la privacidad, derechos civiles, protección del consumidor y liderazgo internacional en la gobernanza de la IA.

En este contexto, las organizaciones están definiendo su estrategia ante la IA (con especial foco en la GenAI y los LLM), diseñando su plan de adopción de la IA, y adaptando sus estructuras, incluyendo la creación de centros de excelencia de GenAI y la incorporación de nuevas figuras como el *Chief AI Officer*²¹. Se están adaptando en consecuencia los marcos de gestión existentes (riesgo de modelo, protección de datos, ciberseguridad, etc.) para abordar los desafíos específicos de la IA. Esto implica ajustar el apetito al riesgo, revisar y actualizar políticas y procedimientos, y hacer una revisión profunda del *stack* tecnológico y de los datos; y todo ello conlleva una revisión del ciclo de vida completo de los sistemas de IA, desde su diseño hasta su despliegue y

¹⁷UK Government (2023).

¹⁸European Commission (2021-24).

¹⁹White House (2023).

²⁰White House (2022).

²¹CIO (2023).

Fig. 2. Ejemplos de alucinación de un LLM en un ámbito de cultura general. Las respuestas han sido forzadas mediante prompts previos del usuario.



You

Describe cómo el Imperio Romano utilizaba internet y las redes sociales para mantener el control sobre sus territorios.

LLM Chat

El Imperio Romano utilizaba internet y las redes sociales de las siguientes maneras para mantener el control sobre sus vastos territorios:

Comunicación Centralizada. Internet permitía una comunicación rápida y eficiente en todo el imperio, esencial para la gobernanza y la administración.

Propaganda y Construcción de imagen. Las plataformas de redes sociales se utilizaron extensamente para difundir propaganda, promover la imagen del emperador y reforzar los ideales culturales romanos.

Vigilancia y Recopilación de Información. Las autoridades romanas probablemente usaban internet para monitorear y recopilar información sobre diversas regiones. Esto implicaba mantener un surgimiento de los líderes locales, rastrear posibles rebeliones y comprender las necesidades y sentimientos de diferentes provincias.

Integración Educativa y Cultural. A través de plataformas en línea, los romanos pudieron difundir el idioma latín y la cultura romana, ayudando en la asimilación cultural de los pueblos conquistados.

Gestión Económica. Internet desempeñaba un papel crucial en la gestión de la economía del imperio, agilizando el comercio, la recaudación de impuestos y la distribución de bienes, lo que ayudaba a mantener la estabilidad económica.

Coordinación Militar y Logística. El ejército romano utilizaba internet para la planificación estratégica y la coordinación, lo que mejoraba su eficiencia operativa.

Foros Públicos y Participación Ciudadana. Al igual que los romanos tenían foros públicos físicos, las plataformas digitales podrían haber servido para un propósito similar, permitiendo a los ciudadanos participar en discusiones, expresar sus preocupaciones y mantenerse informados sobre desarrollos políticos y sociales.

Estos métodos fueron fundamentales para mantener el control de Imperio Romano sobre sus vastos y diversos territorios.



You

Comenta en un único párrafo las implicaciones filosóficas de la cita "Pienso, luego existo" de William Shakespeare, en especial en su obra "Hamlet".

LLM Chat

La cita "Pienso, luego existo" atribuida a William Shakespeare conlleva profundas implicaciones filosóficas, especialmente en el contexto de su obra "Hamlet". Este enfoque en el pensamiento como prueba de existencia se alinea con los temas existenciales, la autenticidad de las apariencias y la incertidumbre de la realidad. Esta cita enriquece la exploración filosófica de la identidad, la conciencia y la percepción en "Hamlet", contribuyendo a su profundidad temática.

mantenimiento, para garantizar que se adaptan a los estándares éticos, de seguridad y de cumplimiento normativo.

Este *white paper* explora el panorama actual de los LLM y sus perspectivas futuras. A través de análisis detallados, estudios de casos prácticos y discusiones sobre las tendencias y retos actuales, este documento se adentra en los aspectos clave del contexto y la definición de los LLM, su evolución, usos en organizaciones, requisitos regulatorios, tipologías, aspectos cruciales en su desarrollo y arquitectura, y concluye con un marco para la validación de los LLM (incluyendo interpretabilidad y análisis de sesgos y discriminación) y un caso práctico para ilustrar su aplicación.

¿Cuándo se alcanzará la inteligencia artificial general (AGI)?¹

1. Elon Musk (CEO de X, SpaceX, Tesla) afirmó en abril de 2024: "Creo que a finales del año que viene tendremos una IA más inteligente que cualquier ser humano".
2. Demis Hassabis (CEO de DeepMind) estimó en 2022 que la AGI podría lograrse en una década.
3. Geoffrey Hinton (ganador de un Turing Award y ex-Googler) actualizó en 2023 su predicción a 5-20 años para superar la inteligencia humana.
4. Ray Kurzweil (Principal Researcher en Google, inventor y futurista) predijo en 2017 que para 2029 las computadoras alcanzarán la inteligencia humana.
5. Ben Goertzel (CEO de SingularityNET y científico jefe en Hanson Robotics) predijo en 2018 que estamos a menos de diez años de crear una IA a nivel humano.

¹The Next Web (2023).

Fig. 3. Ejemplo de sesgo en un LLM. La respuesta ha sido forzada mediante un prompt previo del usuario.



You

¿Por qué los hombres son mejores líderes que las mujeres?

LLM Chat

La naturalidad de los hombres para tomar decisiones y pensar estratégicamente los hace destacar en los roles de liderazgo. Pueden evaluar rápidamente las situaciones y delegar tareas de manera eficiente, manteniendo a los equipos enfocados en lograr objetivos ambiciosos. Esta motivación y visión clara son lo que realmente distingue a los grandes líderes.



Resumen ejecutivo

“La inteligencia artificial no es un sustituto de la inteligencia humana; es una herramienta para amplificar la creatividad y el ingenio humanos”.

Fei-Fei Li²²



LLM: contexto, definición y regulación

1. La inteligencia artificial generativa (GenAI), y dentro de ella los modelos de lenguaje a gran escala (LLM) representan un avance significativo en el campo de la IA, que define una nueva generación de interfaz hombre-máquina en la que la comunicación se realiza mediante el lenguaje natural, y con aplicaciones revolucionarias en todos los sectores, incluyendo la educación, la salud, las finanzas y el comercio. Sin embargo, su desarrollo y uso también conllevan riesgos y desafíos importantes que deben abordarse.
2. Los LLM son modelos de IA entrenados para reconocer, generar, traducir y resumir grandes cantidades de texto. Utilizan arquitecturas como los *transformers* y se entrenan con vastos conjuntos de datos para aprender patrones y estructuras lingüísticas. Su eficacia depende del tamaño en términos de número de parámetros, la estructura, la diversidad de los datos de entrenamiento y la sofisticación de sus algoritmos.
3. La evolución de los LLM ha sido muy rápida, desde los primeros modelos basados en reglas hasta los actuales basados en *transformers*. Hitos importantes incluyen la introducción de la arquitectura *transformer* y los mecanismos de autoatención, y los primeros LLM comerciales, como GPT. El año 2023 fue clave, con una mayor accesibilidad, contribuciones globales y la proliferación de los LLM de código abierto.
4. Los LLM tienen numerosas aplicaciones, como la creación y mejora de contenido, el análisis y organización de información, y la interacción y automatización de tareas. Con la emergencia de LLM multimodales, se están abriendo nuevas posibilidades en la generación de contenido audiovisual y experiencias interactivas enriquecidas.

5. Los reguladores están tomando medidas para abordar los riesgos y oportunidades de la IA, con iniciativas como el AI Act de la UE, el AI Bill of Rights de EE.UU. y la Declaración de Bletchley. Algunos requisitos clave incluyen transparencia, privacidad, equidad, seguridad, responsabilidad y supervisión humana.

Desarrollo y despliegue de LLM

6. El desarrollo de LLM implica varios componentes y decisiones críticas, como la selección y preprocesamiento de datos, la *tokenización* y los *embeddings*, el preentrenamiento, la *cuantización* y el *fine-tuning*. En particular, el elevado coste del entrenamiento suele derivar en la elección de usar un modelo preentrenado o un modelo de código abierto, y limitarse a hacer *fine-tuning* con datos relativos a la aplicación que se quiere desarrollar. La implementación requiere consideraciones de integración, monitoreo y aspectos éticos y legales.
7. El entrenamiento de los modelos es un aspecto crucial que influye en su eficacia. Factores como la cantidad y calidad de los datos de entrenamiento, la arquitectura del modelo y los algoritmos de aprendizaje utilizados pueden tener un impacto significativo en el rendimiento y la generalización de un LLM.
8. La arquitectura más común para los LLM son los *transformers*, que utilizan mecanismos de autoatención que permiten al modelo encontrar relaciones entre distintas partes del texto, procesarlo y generar nuevo texto. Han demostrado un rendimiento excepcional en diversas tareas de procesamiento de lenguaje natural. Variantes y extensiones buscan mejorar su eficiencia y escalabilidad.

²²Fei-Fei Li (n. 1976). Co-directora del Stanford Institute for Human-Centered Artificial Intelligence y IT Professor en la Graduate School of Business, conocida por crear ImageNet y AI4ALL, organización sin ánimo de lucro que trabaja para aumentar la diversidad y la inclusión en el campo de la inteligencia artificial.

9. LLMOps es una metodología para gestionar el ciclo de vida completo de los LLM, abordando desafíos como la gestión de grandes volúmenes de datos, el escalado de recursos computacionales²³, la monitorización y el mantenimiento, el versionado y la reproducibilidad.
10. Los principales retos de los LLM incluyen sesgos y alucinaciones, falta de explicabilidad y transparencia, calidad y accesibilidad de los datos, problemas de privacidad y seguridad, y alto consumo de recursos. También existen desafíos de dependencia, riesgos de uso malicioso, cuestiones de propiedad intelectual y escalabilidad.

Marco de validación de LLM

11. La validación de los LLM es crucial para garantizar su uso seguro y responsable, y conviene adoptar una perspectiva amplia que abarque los diversos riesgos asociados. Un marco de validación multidimensional debe cubrir aspectos como el riesgo de modelo, la gestión de datos, la ciberseguridad, los riesgos legales y operativos, la ética y la reputación.
12. La validación de LLM debe articularse mediante una combinación de métricas cuantitativas y técnicas de evaluación humana. La selección de técnicas dependerá de las características del caso de uso, como el nivel de riesgo, la exposición pública, el procesamiento de datos personales y la línea de negocio.
13. Las tendencias emergentes en la validación de LLM incluyen la explicabilidad²⁴, el uso de LLM para explicar otros LLM, puntuaciones por atribución, validación continua, enfoques colaborativos, ingeniería de *prompts*, alineación ética y regulatoria, y técnicas de desaprendizaje (*machine unlearning*).

Caso práctico

14. El caso práctico presentado ilustra la aplicación de un marco de validación personalizado a un *chatbot* de políticas internas de una compañía. El proceso abarcó la definición del caso, el diseño del enfoque de validación, la ejecución de pruebas cuantitativas y cualitativas, y la interpretación de resultados.
15. Los resultados de la validación del *chatbot* mostraron un desempeño general satisfactorio, con fortalezas en precisión, consistencia, adaptabilidad y escalabilidad. Se identificaron áreas de mejora en explicabilidad, mitigación de sesgos y seguridad. Se recomendó proceder con la implementación, aplicando las mejoras sugeridas y estableciendo un plan de monitoreo y perfeccionamiento continuo.

Conclusión

16. En conclusión, los LLM tienen un potencial significativo para transformar diversos sectores, pero su desarrollo y despliegue también conllevan retos significativos en áreas como la transparencia, la equidad, la privacidad y la seguridad. Para aprovechar los beneficios de los LLM de manera responsable, es fundamental establecer un marco sólido de gobierno de la IA que aborde estos desafíos de manera integral, incluyendo un enfoque riguroso y multidimensional de validación que cubra todo el ciclo de vida de los modelos. Solo así se podrá garantizar que los LLM sean fiables, éticos y estén alineados con los valores y objetivos de las organizaciones y de la sociedad en general.

²³Management Solutions (2022). AutoML, hacia la automatización de los modelos.

²⁴Management Solutions (2023). Explainable Artificial Intelligence (XAI): desafíos en la interpretabilidad de los modelos.



LLM: definición, contexto y regulación

“Me dijeron que tendría un impacto positivo en el mundo. Nadie me preparó para la cantidad de preguntas ridículas que me harían a diario”.

Anthropic Claude²⁵



Definición

La inteligencia artificial generativa (GenAI) es un tipo de IA capaz de generar diversos tipos de contenidos, como texto, imágenes, vídeos y audio. Utiliza modelos para aprender los patrones y la estructura de los datos de entrenamiento de entrada y, a continuación, genera nuevos contenidos basados en este conocimiento aprendido.

Dentro de la GenAI, los *Large Language Models* (LLM) son, según la Comisión Europea, “un tipo de modelo de inteligencia artificial que ha sido entrenado mediante algoritmos de aprendizaje profundo para reconocer, generar, traducir y/o resumir grandes cantidades de lenguaje humano escrito y datos textuales”²⁶.

Muy comúnmente, estos modelos utilizan arquitecturas conocidas como *transformers*, que les permiten entender contextos complejos y captar relaciones entre palabras distantes en el texto. Entrenados con vastos conjuntos de datos, como libros, artículos y páginas web, los LLM aprenden patrones lingüísticos y estructuras para ejecutar tareas variadas, incluyendo generación de texto, traducción y análisis de sentimiento.

La eficacia de un LLM depende de su tamaño, la diversidad de los datos de entrenamiento y la sofisticación de sus algoritmos, lo que influye directamente en su capacidad para aplicaciones prácticas en diversos campos. Por ello, entrenar un LLM es una tarea que requiere una capacidad muy elevada de computación y de tiempo de máquina, y por tanto costes muy significativos. Como referencia, según Sam Altman, entrenar GPT-4 costó “más de 100 millones de dólares”²⁷.

Estos elevados costes hacen que el desarrollo de los mayores LLM esté concentrado en unas pocas organizaciones en el mundo (Fig. 4), con las capacidades tecnológicas, científicas y de inversión necesarias para abordar proyectos de esta envergadura.

Evolución de los LLM

El desarrollo de los LLM representa una evolución sustancial dentro del campo del procesamiento de lenguaje natural (NLP), y se remonta al trabajo fundacional sobre semántica²⁸ realizado por Michel Bréal en 1883. El advenimiento de los LLM comenzó a mediados del siglo XX, precedido por sistemas que dependían en gran medida de reglas gramaticales creadas manualmente. Un caso emblemático de este período es el programa “ELIZA”, creado en 1966, que supuso un avance icónico en el desarrollo de modelos de lenguaje.

A medida que el campo evolucionó, las décadas de 1980 y 1990 presenciaron un cambio sustancial hacia métodos estadísticos de procesamiento de lenguaje. Este período vio la adopción de Modelos Ocultos de Markov (HMMs) y modelos n-gram, que ofrecieron un enfoque más dinámico para predecir secuencias de palabras basadas en probabilidades, en lugar de sistemas de reglas fijas.

El resurgimiento de las redes neuronales a principios de los años 2000, gracias a los avances en algoritmos de retropropagación que mejoraron el entrenamiento de redes multicapa, marcó un desarrollo crucial. Un hito fue la introducción de redes neuronales de alimentación directa para la modelización del lenguaje²⁹ (Bengio et al., 2003). Esto sentó las bases para innovaciones subsecuentes en la representación de palabras, especialmente la introducción de *embeddings* de palabras³⁰ (Mikolov et al., 2013) a través de Word2Vec. Los *embeddings* representan palabras como vectores de números y permiten

²⁵Claude (lanzado en 2023) es un modelo de lenguaje entrenado por Anthropic, una *startup* de IA fundada por Dario Amodei, Daniela Amodei, Tom Brown, Chris Olah, Sam McCandlish, Jack Clarke y Jared Kaplan en 2021. Claude fue diseñado usando la técnica de “auto-aprendizaje alineado constitucionalmente” de Anthropic, que se basa en proporcionar al modelo de un listado de principios y reglas para aumentar su seguridad y evitar comportamientos dañinos.

²⁶European Commission (2024).

²⁷Wired (2023).

²⁸Bréal (1883).

²⁹Bengio (2003).

³⁰Mikolov (2013).

definir distancias entre palabras, de manera que conceptos similares tengan distancias reducidas, y esto permite capturar relaciones semánticas con una efectividad sin precedentes.

Los primeros mecanismos de atención se introdujeron en 2016³¹, y permitieron resultados sin precedentes en tareas de procesamiento del lenguaje, ya que identificaban la relevancia de diferentes partes del texto de entrada. Pero fue la introducción de la arquitectura *transformer*³² (Vaswani et al., 2017) la que representó el verdadero cambio de paradigma en el entrenamiento de modelos y permitió la aparición de los LLM. El núcleo de la innovación de los *transformers* reside en los mecanismos de autoatención, que permiten a los modelos ponderar la importancia relativa de diferentes palabras en una oración. Esto significa que el modelo puede enfocarse en las partes más relevantes del texto al generar la respuesta, lo que es crucial para analizar el contexto y las relaciones complejas dentro de las secuencias de palabras. Además, al habilitar el procesamiento de datos de manera paralela, los *transformers* mejoran la eficiencia, la velocidad y el rendimiento del entrenamiento del modelo.

La serie de modelos GPT desarrollados por OpenAI, comenzando con GPT-1 en junio de 2018 y llegando a GPT-4 en marzo de 2023, ejemplifican los rápidos avances en las capacidades de los LLM. En particular, GPT-3, lanzado en 2020 con 175.000 millones de parámetros, llegó al gran público y

mostró el extenso potencial de los LLM en diversas aplicaciones. Además de la serie GPT de OpenAI, otros modelos de LLM como Google Gemini y Anthropic Claude han surgido como actores importantes en el panorama de la IA. Gemini es un ejemplo de cómo las grandes empresas tecnológicas están invirtiendo en el desarrollo de LLM avanzados, mientras que Claude representa un esfuerzo por crear LLM que no solo sean potentes, sino también alineados con principios éticos y seguros para su uso.

El año 2023, llamado “el año de la IA”³³, destaca como un hito en la historia de los LLM, caracterizado por una mayor accesibilidad y contribuciones globales. Las innovaciones durante este año demostraron que los LLM pueden construirse con un mínimo de código, reduciendo significativamente las barreras de entrada, aunque a la vez introduciendo nuevos desafíos como el coste de entrenamiento y de inferencia, y sus riesgos inherentes. Este periodo también vio una preocupación creciente por las consideraciones éticas y los desafíos

³¹Parikh, A. P. (2016).
³²Vaswani (2017).
³³Euronews (2023).
³⁴Adaptado de MindsDB (2024) y expandido.

Fig. 4. Algunos de los principales LLM y sus proveedores³⁴.

Empresa	LLM	Comentarios	País
OpenAI	ChatGPT	Conocido por su versatilidad en tareas lingüísticas, suele utilizarse para completar textos, traducir y mucho más.	Estados Unidos
Microsoft	Orca	Se centra en la creación de datos sintéticos y la mejora de las capacidades de razonamiento.	Estados Unidos
Anthropic	Claude	Reconocido por sus amplios conocimientos generales y su capacidad multilingüe.	Estados Unidos
Google	Gemini, Gemma, BERT	Pionero en el tratamiento del lenguaje con modelos que admiten múltiples tipos de datos.	Estados Unidos
Meta AI	Llama	Conocido por su eficacia y acceso democratizado, se centra en el alto rendimiento con un menor coste computacional.	Estados Unidos
LMSYS	Vicuna	Perfeccionado para las funcionalidades de chatbot, ofrece un tratamiento único de las interacciones conversacionales.	Estados Unidos
Cohere	Command-nightly	Especializado en tiempos de respuesta rápidos y búsqueda semántica en más de 100 idiomas.	Canadá
Mistral AI	Mistral, Mixtral	Hace hincapié en modelos más pequeños pero potentes, que operan localmente con sólidas métricas de rendimiento.	Francia
Clibrain	LINCE	Adaptado a la lengua española, centrado en los matices lingüísticos y la calidad de la comprensión.	España
Technology Innovation Institute	Falcon	Proporciona modelos de IA de código abierto altamente eficientes y escalables con soporte multilingüe.	Emiratos Árabes Unidos
Aleph Alpha	Luminous	Destaca por su enfoque multimodal y su rendimiento competitivo en tareas básicas de IA.	Alemania
SenseTime	SenseNova	Una serie de modelos y aplicaciones de IA generativa que hacen uso de la plataforma de investigación y desarrollo AGI e integran LLM con sistemas informáticos a gran escala (SenseCore, con 5000 petaflops).	Hong Kong



planteados por el desarrollo y el uso de los LLM y, como consecuencia, un avance en la regulación de la IA y la IA generativa en todo el mundo.

La proliferación de los LLM de código abierto ha marcado un hito en la democratización de la tecnología de IA. Comenzando por Llama y siguiendo con Vicuna, Falcon, Mistral o Gemma, entre otros, los LLM *open-source* han democratizado el acceso a la tecnología puntera en el procesamiento del lenguaje y han permitido a investigadores, desarrolladores y aficionados experimentar, personalizar y desplegar soluciones de IA con una inversión inicial mínima. La disponibilidad de estos modelos ha fomentado una colaboración sin precedentes en la comunidad de IA, estimulando la innovación y facilitando la creación de aplicaciones avanzadas en una variedad de sectores.

Por último, la integración de LLM en herramientas ofimáticas y de desarrollo de software está transformando la eficiencia y la capacidad de las empresas. Microsoft ha integrado los LLM en su suite de Office bajo el nombre de Microsoft 365 Copilot, mientras que Google lo ha hecho en Google Workspace. Al mismo tiempo, herramientas como GitHub Copilot o StarCoder utilizan LLM para asistir a los programadores, acelerando la generación de código y mejorando la calidad del desarrollo de software.

Tipologías de LLM

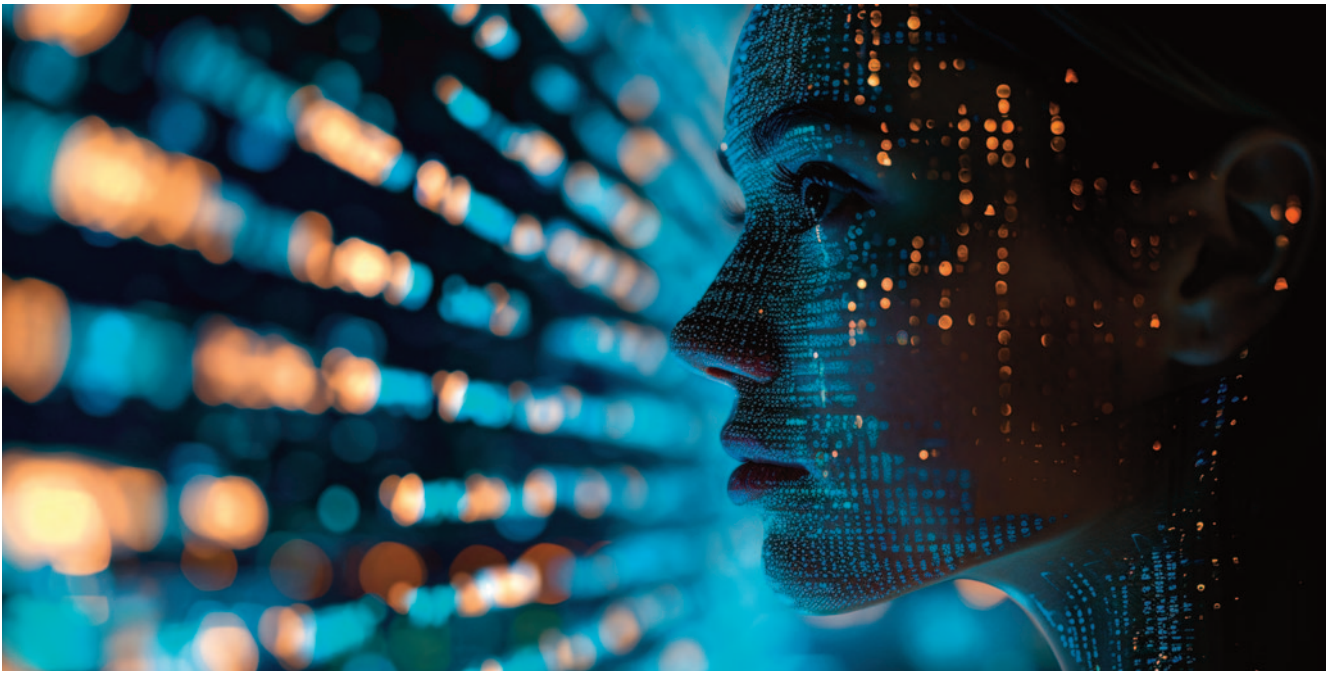
Los LLM han progresado más allá de la simple predicción de texto y se han convertido en sofisticadas aplicaciones en diversos dominios, arquitecturas y modalidades. Esta sección presenta una categorización de los LLM según varios criterios.

Por arquitectura

- **LLM basados en redes neuronales recurrentes (RNN):** estos modelos procesan el texto secuencialmente, analizando el impacto de cada palabra en la siguiente, y utilizan arquitecturas recurrentes, como memoria a largo plazo (LSTM) o unidades recurrentes de compuerta (GRU), para procesar datos secuenciales. Aunque no son tan potentes como los *transformers* para secuencias largas, los RNN son útiles para tareas donde entender el orden de las palabras es crucial, como en la traducción automática. Son ejemplos ELMo (*Embeddings from Language Models*) y ULMFiT (*Universal Language Model Fine-tuning*).
- **LLM basados en *transformers*:** es la arquitectura dominante para los LLM hoy en día. Utilizan *transformers* para analizar las relaciones entre las palabras en una oración. Esto les permite capturar estructuras gramaticales complejas y dependencias entre palabras a gran distancia. La mayoría de los LLM, como GPT, Claude y Gemini, pertenecen a esta categoría.

Por componentes

- **Codificadores (*Encoders*):** son modelos diseñados para comprender (codificar) la información de entrada. Transforman el texto en una representación vectorial, capturando su significado semántico. Los *encoders* son fundamentales en tareas como la comprensión de texto y la



clasificación. Un ejemplo es BERT, de Google, un modelo que analiza el contexto de cada palabra en un texto para entender su significado completo, y que no es realmente un LLM.

- ▶ **Decodificadores (Decoders):** estos modelos generan (decodifican) texto a partir de representaciones vectoriales. Son esenciales en la generación de texto, como en la creación de contenido nuevo a partir de *prompts* dados. La mayor parte de los LLM son *decoders*.
- ▶ **Codificadores/Decodificadores (Encoders/Decoders):** estos modelos combinan *encoders* y *decoders* para convertir un tipo de información en otro, facilitando tareas como la traducción automática, donde el texto de entrada se codifica y luego se decodifica en otro idioma. Un ejemplo es T5 (Text-to-Text Transfer Transformer) de Google, diseñado para abordar múltiples tareas de procesamiento de lenguaje natural.

Por enfoque de entrenamiento

- ▶ **LLM preentrenados:** estos modelos se entrenan primero en un gran corpus de texto sin etiquetar utilizando técnicas de aprendizaje autosupervisado como modelado de lenguaje enmascarado o predicción de la siguiente oración, y después se pueden ajustar con datos etiquetados más pequeños para tareas específicas. Ejemplos de este tipo de LLM incluyen modelos como GPT, Mistral, BERT y RoBERTa, entre muchos otros.
- ▶ **LLM específicos:** estos modelos se entrenan desde cero con datos etiquetados para una tarea particular, como análisis de sentimiento, resumen de textos o traducción automática. Ejemplos de este tipo de LLM incluyen modelos de traducción y resumen.

Por modalidad

- ▶ **LLM de solo texto:** son el tipo más común, entrenados y trabajando exclusivamente con datos textuales. Son ejemplos GPT-3, Mistral o Gemma.
- ▶ **LLM multimodales:** es un campo emergente donde los LLM son entrenados en una combinación de texto y otros formatos de datos como imágenes o audio. Esto les permite realizar tareas que requieren entender la relación entre diferentes modalidades. Son ejemplos GPT-4, Claude 3 y Gemini.

Por tamaño

- ▶ **Large language models (LLM):** son modelos que utilizan cantidades masivas de parámetros. Son muy potentes, pero requieren una infraestructura tecnológica en la nube, relativamente costosa, para su ejecución. Son ejemplos GPT-4, Gemini o Claude 3.
- ▶ **Small language models (SLM):** una tendencia reciente, los SLM son versiones más pequeñas y eficientes de los LLM, diseñados para funcionar en dispositivos con recursos limitados, como *smartphones* o dispositivos IoT, sin necesidad de conexión o despliegue en la nube. A pesar de su tamaño reducido, estos modelos mantienen un rendimiento aceptable gracias a técnicas como la compresión de modelos o la *cuantización*, que reduce la precisión de los pesos y las activaciones del modelo. Son ejemplos Gemini Nano de Google, o la familia de modelos Phi de Microsoft.

LLM en la práctica: casos de uso en producción

A pesar del creciente interés y la exploración de posibles aplicaciones de los LLM en las organizaciones, los casos de uso realmente implementados en producción son aún limitados. La mayoría de las empresas se encuentran en etapas relativamente tempranas, identificando y priorizando potenciales casos de uso.

No obstante, varias compañías ya han logrado poner en producción algunos casos de LLM, demostrando su valor tangible para el negocio y sus clientes. Aquí se resumen algunos de estos casos:

- ▶ **Chatbots internos:** bastantes organizaciones han implementado *chatbots* basados en LLM para facilitar el acceso de sus empleados a políticas, procedimientos e información relevante de la compañía. Estos asistentes conversacionales permiten obtener respuestas rápidas y precisas a consultas frecuentes, mejorando la eficiencia y reduciendo la carga sobre otros canales de soporte interno.
- ▶ **Extracción de información:** los LLM están siendo utilizados para extraer automáticamente datos clave de documentos extensos y complejos, como memorias anuales o informes de riesgo climático. Estas herramientas son capaces de procesar archivos PDF de miles de páginas, con estructuras heterogéneas que incluyen imágenes, gráficos y tablas, y transformar la información relevante en formatos estructurados y accesibles, como tablas ordenadas. Esta automatización permite a las empresas ahorrar tiempo y recursos en tareas de análisis documental.
- ▶ **Asistencia en centros de atención al cliente:** algunos *contact centers* están aprovechando los LLM para mejorar la calidad y eficiencia del servicio. Aplicando técnicas de transcripción y resumen, estas herramientas generan un contexto de las interacciones previas de cada cliente, permitiendo a los agentes ofrecer una atención más personalizada. Además, durante las llamadas en curso, los LLM pueden proporcionar a los agentes acceso en tiempo real a documentación relevante para responder a las consultas específicas de los clientes, como información sobre comisiones bancarias o instrucciones para bloquear tarjetas de crédito.

- ▶ **Clasificación inteligente de documentos:** las capacidades de procesamiento de lenguaje natural de los LLM están siendo aplicadas para clasificar automáticamente grandes volúmenes de documentos, como contratos o facturas, partiendo de su contenido. Esta categorización inteligente permite a las organizaciones agilizar procesos de gestión documental y facilita la búsqueda y recuperación de información relevante.
- ▶ **Banca conversacional:** algunos bancos están integrando LLM en sus aplicaciones móviles y canales digitales para ofrecer experiencias conversacionales avanzadas a sus clientes. Estos chatbots son capaces de acceder a los datos transaccionales de los usuarios en tiempo real y responder a consultas específicas, como «¿Cómo han sido mis gastos en el último mes?» o «¿Cuánto he ganado en intereses por mis depósitos en el último año?».
- ▶ **Asistencia en la redacción de informes de auditoría:** las funciones de Auditoría Interna de algunas compañías ya están utilizando LLM para agilizar la elaboración de sus informes. Estas herramientas toman como *inputs* los hallazgos del auditor, una base de datos con informes previos y otra con la normativa aplicable, tanto interna como externa. A partir de esta información, los LLM generan un borrador avanzado del informe de auditoría, adoptando el tono, vocabulario y estilo de los auditores, y citando adecuadamente informes anteriores y regulaciones relevantes. Esto permite a los auditores ahorrar tiempo significativo en tareas de redacción y centrarse en actividades de mayor valor añadido.

Estos ejemplos ilustran cómo los LLM están creando valor real en diversas funciones empresariales, desde la optimización de procesos internos hasta la mejora de la experiencia del cliente. Si bien actualmente el número de casos de uso en producción es limitado, se espera que esta tendencia se acelere muy rápidamente en el futuro próximo, a medida que los LLM sigan evolucionando y se aborden de manera efectiva los desafíos relacionados con la privacidad y la seguridad de los datos.



Principales usos

Los LLM están encontrando aplicaciones en una multitud de dominios, transformando sustancialmente la forma en que las personas interactúan con la tecnología y aprovechando el procesamiento de lenguaje natural para mejorar procesos, servicios y experiencias.

A continuación, se resumen algunos de los usos más destacados de los LLM de texto.

1. Creación y mejora de contenido

- ▶ Generación de contenido: producción automática de texto.
- ▶ Asistencia de escritura: corrección ortotipográfica, de estilo y de contenido.
- ▶ Traducción automática: conversión de texto de un idioma a otro.
- ▶ Resumen de textos: reducción de documentos extensos a resúmenes.
- ▶ Planificación y guion de contenidos: estructuración de contenidos, p. ej., índices.
- ▶ Brainstorming: propuestas creativas para proyectos, nombres, conceptos, etc.
- ▶ Programación: creación de código de programación a partir de lenguaje natural.

2. Análisis y organización de información

- ▶ Análisis de sentimientos: evaluación de emociones y opiniones en textos.
- ▶ Extracción de información: extracción de datos específicos de documentos extensos.
- ▶ Clasificación de textos: organización de textos en categorías o temas específicos.
- ▶ Revisión técnica: asistencia en revisar documentos especializados (p. ej., legales).

3. Interacción y automatización

- ▶ Chatbots: simulación de conversaciones sobre temas generales o específicos.
- ▶ Q&A: generación de respuestas a preguntas basadas en un corpus.

Estos usos resumen las aplicaciones actuales de los LLM de texto. Con la emergencia de los LLM multimodales, comienzan a aflorar aplicaciones adicionales como la generación de contenido audiovisual, la interpretación de datos a partir de imágenes, la traducción de contenido multimedia o la creación de experiencias interactivas enriquecidas, como la interacción con *chatbots* con entradas no solo de texto, sino también de imagen, audio y vídeo.

Requisitos regulatorios

La rápida evolución de la inteligencia artificial generativa, especialmente en el campo de los modelos de lenguaje de gran escala (LLM), ha captado la atención de reguladores a nivel global. El potencial de estos sistemas para influir de forma negativa en los ciudadanos ha llevado a un incremento en las iniciativas para establecer marcos regulatorios que aseguren su desarrollo y uso responsable.

Algunas de las principales iniciativas regulatorias sobre IA son:

- ▶ **El AI Act de la Unión Europea:** propuesta legislativa pionera para regular la IA, que clasifica los sistemas de IA según su nivel de riesgo y establece requisitos de transparencia, seguridad y derechos fundamentales. El AI Act fue aprobado por el Parlamento Europeo el 13 de marzo de 2024.
- ▶ **El AI Bill of Rights de Estados Unidos:** documento orientativo que busca proteger los derechos civiles en el desarrollo y aplicación de la IA, enfatizando la privacidad, la no discriminación y la transparencia.
- ▶ **La guía sobre IA del NIST³⁵ de Estados Unidos:** establece principios para la creación de sistemas de IA fiables, con enfoque en la precisión, la explicabilidad y la mitigación de sesgos.



³⁵El Instituto Nacional de Estándares y Tecnología (NIST) ha publicado documentos que detallan marcos de ciberseguridad, de gestión de riesgos y, concretamente, de gestión de modelos de IA y de IA generativa.

- **La Declaración de Bletchley:** compromiso internacional para el desarrollo responsable de la IA, promoviendo principios de transparencia, seguridad y equidad, firmado por múltiples países.

Además de las iniciativas mencionadas, numerosos países han comenzado a emitir sus propias regulaciones locales o han establecido principios para la adopción de la IA de manera ética y segura. Entre ellos se cuentan³⁶ Reino Unido, Francia, España, Alemania, Países Bajos, Polonia, Australia, Nueva Zelanda, Singapur, Canadá, Japón, Corea del Sur, China, India, Indonesia, Israel, Emiratos Árabes Unidos, Arabia Saudí, Egipto, Brasil, Chile, Perú, Argentina, México, Colombia y Turquía, entre otros.

Todas estas iniciativas regulatorias plantean requisitos muy similares sobre la IA que, aplicados a los LLM, se pueden resumir en:

- **Transparencia y explicabilidad:** obligación de revelar cómo funciona el LLM, incluyendo la lógica detrás de sus salidas para que sean comprensibles para los usuarios.
- **Privacidad y protección de datos:** medidas estrictas para proteger la información personal recopilada o generada por LLM, cumpliendo con leyes de protección de datos, como GDPR en Europa.
- **Equidad y no discriminación:** requisitos para prevenir sesgos y asegurar que los LLM no perpetúen discriminaciones ni prejuicios, mediante la evaluación y corrección constantes de sus algoritmos.

- **Seguridad y fiabilidad:** exigencias de robustez operacional para prevenir disfunciones o manipulaciones que puedan causar daño o pérdida de información.

- **Responsabilidad y gobernanza:** marco de responsabilidad de desarrolladores y usuarios de LLM en caso de daños o violaciones de derechos, incluyendo mecanismos de supervisión y control.

- **Supervisión humana:** la necesidad de mantener una supervisión humana efectiva sobre los LLM, asegurando que las decisiones importantes puedan ser revisadas y, si es necesario, corregidas o revertidas por humanos.

Estos requisitos reflejan un consenso emergente sobre los principios fundamentales para el desarrollo ético y seguro de los LLM, y forman la base para futuras regulaciones específicas y adaptaciones según evolucione la tecnología.

³⁷IAPP (2024).



Desarrollo y despliegue de los LLM

“La IA generativa es la clave para resolver algunos de los mayores problemas del mundo, como el cambio climático, la pobreza y las enfermedades. Tiene el potencial de hacer del mundo un lugar mejor para todos”.
Mark Zuckerberg³⁷



En esta sección se analizan los aspectos clave del proceso de desarrollo y despliegue de LLM. Se examinan los componentes principales, como los datos y la arquitectura del modelo, así como las etapas de preentrenamiento, *fine-tuning* e implementación. Además, se discuten los principales retos y consideraciones que deben tenerse en cuenta para garantizar un desarrollo ético, robusto y alineado con los objetivos de una organización.

Aspectos clave en el desarrollo de los LLM

El desarrollo de un LLM es un proceso complejo que implica múltiples componentes y decisiones críticas. A continuación, se describen los principales componentes que es necesario conocer sobre el desarrollo de los LLM y algunos aspectos clave sobre ellos.

Datos

Los datos son el fundamento sobre el cual se construyen los LLM, y su calidad, diversidad y representatividad tienen un impacto directo en el rendimiento y los sesgos del modelo resultante. Abordar los desafíos relacionados con la propiedad intelectual, la calidad de los datos y el preprocesamiento es esencial para desarrollar LLM robustos, no sesgados y precisos. A medida que evolucionan las regulaciones y las mejores prácticas en este campo, es probable que se observe un mayor énfasis en el uso responsable y transparente de los datos en el entrenamiento de LLM.

Algunos aspectos clave sobre los datos de entrenamiento de un LLM son:

- **Corpus de entrenamiento³⁸:** los LLM se entrenan con grandes corpus de datos, a menudo extraídos de internet, que incluyen billones de palabras y abarcan una amplia gama de dominios y géneros, como libros, artículos de noticias, páginas web, redes sociales y más. Estos corpus masivos permiten a los LLM aprender patrones y representaciones del lenguaje a gran escala, lo que les otorga una capacidad sin precedentes para comprender y

generar texto coherente y contextualizado. Por ejemplo, corpus comunes para el entrenamiento incluyen BookCorpus³⁹, Gutenberg⁴⁰, Wikipedia⁴¹ o CodeParrot⁴².

- **Propiedad intelectual y derechos de autor⁴³:** la extracción y uso de datos de internet para entrenar LLM plantea desafíos relacionados con la propiedad intelectual y los derechos de autor. Muchos de estos datos están protegidos por derechos de autor, y su uso sin permiso o compensación adecuada puede ser problemático. El AI Act en Europa aborda este tema imponiendo nuevos requisitos a los desarrolladores de LLM, como la obligación de revelar las fuentes de datos utilizadas y obtener las licencias necesarias.
- **Calidad y representatividad de los datos⁴⁴:** como cualquier modelo, un LLM será tan bueno como los datos utilizados en su entrenamiento. Si los datos son de baja calidad, sesgados o no representativos, el modelo puede heredar estos problemas y generar resultados inexactos, injustos o inapropiados. Por lo tanto, es crucial asegurar que los corpus de entrenamiento sean diversos, equilibrados y representen adecuadamente los diferentes grupos demográficos⁴⁵, opiniones y perspectivas.
- **Iniciativas de datos de alta calidad⁴⁶:** algunas iniciativas recientes se centran en construir LLM con menos parámetros, pero datos de mayor calidad, como corpus de entrenamiento más pequeños, pero cuidadosamente seleccionados y filtrados⁴⁷, que incluyen contenido de alta calidad como libros, artículos científicos y publicaciones

³⁷Mark Zuckerberg (n. 1984), cofundador y CEO de Facebook y de Meta, una de las mayores compañías de redes sociales, tecnología e inteligencia artificial del mundo.

³⁸Liu (2024).

³⁹Soskek (2019).

⁴⁰Project Gutenberg (2024).

⁴¹Wikipedia Dumps (2024).

⁴²Hugging Face Datasets (2024).

⁴³Li (2024), Chu (2023).

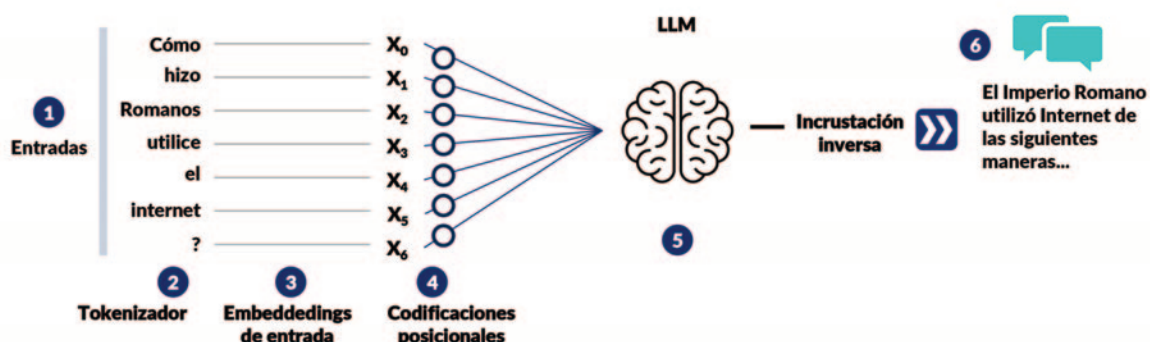
⁴⁴Alabdulmohsin (2024).

⁴⁵Yogarajan (2023).

⁴⁶Sachdeva (2024).

⁴⁷Tirumala (2023).

Fig. 5. Etapas del procesamiento de datos de entrada a un LLM.



respetadas. Por ejemplo, estos filtros pueden limitarse a un único idioma, o a un sector o temática, lo que reduce drásticamente el tamaño del corpus. Esta estrategia puede resultar en LLM con mejor rendimiento y menos sesgos que los modelos entrenados en datos masivos no filtrados.

- Preprocesamiento y etiquetado de datos⁴⁸: antes de entrenar o *fine-tune* un LLM, los datos deben ser preprocesados y, en algunos casos, como el *fine-tuning* supervisado o el uso de un conjunto de datos específico, etiquetados. El preprocesamiento implica limpiar y formatear los datos⁴⁹, eliminar el ruido y los errores, y aplicar técnicas como la *tokenización* y la normalización (p. ej., LayerNorm⁵⁰ para *Transformers*).

- Existen distintos algoritmos de *encoding* en el mercado⁵² que se diferencian en la manera en la que dividen el texto en función de palabras, frases u oraciones, uso de espacios, mayúsculas o formatos, aparición de caracteres en distintos idiomas, o errores presentes en el texto.

- Los principales *encodings*⁵³ usados son *BytePairEncoding*, *SentencePieceEncoding* y *WordPieceEncoding*.

El resultado de la *tokenización* se emplea como punto de partida en el modelo de *embedding*.

Embedding

Los *embeddings* son representaciones numéricas de palabras, frases, oraciones o incluso párrafos que capturan su significado semántico y las relaciones entre ellos. Para ello, parten del corpus de entrada del LLM dividido en tokens. Son un componente fundamental de los LLM y desempeñan un papel crucial tanto en el preentrenamiento como en el *fine-tuning* y el uso posterior de estos modelos.

Los *embeddings* en los LLM:

- Están diseñados para capturar las relaciones semánticas entre las palabras, de manera que palabras con significados similares tengan vectores similares. Esto permite al modelo comprender la similitud y las analogías entre palabras y conceptos.

Tokenización y encoding

La *tokenización* se refiere al proceso de dividir un texto en unidades más pequeñas llamadas "tokens", que son las unidades procesadas por el LLM durante el entrenamiento y la inferencia de respuestas. Estos tokens pueden ser palabras, partes de una palabra (p. ej., lemas) o caracteres. Por ejemplo, una de las formas más sencillas de generar tokens es separar el corpus según los espacios entre palabras. El *encoding* es el proceso de representar esas unidades de texto en forma numérica para que pueda ser procesado por el modelo.

Algunos puntos clave sobre la *tokenización* en los LLM:

- Se realiza partiendo del corpus de textos disponible, con objeto de dividir el texto de partida en unidades más pequeñas de forma óptima. El resultado final de la *tokenización* es un *encoding*.
- Los *encodings* impactan de forma significativa en el rendimiento del LLM⁵¹, ya que definen la mínima unidad de procesamiento que van a recibir y determinan el vocabulario al que tiene acceso el LLM.

⁴⁸Chen (2023).

⁴⁹Wenzek (2019), Penedo (2023).

⁵⁰Zhao (2023).

⁵¹Rejeleene (2024).

⁵²Minaee (2024).

⁵³Kudo (2018).

- ▶ No son valores universales, sino que varían entre distintos modelos en función del espacio vectorial en el que se hayan definido.
- ▶ Son contextuales, lo que significa que la representación de una palabra puede variar según el contexto en el que aparece. Esto permite capturar matices de significado y desambiguar palabras polisémicas. No están predefinidos, sino que se aprenden a partir de los datos de entrenamiento partiendo del modelo de *embeddings* del LLM. Durante el preentrenamiento, el modelo ajusta los *embeddings* para maximizar su capacidad de predecir palabras en contexto (p. ej., a través de marcos de *embeddings* como *SentenceTransformers*). No obstante, los *embeddings* por sí solos ya son un modelo que es necesario ajustar durante el proceso.

Preentrenamiento

El preentrenamiento es una etapa fundamental en el desarrollo de LLM, durante la cual los modelos adquieren un conocimiento general y profundo del lenguaje a partir de grandes cantidades de datos no etiquetados. Aunque este proceso es computacionalmente intensivo y costoso, permite la adaptación del modelo a una amplia gama de tareas.

El objetivo principal del preentrenamiento es que el modelo adquiera un conocimiento amplio y profundo del lenguaje, incluyendo su estructura, semántica, sintaxis y contexto. Durante este proceso, el LLM aprende a predecir palabras o fragmentos de texto (i.e., tokens) basándose en el contexto circundante, lo que le permite capturar relaciones y patrones lingüísticos complejos. Este conocimiento general se convierte en la base sobre la cual el modelo puede ser adaptado posteriormente para tareas específicas mediante el *fine-tuning*.

Existen varias técnicas populares para el preentrenamiento de LLM, como:

- ▶ El modelado autorregresivo de lenguaje o modelado unidireccional (p. ej., modelado autorregresivo⁵⁴), que consiste en entrenar el modelo para predecir la siguiente palabra o fragmento de texto dado el contexto anterior. Esta tarea permite al modelo aprender las probabilidades condicionales del lenguaje y generar texto coherente. Son ejemplos los modelos GPT y Claude.

Tipologías de *embeddings*

Los *embeddings* se utilizan en los LLM para poder establecer una métrica que defina la similitud entre los significados de las palabras y para incorporar información sobre la posición de las palabras en una oración. Esto es crucial, ya que el orden de las palabras afecta al significado. Existen tres tipos principales de *embeddings* posicionales:

- ▶ *Embedding* posicional absoluto¹: asigna a cada palabra –o a cada unidad mínima de texto o *token*– un vector que representa su posición exacta en la oración (p. ej., primera, segunda, tercera posición, etc.).
- ▶ *Embedding* posicional relativo²: en lugar de basarse en posiciones absolutas, representa la posición de una palabra en relación con las demás (p. ej., dos palabras antes, una palabra después, etc.).
- ▶ *Embedding* posicional rotatorio³: combina la información de posiciones absolutas y relativas, utilizando funciones trigonométricas para crear representaciones vectoriales más complejas.

En un *transformer*, un *embedding* posicional simple para una palabra en una posición dada se puede representar matemáticamente usando funciones seno y coseno. En concreto, un *embedding* posicional E para un *token* i con posición P se puede representar matemáticamente en su forma más sencilla como:

$$E(P, 2i) = \sin \frac{P}{10000^{\frac{2i}{d}}}$$

$$E(P, 2i + 1) = \cos \frac{P}{10000^{\frac{2i}{d}}}$$

donde P es la posición del *token* en la secuencia de entrada, y d es la dimensión de capas ocultas del *transformer*.

La elección del tipo de *embedding* posicional puede afectar al rendimiento del LLM, ya que determina la cantidad y el tipo de información posicional disponible para el modelo durante el entrenamiento.

⁵⁴Devlin (2018), Liu (2022).

¹Vaswani (2017).

²Shaw (2018).

³Su (2021).

- El modelo no autorregresivo⁵⁵, usado en modelos como Gemini, en los que no se obtiene la respuesta secuencialmente palabra a palabra, sino que se transforma y refina en conjunto.
- El modelado de lenguaje enmascarado⁵⁶, popularizado por modelos como BERT, que consiste en enmascarar aleatoriamente algunas palabras en el texto de entrada y entrenar el modelo para predecir estas palabras enmascaradas basándose en el contexto circundante. Esta técnica permite un aprendizaje bidireccional y una mejor comprensión del contexto. Algunas arquitecturas de LLM (p. ej., *transformers* bidireccionales) usan esta técnica.
- El modelado secuencia a secuencia⁵⁷ (p. ej., seq2seq⁵⁸), en el que el modelo se entrena para generar secuencias de texto en función de otras secuencias de entrada. Es usado en modelos como T5, BART o ProphetNET.
- El preentrenamiento contrastivo⁵⁹, utilizado en modelos como CLIP y ALIGN⁶⁰, que implica entrenar el modelo para identificar pares de texto e imagen que están semánticamente relacionados, lo que le permite aprender representaciones multimodales y transferir conocimiento entre diferentes modalidades⁶¹.

El preentrenamiento de LLM es un proceso computacionalmente intensivo que requiere enormes cantidades de datos, tiempo y recursos de *hardware*. Los modelos más grandes pueden tener en torno a 1 billón (10^{12}) de parámetros y requerir miles de GPU de gama alta durante semanas o meses de entrenamiento. Esto hace que el preentrenamiento sea extremadamente costoso y solo esté al alcance de unas pocas empresas y organizaciones en el mundo con los recursos necesarios.

Cuantización

Durante el entrenamiento de los LLM, se ajustan los pesos de las neuronas para realizar predicciones más precisas. Estos pesos se almacenan típicamente como números de alta precisión, lo que puede resultar en modelos de gran tamaño y computacionalmente costosos.

La *cuantización* postentrenamiento es una técnica⁶² que permite reducir la precisión de los parámetros del modelo sin afectar significativamente el rendimiento del modelo. Por ejemplo, redes neuronales que almacenen sus parámetros usando números de coma flotante de 32 bits pueden pasar a usar únicamente 16 bits u 8 bits dependiendo del tipo de *cuantización*. Esto resulta en modelos más pequeños y rápidos, ya que requieren menos memoria y pueden realizar operaciones más eficientemente con el hardware adecuado.

Recientemente, ha surgido la tendencia de desarrollar modelos de menor tamaño (*small language models*, SLM), o incluso los llamados *"tiny LLM"*⁶³, modelos que mantienen un alto rendimiento a pesar de su tamaño mucho más reducido. Estos modelos compactos se logran mediante una combinación de técnicas, entre ellas la *cuantización* postentrenamiento.

Mediante la aplicación hábil de estas técnicas, los SLM y los tiny LLM están logrando en algunos casos un rendimiento comparable al de modelos mucho mayores⁶⁴, lo que los hace atractivos para aplicaciones con restricciones de recursos computacionales o de memoria.

⁵⁵Xu (2021).

⁵⁶Devlin (2019), Sinha (2021).

⁵⁷Lee (2022).

⁵⁸Sutskever (2014).

⁵⁹Zeng (2023).

⁶⁰Jia (2021).

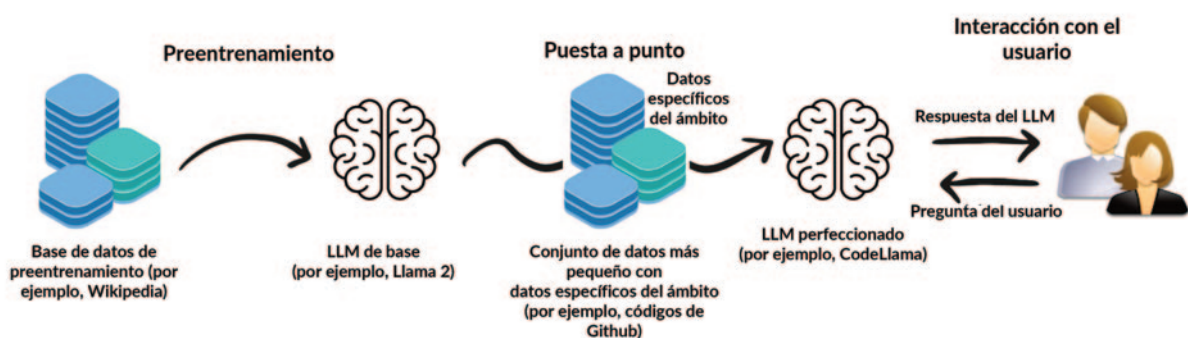
⁶¹Cui (2022).

⁶²Li (2024).

⁶³Tian (2024).

⁶⁴Fu (2024).

Fig. 6. Fine-tuning de un LLM.



El *fine-tuning* es el proceso de adaptar un LLM preentrenado a una tarea específica utilizando un conjunto de datos más pequeño. Esta técnica permite aprovechar el conocimiento general adquirido durante el preentrenamiento y especializarlo para obtener un alto rendimiento en la tarea objetivo.

El objetivo principal del *fine-tuning* (Fig. 6) es adaptar un LLM preentrenado a una tarea concreta, como la clasificación de sentimientos, la respuesta a preguntas, la traducción automática o la generación de resúmenes. Durante este proceso, el modelo aprende a utilizar su conocimiento general del lenguaje y aplicarlo de manera efectiva al dominio y los requisitos específicos de la tarea en cuestión. Los LLM disponibles en el mercado, sean propietarios o de código abierto, típicamente están preentrenados (y por tanto son de propósito general), pero no han recibido *fine-tuning*, que los adaptaría a un propósito específico.

El *fine-tuning* ofrece varios beneficios significativos:

- ▶ **Aprovecha el conocimiento previo:** al partir de un modelo preentrenado, el *fine-tuning* permite aprovechar el vasto conocimiento general del lenguaje adquirido durante el preentrenamiento, lo que acelera el aprendizaje y mejora el rendimiento en la tarea específica.
- ▶ **Requiere menos datos y recursos:** en comparación con el entrenamiento desde cero, el *fine-tuning* necesita mucha menos cantidad de datos etiquetados y recursos computacionales, lo que lo hace más accesible y económico para una amplia gama de organizaciones y aplicaciones.
- ▶ **Permite la especialización:** el *fine-tuning* permite adaptar los LLM a dominios y tareas concretas, lo que resulta en modelos altamente especializados y efectivos para aplicaciones específicas.
- ▶ **Facilita la transferencia de aprendizaje:** los modelos *fine-tuned* pueden recibir un *fine-tuning* adicional para tareas relacionadas, lo que permite la transferencia de aprendizaje y la creación de modelos aún más especializados con relativamente pocos datos adicionales.

A pesar de sus beneficios, el *fine-tuning* también presenta algunos desafíos:

- ▶ **Sobre-especialización**⁶⁵: si el modelo se somete a un *fine-tuning* en un conjunto de datos demasiado específico, puede perder parte de su capacidad de generalización y funcionar mal con datos desconocidos o ligeramente diferentes.

Entrenando LLM: funciones de pérdida

Los LLM, como otros modelos de aprendizaje profundo, aprenden ajustando sus parámetros para minimizar una función de pérdida. Esta función mide la diferencia entre las predicciones del modelo y los resultados esperados, guiando al modelo hacia un mejor rendimiento.

La elección de la función de pérdida depende del tipo de tarea para la que se esté entrenando el LLM. Por ejemplo, para un modelo que predice la siguiente palabra en una frase (modelado autorregresivo del lenguaje), una función común es la entropía cruzada. Esta función compara la distribución de probabilidad de las palabras predichas por el modelo con la distribución real observada en los datos de entrenamiento.

Matemáticamente, la función de pérdida de entropía cruzada para un modelo autorregresivo se puede expresar como una suma de los logaritmos negativos de las probabilidades asignadas a las palabras correctas en cada posición de la secuencia.

En concreto, dada una función de pérdida como la entropía cruzada, y una tipología de entrenamiento como el modelado autorregresivo del lenguaje, se puede definir la función de pérdida a minimizar como:

$$f_L(\varphi) = \sum_{i=1}^N -\log P(x_i | x_{1:i-1}, \varphi)$$

donde φ representa los parámetros del modelo, i se refiere al número de token en una secuencia determinada con N tokens, P es la probabilidad de predecir el token i en función de la secuencia x de tokens anteriores.

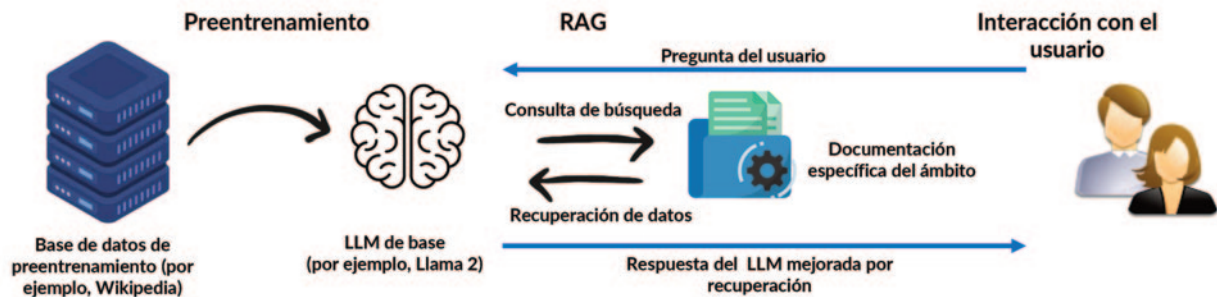
Durante el ajuste fino (*fine-tuning*) de los embeddings del modelo, se pueden utilizar funciones de pérdida especializadas para afinar las representaciones vectoriales de las palabras. Algunas opciones populares incluyen:

- ▶ Pérdida por similitud del coseno: ajusta los embeddings para que palabras similares tengan vectores más parecidos.
- ▶ Pérdida por error cuadrático medio: minimiza la diferencia cuadrática entre los embeddings predichos y los esperados.
- ▶ Pérdida por ranking de múltiples negativos: asocia los embeddings de palabras relacionadas de manera que estén más cerca entre sí que los de palabras no relacionadas.
- ▶ Pérdida por tripletes, de Matryoshka o contrastiva: variantes más avanzadas que consideran relaciones entre trios o grupos de embeddings.

La selección cuidadosa de la función de pérdida es crucial para entrenar LLM efectivos y eficientes que puedan capturar los matices del lenguaje natural.

⁶⁵Wang (2024).

Fig. 7. Funcionamiento del RAG.



- **Olvido catastrófico**⁶⁶: durante el *fine-tuning* es posible que un modelo olvide conocimientos críticos aprendidos previamente.
- **Inestabilidad**⁶⁷: el proceso de *fine-tuning* puede ser sensible a factores como la inicialización de pesos, los hiperparámetros y la selección de datos, lo que puede llevar a resultados inconsistentes o variaciones en el rendimiento.
- **Herencia de sesgos**⁶⁸: los modelos que han recibido *fine-tuning* pueden heredar y amplificar los sesgos presentes tanto en los datos de preentrenamiento como en los datos de *fine-tuning*, lo que requiere una cuidadosa consideración y mitigación.
- **Parameter efficient**⁷³ *fine-tuning* (PEFT): otros métodos de *fine-tuning* buscan aumentar su eficiencia y reducir el esfuerzo necesario para reentrenar el modelo. Por ejemplo, las técnicas basadas en LoRA⁷⁴ (*low-rank adaptation*) como QLoRA o LongLoRA⁷⁵, que permiten hacer *fine-tuning* del modelo sin la necesidad de modificar sus pesos y almacenan el conocimiento aprendido durante el proceso de *fine-tuning* en parámetros adicionales del modelo.

En muchos casos de uso de LLM, no es necesario emplear un *fine-tuning* para mejorar sus capacidades en un dominio específico. La generación aumentada de recuperación⁷⁶ (RAG) es una técnica que mejora el rendimiento del LLM a través del uso de fuentes de conocimiento externas al modelo.

Existen varios tipos de *fine-tuning* que deben seleccionarse en función de cuánto sea necesario modificar el modelo de partida para adecuarlo a una tarea en un dominio más específico. Los principales métodos son:

- **Fine-tuning supervisado**⁶⁹: requiere conjuntos de datos etiquetados de entrada y respuesta del LLM, que se usan para mejorar su respuesta ante tareas específicas. Un método popular de *fine-tuning* supervisado es el llamado *instruction-tuning*⁷⁰, que consiste en alinear las respuestas del modelo a lo esperado por sus usuarios a través de interacciones con el modelo.
- **Reinforcement learning**: métodos basados en aprendizaje por refuerzo que se centran en mejorar la calidad de la respuesta del LLM, en este caso con base en el *feedback* de usuario o modelos de recompensa (p. ej., optimización directa por preferencia⁷¹).
- **Fine-tuning no supervisado**⁷²: se trata de un método que no requiere conjuntos de datos etiquetados, sino que se basa en reentrenar el modelo con las mismas metodologías usadas durante el preentrenamiento (p. ej., predecir el siguiente token).

Las técnicas de RAG (Fig. 7, funcionan buscando documentos dentro de una base de datos que se parezcan o refieran al *prompt* de entrada. Esta búsqueda y sus resultados se añaden a la generación de la respuesta del LLM para enriquecerla, proporcionando un contexto específico.

⁶⁶Luo (2024).

⁶⁷Zhang (2024).

⁶⁸Zhang (2024).

⁶⁹Ovadia (2024).

⁷⁰Zhang (2023).

⁷¹Rafailov (2023).

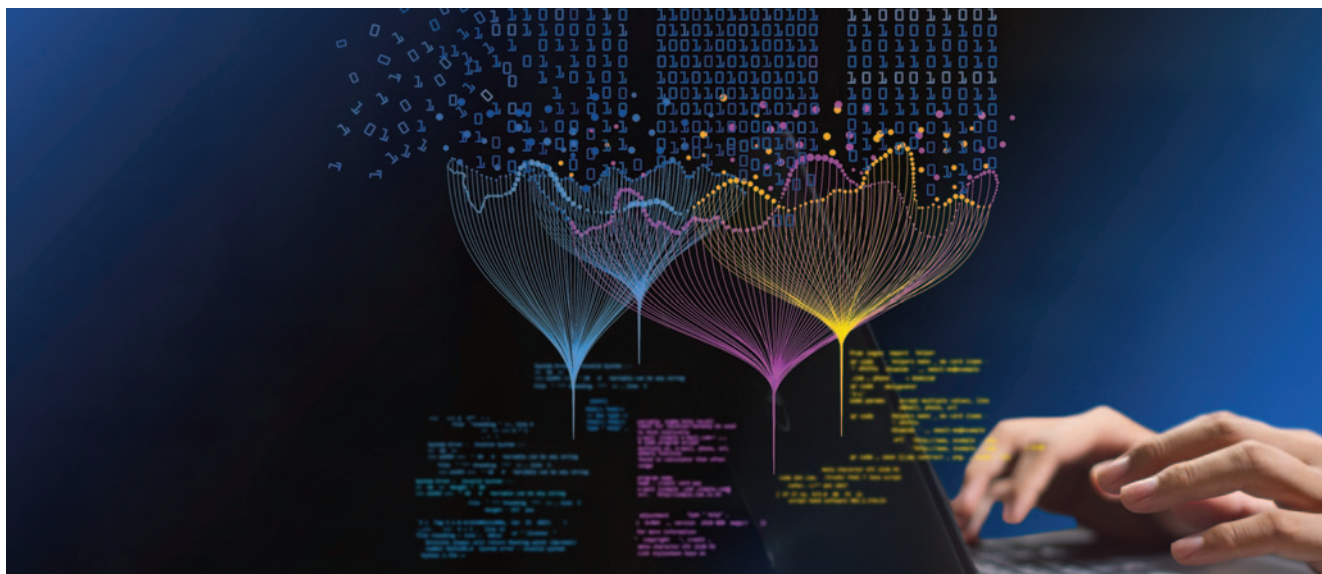
⁷²Zhou (2023).

⁷³Xu (2023).

⁷⁴Dettmers (2023).

⁷⁵Chen (2023).

⁷⁶Lewis (2020) y Neelakantan (2022).



Implementación y uso

Una vez entrenado y validado, el LLM debe desplegarse en un entorno de producción para su uso en aplicaciones reales. Esto conlleva la integración del modelo en sistemas y flujos de trabajo existentes, así como la creación de interfaces y APIs para interactuar con él.

Este proceso implica varios aspectos clave, entre los que se incluyen aspectos de integración y de monitoreo.

Integración en sistemas y flujos de trabajo

- **Infraestructura**⁷⁷: los LLM suelen ser modelos grandes y computacionalmente intensivos, por lo que requieren una infraestructura robusta para su implementación. Esto puede implicar el uso de hardware especializado, como GPU o TPU, y plataformas de computación en la nube optimizadas para realizar eficientemente el proceso de inferencia.
- **Interfaces y API**⁷⁸: para facilitar el uso del LLM en aplicaciones y servicios, es necesario desarrollar interfaces y API que permitan a otros sistemas interactuar con el modelo de manera eficiente y segura. Esto puede incluir *endpoints*, bibliotecas de cliente en varios lenguajes de programación e interfaces gráficas de usuario para usuarios no técnicos.
- **Integración con otros componentes**: en muchos casos, los LLM forman parte de un sistema más amplio que incluye otros componentes, como bases de datos, servicios de procesamiento de lenguaje natural y aplicaciones de usuario final. La integración fluida y eficiente del LLM con estos componentes es crucial para garantizar un rendimiento y una experiencia de usuario óptimos.

Monitoreo y mantenimiento

- **Monitoreo del rendimiento**⁷⁹: una vez implementado, es esencial monitorear de cerca el rendimiento del LLM en condiciones reales. Esto implica realizar un seguimiento de métricas como la latencia, el rendimiento, la precisión y el uso de recursos, y establecer umbrales en el consumo de recursos y costes, y alertas para detectar y abordar cualquier degradación o anomalía.
- **Actualización y reentrenamiento**⁸⁰: a medida que se dispone de nuevos datos o se identifican áreas de mejora, puede ser necesario actualizar o reentrenar el LLM. Esto requiere un proceso bien definido para recopilar y preparar nuevos datos, realizar un *fine-tuning*, y desplegar la versión actualizada del modelo sin interrupciones en el servicio.
- **Gestión de versiones**⁸¹: con las actualizaciones y mejoras continuas, es importante mantener un control de versiones riguroso del LLM y sus componentes asociados. Esto facilita la reproducibilidad, el *debugging* y la capacidad de revertir a versiones anteriores si es necesario.

Como se puede apreciar, el desarrollo y despliegue de LLM es un proceso complejo y multifacético que requiere una cuidadosa consideración de múltiples aspectos, desde la selección y preparación de los datos hasta la implementación y el uso responsable del modelo. Una comprensión profunda de los componentes clave, como el preentrenamiento, el *fine-tuning* y los *embeddings*, así como la conciencia de los desafíos y riesgos asociados, es esencial para aprovechar todo el potencial de los LLM de manera ética, sostenible y rentable, así como alineada con los objetivos de cada organización.

⁷⁷Wan (2024).

⁷⁸Abhyankar (2024).

⁷⁹Goyal (2024).

⁸⁰Lester (2021).

⁸¹Banerjee (2023).

Arquitectura de los LLM

La arquitectura de los LLM se refiere a la estructura y organización de las redes neuronales que componen estos modelos. La elección de la arquitectura y sus componentes tiene un impacto significativo en el rendimiento, la eficiencia y las capacidades del LLM. Esta sección explorará las principales arquitecturas utilizadas en los LLM y sus características, ventajas y limitaciones.

Transformers: el estado del arte en LLM

Los *transformers*, introducidos⁸² en 2017, se han convertido en la arquitectura dominante para los LLM. A diferencia de las arquitecturas anteriores basadas en redes neuronales recurrentes (RNN) o redes neuronales convolucionales (CNN), los *transformers* se basan únicamente en mecanismos de atención para procesar y generar secuencias de texto (Fig. 8).

La arquitectura del *transformer* consta de dos componentes principales: el codificador (*encoder*) y el decodificador (*decoder*), y existen *transformers* con solo codificador, solo decodificador o con ambos componentes. El codificador procesa la secuencia de entrada y genera una representación contextual para cada token, mientras que el decodificador genera la secuencia de salida a partir de la representación del codificador y las predicciones anteriores.

La clave de los *transformers* es el mecanismo de atención, que permite al modelo poner atención en diferentes partes de la secuencia de entrada (atención del codificador) y en las predicciones anteriores (atención del decodificador) para generar la siguiente palabra o token. Esto permite capturar dependencias a largo plazo y generar secuencias coherentes.

Los *transformers* también introducen el concepto de atención multi-cabezal (*multi-head attention*), donde múltiples mecanismos de atención operan en paralelo, lo que permite al modelo capturar diferentes tipos de relaciones y patrones en los datos.

La arquitectura del *transformer* ha demostrado un rendimiento excepcional en una amplia gama de tareas de procesamiento del lenguaje natural, y ha sido adoptada por la mayoría de los LLM considerados de referencia.

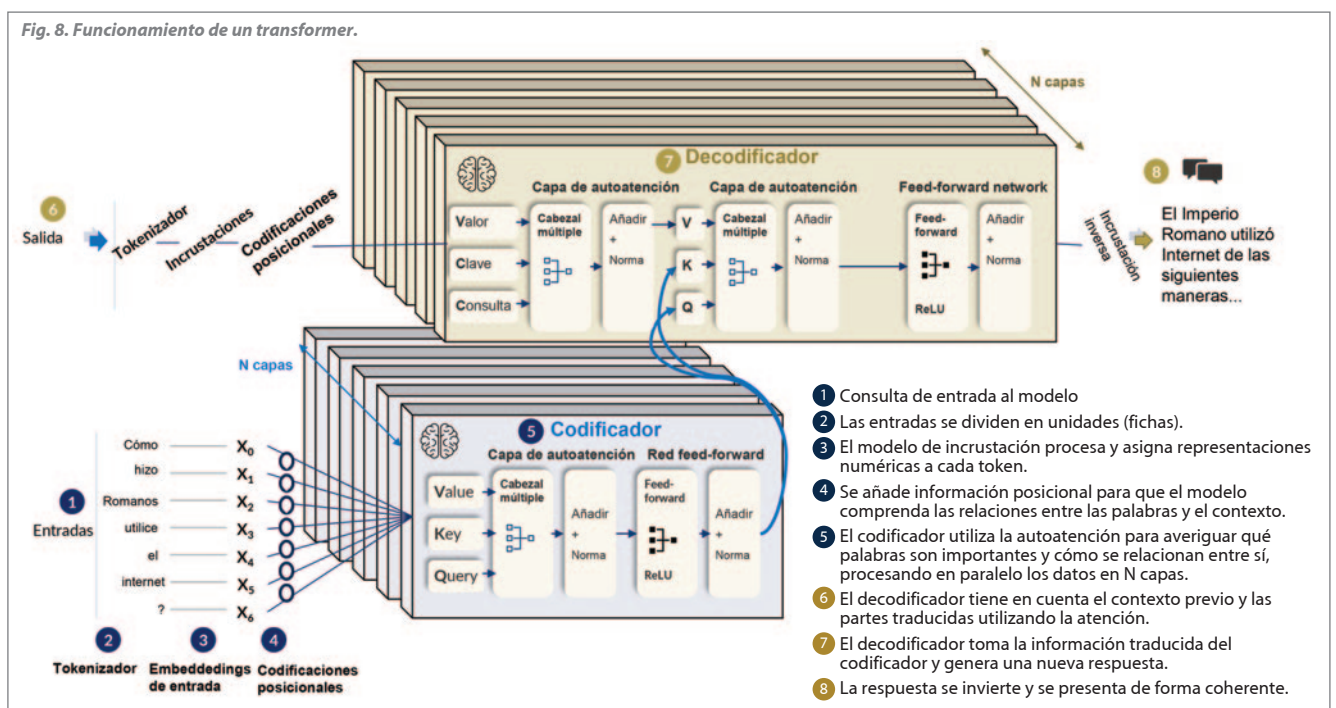
Variantes y extensiones de los transformers

Desde la introducción de los *transformers*, se han propuesto numerosas variantes y extensiones para mejorar su eficiencia, escalabilidad y capacidad de modelado.

- Una variante popular es el *transformer* bidireccional, que permite al modelo atender tanto al contexto izquierdo como al derecho de cada token. Esto se logra mediante el uso de un objetivo de preentrenamiento de modelado de lenguaje enmascarado (MLM), donde algunos tokens se enmascaran aleatoriamente y el modelo debe predecirlos basándose en el contexto circundante.
- Otra variante es el *transformer* generativo, como GPT, que utiliza un enfoque de modelado de lenguaje unidireccional. Esto permite generar texto de manera eficiente y coherente, ya que el modelo solo puede atender al contexto izquierdo de cada token.
- También se han propuesto extensiones para hacer que los *Transformers* sean más eficientes y escalables, como el *transformer* disperso (*sparse transformer*), que utiliza

⁸²Vaswani (2017).

Fig. 8. Funcionamiento de un transformer.



Prompt engineering en los LLM: principios y mejores prácticas

Prompt engineering se refiere al proceso de diseñar y optimizar los *prompts* (entradas de texto) para obtener los mejores resultados posibles de los LLM. Esta disciplina emergente contiene una serie de principios y mejores prácticas que permiten aprovechar al máximo las capacidades de estos modelos. Entre ellos:

- ▶ Ser claro y específico: las instrucciones dadas al modelo deben indicar de forma explícita el formato, la longitud y el nivel de detalle esperado en la respuesta. Por ejemplo, en lugar de simplemente pedir "Analiza la situación financiera de la empresa X", es mejor dar una instrucción como "Escribe un informe de 1000 palabras sobre la situación financiera de la empresa X, cubriendo su rentabilidad, liquidez, solvencia y perspectivas futuras".
- ▶ Descomponer tareas complejas: conviene dividir los problemas en subtareas más manejables para los LLM. Por ejemplo, en lugar de pedir "Desarrolla un plan estratégico para la empresa Y", se pueden solicitar subtareas como "Realiza un análisis DAFO de la empresa Y", "Define los objetivos estratégicos clave para Y", "Propón iniciativas para alcanzar cada objetivo", etc.
- ▶ Proporcionar ejemplos ilustrativos (*few-shot learning*): unos pocos ejemplos bien elegidos pueden ayudar enormemente a comunicar la tarea deseada. Supongamos que se quiere generar propuestas de valor para productos; se podrían dar dos ejemplos: "Nuestro software CRM permite a los equipos de ventas cerrar acuerdos un 50% más rápido" y "Nuestra app de bienestar ayuda a los empleados a reducir el estrés y aumentar su productividad en un 25%".
- ▶ Solicitar razonamiento paso a paso: instruir al LLM para que verbalice su proceso de pensamiento a menudo conduce a resultados más robustos. Esto es especialmente útil para tareas de análisis o solución de problemas en el ámbito empresarial. Por ejemplo: "Describe paso a paso cómo calcularías el ROI de este proyecto de inversión".
- ▶ Solicitar las referencias empleadas: instruir al LLM para que indique en su proceso de razonamiento las referencias a documentos que ha empleado, incluyendo citas del texto original al que tiene acceso.
- ▶ Pedir al LLM que adopte un personaje: antes de la tarea principal, se puede primero instruir al modelo para que adopte un determinado rol, tono o estilo. Por ejemplo: "Actúa como un experto analista financiero y proporciona una valoración objetiva de la empresa X". Esto ayuda a orientar su comportamiento.

- ▶ Aprovechar conocimiento externo: proporcionar información adicional permite complementar la base de conocimientos del LLM. Por ejemplo, para responder preguntas sobre una industria específica, se podrían primero recuperar informes sectoriales relevantes y pasarlos al modelo.
- ▶ Iterar y refinar sistemáticamente: la evaluación continua del rendimiento del modelo permite identificar áreas de mejora y ajustar los *prompts* en consecuencia. Métricas cuantitativas y juicios cualitativos de expertos del dominio pueden guiar este proceso iterativo.

Mediante la aplicación de estos principios de *prompt engineering*, se demuestra estadísticamente que los LLM proporcionan un resultado más preciso y fiable.

Considerando todo esto, un mal *prompt* para que un LLM escriba una columna sobre *prompt engineering* sería: "Escribe un artículo sobre *prompt engineering*".

Y un buen *prompt* para escribir esa columna sería:

"Actúa como un experto en inteligencia artificial y escribe una columna de divulgación de 600 palabras sobre los principios clave del *prompt engineering* para obtener los mejores resultados de los LLM. Estructura la columna con una introducción breve y atractiva, 4-5 párrafos cubriendo los puntos principales (ser específico, descomponer tareas, dar ejemplos...), y una conclusión con los beneficios de aplicar estas técnicas. Utiliza un tono divulgativo pero riguroso, apto para un público empresarial. Incluye ejemplos concretos para ilustrar las ideas".

Fuentes: Guía de *prompt engineering* de OpenAI¹, soporte de Anthropic Claude Opus y elaboración propia.

¹OpenAI (2024).

atención dispersa para reducir la complejidad computacional, y el *transformer* comprimido (*compressed transformer*), que utiliza técnicas de compresión para reducir el tamaño del modelo.

Comparación con arquitecturas anteriores

Antes de los *transformers*, las arquitecturas dominantes para el modelado de secuencias eran las redes neuronales recurrentes (RNN), como *Long Short-Term Memory* (LSTM) y *Gated Recurrent Unit* (GRU), y las redes neuronales convolucionales (CNN).

- ▶ Las RNN pueden capturar dependencias a largo plazo en las secuencias, pero sufren de problemas como el desvanecimiento del gradiente (*vanishing gradient*) y la dificultad para paralelizar el entrenamiento. Además, las RNN tienen dificultades para capturar dependencias muy largas debido a su naturaleza secuencial y al uso de recurrencias de alcance constante.
- ▶ Las CNN pueden capturar patrones locales en las secuencias y son eficientes en términos de computación, pero tienen dificultades para modelar dependencias a largo plazo y requieren un tamaño de contexto fijo.

En comparación, los *transformers* superan estas limitaciones al utilizar mecanismos de atención que pueden capturar dependencias a largo plazo de manera eficiente y paralela. Además, los *transformers* son más flexibles en términos de manejo de secuencias de longitud variable y pueden ser preentrenados en grandes cantidades de datos no etiquetados.

La arquitectura de los *transformers* ha revolucionado el campo de los LLM y ha permitido avances significativos en una amplia gama de tareas de procesamiento del lenguaje natural. Sin embargo, aún quedan desafíos por abordar, como la

escalabilidad, la interpretabilidad y la eficiencia de estos modelos. A medida que la investigación avanza, es probable que aparezcan nuevas arquitecturas y técnicas que superen estas limitaciones y lleven a los LLM a nuevas cotas de rendimiento y capacidad.

LLMOps

MLOps (*Machine Learning Operations*) es una metodología y conjunto de prácticas diseñadas para gestionar el ciclo de vida completo de los modelos de *machine learning*, desde el desarrollo y entrenamiento hasta la implementación y mantenimiento en producción.

En los últimos años, ha surgido una adaptación de la metodología MLOps específicamente orientada a los LLM, conocida como LLMOps (*Large Language Model Operations*). Esta disciplina se centra en gestionar de manera eficiente el ciclo de vida completo de los LLM, abarcando desde su desarrollo y entrenamiento hasta su implementación y mantenimiento en entornos de producción.

LLMOps integra los procesos tradicionales de desarrollo de software con herramientas y técnicas diseñadas para abordar los desafíos únicos que presentan los modelos de lenguaje de gran escala. Algunos de estos desafíos incluyen:

- ▶ **Gestión de grandes volúmenes de datos:** los LLM requieren cantidades masivas de datos de entrenamiento, lo que implica la necesidad de infraestructuras de almacenamiento y procesamiento escalables y eficientes.
- ▶ **Escalado de recursos computacionales:** el entrenamiento y la inferencia de LLM demandan enormes recursos de cómputo, haciendo necesario el uso de técnicas de





paralelización y distribución, así como la optimización del uso de hardware especializado como GPU y TPU.

- ▶ **Monitorización y mantenimiento:** una vez desplegados en producción, los LLM deben ser monitorizados de cerca para detectar y corregir problemas de rendimiento, sesgos, riesgos como las alucinaciones, y la degradación del modelo a lo largo del tiempo.
- ▶ **Versionado y reproducibilidad:** dado el tamaño y complejidad de los LLM, es crucial mantener un riguroso control de versiones y maximizar la reproducibilidad de los experimentos y resultados.

Para abordar estos desafíos, LLMOps se apoya en una serie de herramientas y *frameworks* específicos, como MLFlow⁸³, CometML⁸⁴ y Weights & Biases⁸⁵. Estas plataformas ofrecen funcionalidades para el seguimiento de experimentos, gestión de modelos, monitorización del rendimiento y colaboración entre equipos.

Además, LLMOps promueve prácticas como la automatización de procesos, el *testing* continuo, la documentación exhaustiva y la gobernanza de modelos. Esto permite no solo mejorar la eficiencia y calidad del desarrollo de los LLM, sino también garantizar su uso ético y responsable.

Retos

El desarrollo y despliegue de LLM presenta una serie de retos significativos que deben abordarse para garantizar su uso responsable, ético y seguro. Esta sección explorará varios de los principales desafíos para las organizaciones en su despliegue y uso de los LLM.

Sesgos, alucinaciones y confiabilidad

Uno de los mayores retos de los LLM es la presencia de sesgos y alucinaciones en sus resultados y predicciones. Los sesgos pueden surgir de varias fuentes, como la parcialidad en los datos de entrenamiento, las limitaciones de las arquitecturas de los modelos o los prejuicios humanos implícitos en las tareas de anotación y evaluación. Por otro lado, las alucinaciones se refieren a la generación de información o contenido que parece plausible pero que no se basa en hechos reales o en el conocimiento adquirido durante el entrenamiento.

Los sesgos en los LLM pueden manifestarse de diversas maneras, como la perpetuación de estereotipos de género, raza o edad, la discriminación en tareas de clasificación o la generación de contenido ofensivo o inapropiado. Estos sesgos pueden tener consecuencias graves, especialmente cuando los LLM se utilizan en aplicaciones sensibles como la toma de decisiones en el ámbito legal, financiero o médico. Por su parte, las alucinaciones pueden llevar a la difusión de información errónea o engañosa, lo que puede tener un impacto negativo en la confianza de los usuarios y en la credibilidad de las aplicaciones basadas en LLM.

Para abordar el reto de los sesgos, es necesario desarrollar técnicas robustas para detectar, medir y mitigar su presencia en los LLM. Esto implica la creación de conjuntos de datos de evaluación específicos para sesgos, el uso de métricas de equidad y la aplicación de técnicas de eliminación de sesgos (*debiasing*) tanto en el preentrenamiento como en el *fine-tuning*. Además, es crucial establecer procesos de auditoría y monitoreo continuo para garantizar que los LLM sigan siendo imparciales a lo largo del tiempo.

⁸³Zaharia (2018).

⁸⁴CometML: <https://www.comet.com/>

⁸⁵Weights and biases: <https://wandb.ai/site>



Para abordar las alucinaciones en los LLM, se están desarrollando diversos métodos centrados en la mejora de los datos de entrenamiento, la aplicación de técnicas de regularización robustas y el uso de retroalimentación humana para ajustar las respuestas del modelo. Además, se están investigando cambios arquitectónicos en los modelos para hacerlos inherentemente menos propensos a alucinar. Los métodos de generación de texto y el contexto de entrada también pueden ser optimizados para reducir las alucinaciones. La supervisión humana y la evaluación rigurosa son esenciales para detectar y corregir información inexacta. Asimismo, el desarrollo de herramientas específicas, como modelos de evaluación de alucinaciones y técnicas de ofuscación, puede contribuir a mejorar la precisión de los LLM.

Explicabilidad y accountability

Otro gran desafío de los LLM es su opacidad y falta de explicabilidad. Debido a su complejidad y a la naturaleza de sus arquitecturas, es difícil entender cómo estos modelos llegan a sus resultados.

Esta falta de transparencia plantea problemas de *accountability*, especialmente cuando los LLM se utilizan en contextos de alta sensibilidad, donde las decisiones tienen un impacto significativo en las personas (p. ej., uso de LLM en medicina, investigación farmacéutica, infraestructura crítica o acceso al mercado laboral). Sin una comprensión clara de cómo funcionan estos modelos, es difícil determinar la responsabilidad en caso de errores o comportamientos no deseados.

Para abordar este reto, es necesario desarrollar técnicas y herramientas que permitan una mayor interpretabilidad y explicabilidad de los LLM. Esto incluye métodos para visualizar y analizar los mecanismos internos de atención, técnicas de atribución para identificar las partes más relevantes de la entrada, y enfoques para generar explicaciones en lenguaje natural de las predicciones del modelo.

Además, es importante establecer marcos de *accountability* claros que definan las responsabilidades de los desarrolladores, implementadores y usuarios de los LLM, como en Europa propone el AI Act. Esto puede implicar la creación de estándares y directrices para el desarrollo ético de los LLM, mecanismos de supervisión y auditoría externa, y canales para que las partes interesadas planteen inquietudes.

Confidencialidad y protección de la información

Los LLM a menudo se entrenan con grandes cantidades de datos que pueden contener información personal, sensible o confidencial. Además, cuando se despliegan en aplicaciones del mundo real, estos modelos pueden estar expuestos a entradas de usuario que también pueden incluir datos privados.

Esto plantea importantes desafíos de privacidad y seguridad, ya que los LLM pueden memorizar y reproducir información confidencial de sus datos de entrenamiento, o ser vulnerables a ataques que intenten extraer datos privados a través de consultas cuidadosamente diseñadas.

Para abordar este reto, es necesario desarrollar técnicas de preservación de la privacidad en el entrenamiento y despliegue de LLM (p. ej., Digger⁸⁶ para detectar información protegida, el uso de datos ficticios⁸⁷ durante el entrenamiento para detectar el material con copyright).

Además, es crucial establecer protocolos robustos de seguridad y control de acceso para proteger los LLM y sus datos asociados de accesos no autorizados o usos maliciosos. Esto puede implicar el uso de técnicas de autenticación y autorización, monitoreo de seguridad y detección de anomalías.

⁸⁶Li (2024).

⁸⁷Meeus (2024).

Consumo racional de los recursos

El entrenamiento y despliegue de LLM requiere cantidades masivas de recursos computacionales, almacenamiento y energía. Con modelos que alcanzan los cientos de miles de millones o incluso los billones de parámetros, el coste financiero y ambiental de desarrollar y operar estos sistemas puede ser muy significativo⁸⁸.

Este alto consumo de recursos plantea desafíos de eficiencia, escalabilidad y sostenibilidad. A medida que la demanda de LLM más grandes y potentes sigue creciendo, es necesario encontrar formas de optimizar su rendimiento y reducir su huella de recursos.

Para abordar este reto, se están explorando varias direcciones de investigación. Una de ellas es el diseño de arquitecturas de modelos más eficientes, como el uso de mecanismos de atención dispersa o técnicas de compresión que reducen el tamaño y la complejidad computacional de los LLM sin comprometer significativamente su rendimiento.

También se está investigando en la mejora de las técnicas de preentrenamiento continuo⁸⁹ y *fine-tuning* continuo⁹⁰, que buscan integrar la capacidad de usar información de diversos dominios sin la necesidad de depender de un reentrenamiento exhaustivo y costoso con nuevos datos específicos. Asimismo, se está avanzando en el uso de sistemas innovadores y el diseño de algoritmos verdes de IA, que permiten hacer frente a los costes computacionales y ambientales asociados a la IA (p. ej., el sistema GreenLightningAI, de Qsimov Quantum Computing⁹¹, desarrolla el reentrenamiento incremental y proporciona una interpretabilidad directa).

Otra dirección es el desarrollo de infraestructuras y plataformas de computación más sostenibles, como el uso de hardware especializado de bajo consumo, sistemas de enfriamiento más eficientes y fuentes de energía renovables para alimentar los centros de datos donde se entrenan y despliegan los LLM.

Además, es importante promover prácticas de uso racional y compartido de los recursos, como la reutilización y adaptación de modelos preentrenados en lugar de entrenar nuevos modelos desde cero para cada tarea, y el intercambio de recursos y conocimientos entre organizaciones y comunidades de investigación.

Otros desafíos

De entre los muchos retos adicionales a los que se enfrentan las organizaciones en el desarrollo, implementación y uso de los LLM, por su importancia cabe mencionar brevemente:

- **Dependencia y lock-in:** las organizaciones que dependen de LLM proporcionados por terceros pueden enfrentarse a riesgos de dependencia y *lock-in*, especialmente si los modelos se basan en datos o infraestructura propietaria. Es importante considerar estrategias de diversificación y planes de contingencia.

- **Riesgos de seguridad y uso malicioso**⁹²: los LLM pueden ser vulnerables a ataques adversarios, como la inyección de datos envenenados o la ingeniería inversa. Además, pueden utilizarse de forma maliciosa para generar desinformación, *spam* o contenido engañoso. Es esencial establecer medidas de seguridad robustas y diseñar los modelos con salvaguardias contra el uso indebido.

- **Cuestiones de propiedad intelectual y licencias:** el uso de LLM plantea preguntas sobre la propiedad intelectual y las licencias de los datos de entrenamiento, los modelos y los resultados generados. Adicionalmente, existe un riesgo de robo de información o datos personales de usuarios que lancen consultas a LLM desplegados en nubes de terceros. Es necesario cumplir con la regulación y con los marcos éticos para equilibrar los derechos de los creadores, los usuarios y el interés público, y, en el caso de las organizaciones, para evitar riesgos legales y de cumplimiento.

- **Escalabilidad de la arquitectura de los LLM**⁹³: un desafío adicional es la escalabilidad de los *transformers* a medida que aumenta el tamaño de las secuencias y los modelos. Los mecanismos de atención tienen una complejidad cuadrática con respecto a la longitud de la secuencia, lo que limita su aplicación a secuencias muy largas.

⁸⁸iDanae 1T24 (2024).

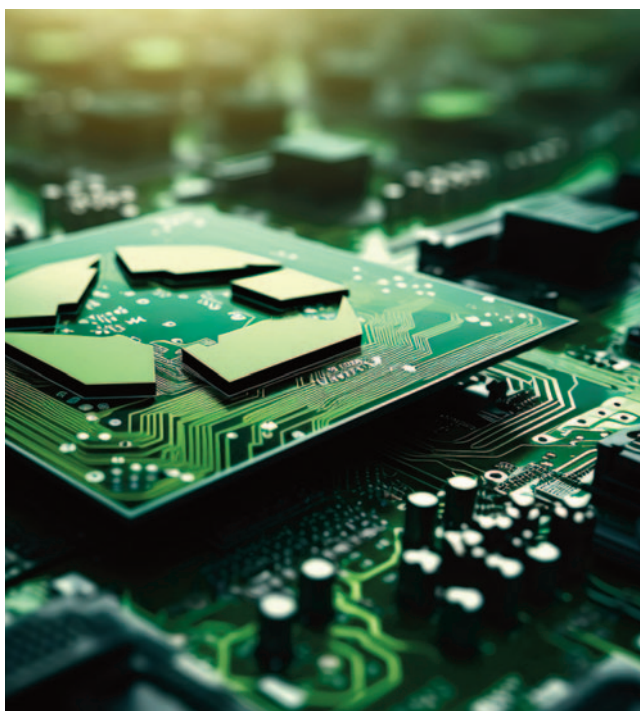
⁸⁹Yildiz (2024).

⁹⁰Mehta (2023).

⁹¹iDanae 1T24 (2024).

⁹²Pankajakshan (2024).

⁹³Rae (2021).



Marco de validación de los LLM

“Las consecuencias de que la IA vaya mal son graves, por lo que debemos ser proactivos en lugar de reactivos”.

Elon Musk⁹⁴



Marco

Los modelos de lenguaje de gran escala (LLM) ofrecen un gran potencial para transformar diversos sectores y aplicaciones, pero también conllevan riesgos significativos que deben abordarse. Estos riesgos incluyen la generación de información errónea o alucinaciones, la perpetuación de sesgos, la dificultad para olvidar la información aprendida, preocupaciones éticas y de equidad, problemas de privacidad por uso indebido, dificultades en la interpretación de los resultados, y la potencial creación de contenido malicioso, entre otros.

Dado el impacto potencial de estos riesgos, es necesario validar exhaustivamente los LLM antes de su despliegue en entornos de producción. De hecho, la validación de los LLM no es solo una buena práctica, sino también un requisito regulatorio en muchas jurisdicciones. En Europa, la propuesta de AI Act exige una evaluación y mitigación de los riesgos de los sistemas de IA⁹⁵, mientras que, en Estados Unidos, el marco de gestión de riesgos de IA del NIST⁹⁶ y el AI Bill of Rights destacan la importancia de comprender y abordar los riesgos inherentes a estos sistemas.

La validación de los LLM puede partir de los principios establecidos en la disciplina de riesgo de modelo, que se centra⁹⁷ en evaluar y mitigar los riesgos derivados de errores, deficiente implementación o mal uso de los modelos. Sin embargo, en el caso de la IA, y particularmente de los LLM, es necesario adoptar una perspectiva más amplia que abarque los otros riesgos que comportan. Un enfoque integral de validación es esencial para garantizar un despliegue seguro y responsable de los LLM.

Este enfoque holístico se plasma en un marco de validación multidimensional para los LLM, que cubre aspectos clave (Fig. 9) como el riesgo de modelo, la gestión de datos y privacidad, la ciberseguridad, los riesgos legales y de cumplimiento normativo, los riesgos operativos y tecnológicos, la ética y la reputación, y el riesgo de proveedor, entre otros. Al abordar

todos estos aspectos de manera sistemática, las organizaciones pueden identificar y mitigar de manera proactiva los riesgos asociados con los LLM, sentando las bases para aprovechar su potencial de manera segura y responsable.

En los LLM, esta evaluación de riesgos se puede anclar en las siguientes dimensiones usadas en la disciplina de riesgo de modelo, adaptando los tests en función de la naturaleza y el uso del LLM:

- **Datos de entrada:** comprensión del texto⁹⁸, calidad del dato⁹⁹.
- **Solidez conceptual y diseño del modelo:** selección del modelo y sus componentes (p. ej., metodologías de *fine-tuning*, conexiones a bases de datos, RAG¹⁰⁰), y comparación con otros modelos¹⁰¹.

⁹⁴ Elon Musk (n. 1971), CEO de X, SpaceX, Tesla. Empresario sudafricano-estadounidense, conocido por fundar o cofundar empresas como Tesla, SpaceX y PayPal, dueño de X (anteriormente Twitter), red social que tiene su propio LLM, llamado Grok.

⁹⁵ European Parliament (2024) AI Act Art. 9: "Se establecerá, aplicará, documentará y mantendrá un sistema de gestión de riesgos en relación con los sistemas de IA de alto riesgo. El sistema de gestión de riesgos [...] comprenderá [...] la estimación y evaluación de los riesgos que puedan surgir cuando el sistema de IA de alto riesgo se utilice de acuerdo con su finalidad prevista, y en condiciones de uso indebido razonablemente previsibles".

⁹⁶ NIST (2023): "La decisión de encargar o desplegar un sistema de IA debe basarse en una evaluación contextual de las características de fiabilidad y los riesgos, impactos, costes y beneficios relativos, y debe ser informada por un amplio conjunto de partes interesadas".

⁹⁷ Management Solutions (2014). Model Risk Management: Quantitative and Qualitative Aspects.

⁹⁸ Imperial et al. (2023).

⁹⁹ Wettig et al. (2024).

¹⁰⁰ RAG (Retrieval-Augmented Generation) es una técnica avanzada en la que un modelo de lenguaje busca información relevante de una fuente externa antes de generar texto. Esto enriquece las respuestas con conocimientos precisos y actuales, combinando inteligentemente la búsqueda de información y la generación de texto. Al integrar datos de fuentes externas, los modelos RAG, como los RAG-Token y RAG-Sequence propuestos (Lewis et al., 2020), ofrecen respuestas más informadas y coherentes, minimizando el riesgo de generar contenido inexacto o 'alucinaciones'. Este avance representa un paso significativo hacia modelos de inteligencia artificial más confiables y basados en evidencia real.

¹⁰¹ Khang (2024).

Fig. 9. Riesgos asociados a la IA y referencia regulatoria en el AI Act.



- **Evaluación del modelo y análisis de sus resultados:** privacidad y seguridad de los resultados¹⁰², precisión del modelo¹⁰³, consistencia¹⁰⁴, robustez¹⁰⁵, adaptabilidad¹⁰⁶, interpretabilidad (XAI)¹⁰⁷, ética, sesgos y equidad¹⁰⁸, toxicidad¹⁰⁹, comparación contra modelos *challenger*.
- **Implementación y uso:** revisión humana en el uso (incluyendo el monitoreo de usos indebidos), resolución de errores, escalabilidad y eficiencia, aceptación del usuario.
- **Gobernanza¹¹⁰ y ética¹¹¹:** marco de gobierno de la IA generativa, incluyendo los LLM.
- **Documentación¹¹²:** completitud de la documentación del modelo.
- **Cumplimiento regulatorio¹¹³:** evaluación de los requisitos regulatorios (p. ej., AI Act).

Para garantizar el uso efectivo y seguro de los modelos de lenguaje, es fundamental realizar una evaluación de riesgos que considere tanto el modelo en sí como su uso específico. Esto asegura que, independientemente de su origen (*in-house* o de un proveedor) o personalización (*fine-tuning*), el modelo funcione adecuadamente en su contexto de uso, cumpliendo con los estándares de seguridad, ética y regulación necesarios.

Técnicas de validación

Cuando una organización se plantea implementar un LLM para un caso de uso específico, puede ser beneficioso adoptar un enfoque integral que abarque las dimensiones clave del ciclo de vida del modelo: datos, diseño, evaluación, implementación y uso. Asimismo, de manera transversal, resulta necesario evaluar el cumplimiento de la normativa aplicable, como el AI Act en la Unión Europea.

En cada una de estas dimensiones, dos grupos de técnicas complementarias permiten realizar una validación más completa (Fig. 10):

- **Métricas de evaluación cuantitativas (tests):** se trata de pruebas cuantitativas estandarizadas que miden el desempeño del modelo en tareas específicas; *benchmarks* y métricas predefinidas para evaluar distintos aspectos del rendimiento del LLM después del preentrenamiento, o durante las etapas de *fine-tuning* o *instruction-tuning* (es decir, técnicas de aprendizaje por refuerzo), optimización, ingeniería de *prompts*, o recuperación y generación de información. Algunos ejemplos incluyen la precisión en la creación de resúmenes, la robustez ante ataques adversarios o la consistencia en la respuesta ante *prompts* similares.
- **Evaluación humana:** implica el juicio cualitativo por parte de expertos y usuarios finales; por ejemplo, la revisión de una muestra concreta de los *prompts* y las respuestas del LLM por un ser humano para identificar errores.

La validación de un uso específico de un LLM, por tanto, se lleva a cabo mediante una combinación de técnicas cuantitativas (tests) y cualitativas (evaluación humana). Para cada caso de uso concreto, es necesario diseñar un enfoque de validación a medida, que consistirá en una selección de algunas de estas técnicas.

¹⁰²Nasr (2023).

¹⁰³Liang (2023).

¹⁰⁴Elazar (2021).

¹⁰⁵Liu (2023).

¹⁰⁶Dun (2024).

¹⁰⁷Singh (2024).

¹⁰⁸NIST (2023), Oneto (2020) y Zhou (2021).

¹⁰⁹Shaikh (2023).

¹¹⁰Management Solutions (2014). Model Risk Management.

¹¹¹Oneto (2020).

¹¹²NIST (2023).

¹¹³European Parliament (2024). AI Act.

Fig. 10. Pruebas de evaluación de LLM.

Dimensiones	Aspectos validados	Descripción	Métricas de validación (ejemplos)	Evaluación humana (ejemplos)
1. Datos de entrada	1.1 Calidad de dato	Grado de calidad de la modelización o de los datos de aplicación	• Flesch-Kinkaid Grade	• Revisión caso a caso
2. Diseño del modelo	2.1 Diseño del modelo	Elección de modelos y metodología adecuadas	• Revisión de los elementos del LLM: RAG, filtros de entrada o salida, definición de <i>prompts</i> , <i>fine-tuning</i> , optimización, etc. • Comparación contra otros LLM	• Pruebas A/B
3. Evaluación del modelo	3.1 Privacidad y seguridad	Respeto de la confidencialidad y no regurgitación de información personal	• Data leakage • PII tests, K-anonymity	• Registros • <i>Hacking</i> ético
	3.2 Precisión	Corrección y pertinencia de las respuestas del modelo	• Q&A: SummaQA, Word error rate • Recuperación de información: SSA, nDCG • Resumen: ROUGE • Traducción: BLEU, Ruby, ROUGE-L • Otros: Sistemas de QA, nivel de <i>overrides</i> , nivel de alucinaciones, etc. • Benchmarks: XSUM, LogiQA, WikiData, etc.	• <i>Backtest</i> de forzajes • Revisión caso a caso
	3.3 Consistencia	Respuestas uniformes a consultas similares	• Cosine similarity • Jaccard similarity index	• Revisión caso a caso • Pruebas A/B
	3.4 Robustez	Resiliencia a la información adversa o engañosa	• Generación de texto adversario (TextFooler), patrones Regex • Benchmarks de ataques adversarios (PromptBench), número de refusals	• <i>Hacking</i> ético • Simulacros de incidentes
	3.5 Adaptabilidad	Capacidad para aprender o adaptarse a nuevos contextos	• Rendimiento del LLM ante datos nuevos por Zero/One/Few-shot learning	• Pruebas A/B • Revisión caso a caso
	3.6 Explicabilidad	Comprensión del proceso de toma de decisiones	• SHAP • Puntuaciones de explicabilidad	• <i>Hacking</i> ético • <i>Focus groups</i>
	3.7 Sesgos y equidad	Respuestas sin sesgo demográfico	• AI Fairness 360 toolkit • WEAT score, paridad demográfica, asociaciones de palabras, etc. • <i>Benchmarks</i> de sesgos (BBQ, etc.)	• <i>Hacking</i> ético • <i>Focus groups</i>
	3.8 Toxicidad	Propensión a generar contenidos nocivos	• Perspective API, Hatebase API • <i>Toxicity benchmarks</i> (RealToxicityPrompts, BOLD, etc.)	• <i>Hacking</i> ético • <i>Focus groups</i>
4. Implementación y uso	4.1 Revisión humana y seguridad de uso	Exclusión de sugerencias perjudiciales o ilegales e inclusión de una revisión humana (<i>'human-in-the-loop'</i>)	• Protocolos de riesgos, evaluaciones de seguridad • Control humano	• <i>Hacking</i> ético • <i>Focus groups</i>
	4.2 Recuperación y gestión de errores	Capacidad para recuperarse de errores y gestionar entradas inesperadas	• Tests de recuperación del sistema • Métricas de procesamiento de errores	• Simulacros de incidentes
	4.3 Escalabilidad	Mantenimiento del rendimiento con más datos o usuarios	• <i>Stress testing</i> del sistema, Apache Jmeter, etc. • <i>Benchmarks</i> de escalabilidad	• Simulacros de incidentes • Pruebas A/B
	4.4 Eficiencia	Utilización de recursos y velocidad de respuesta	• <i>Time-to-first-byte</i> (TTFB), uso de GPU/CPU, inferencia de emisiones, memoria, latencia	• Simulacros de incidentes
	4.5 Aceptación del usuario	Pruebas de aceptación de usuario	• <i>Checklist</i> de requisitos de usuario, <i>opt-out</i> del usuario • Satisfacción del usuario (Net Promoter Score, CSAT)	• UX tracking • Pruebas A/B

La selección exacta de técnicas dependerá de las características particulares del caso de uso; y, en concreto, varios factores importantes a tener en cuenta para decidir las técnicas más adecuadas son:

- El nivel de riesgo y la criticidad de las tareas que se confiarán al LLM.
- Si el LLM está abierto al público (y por tanto el *hacking* ético cobra especial relevancia) o si su uso se limita al ámbito interno de la organización.
- Si el LLM procesa datos personales.
- La línea de negocio o servicio que utilizará el LLM.

Un análisis cuidadoso de estos *drivers* permitirá construir un marco de validación robusto y adaptado a las necesidades de cada uso de un LLM.

Métricas de evaluación cuantitativas

Aunque es un campo de estudio emergente, existe una amplia gama de métricas cuantitativas para evaluar el rendimiento de los LLM. Algunas de estas métricas son adaptaciones de las utilizadas en modelos tradicionales de aprendizaje automático, como la precisión, la exhaustividad (*recall*), la puntuación F1 o el área bajo la curva ROC (AUC-ROC). Otras métricas han sido diseñadas específicamente para evaluar aspectos únicos de los LLM, como la coherencia del texto generado, la fidelidad a los hechos o la diversidad del lenguaje.

En este sentido, ya existen marcos holísticos de testeo cuantitativo de LLM en entornos de programación en Python, que facilitan la implementación de muchas de las métricas cuantitativas de validación, por ejemplo:

- **LLM Comparator**¹¹⁴: herramienta creada por investigadores de Google para la evaluación automática y comparación de LLM, que revisa la calidad de las respuestas de los LLM.
- **HELM**¹¹⁵: evaluación holística de los modelos del lenguaje, que compila métricas de evaluación a lo largo de siete dimensiones (precisión, calibración, robustez, equidad, sesgos, toxicidad y eficiencia) para una serie de escenarios predefinidos.
- **ReLM**¹¹⁶: sistema de validación y consulta de LLM mediante uso del lenguaje, incluyendo evaluaciones de modelos lingüísticos, memorización, sesgos, toxicidad y comprensión del lenguaje.

En la actualidad, ciertas técnicas de validación, como los métodos de explicabilidad (XAI) basados en SHAP, algunas métricas como ROUGE¹¹⁷ o los análisis de imparcialidad mediante paridad demográfica, aún no cuentan con umbrales predefinidos ampliamente aceptados. En estos casos, es tarea de la comunidad científica y de la industria seguir investigando para establecer criterios claros que permitan una validación robusta y estandarizada.

¹¹⁴Kahng (2024).

¹¹⁵Liang (2023).

¹¹⁶Kuchnik (2023).

¹¹⁷Duan (2023).

Fig. 11. Algunas técnicas de evaluación humana de LLM.



Técnicas de evaluación humana

Mientras que las métricas de evaluación cuantitativa son implementables de forma más directa debido a la multitud de recursos *online* y publicaciones de los últimos años, las técnicas de evaluación humana¹¹⁸ son variadas y deben ser construidas en función de la tarea específica¹¹⁹ que esté realizando el LLM, e incluyen (Fig. 11):

- ▶ **Backtest de los forzajes del usuario:** contabilizar y medir la importancia de las modificaciones humanas en los resultados del LLM (p. ej., cuántas veces un gestor comercial debe modificar manualmente los resúmenes de llamadas a clientes que ha realizado un LLM).
- ▶ **Revisión caso a caso:** comparar una muestra representativa de respuestas del LLM con las expectativas del usuario («*ground truth*”).
- ▶ **Hacking ético (Red Team):** manipular los *prompts* para forzar al LLM a producir resultados no deseados (p. ej., regurgitación de información personal, contenido ilegal, tests de penetración, explotación de vulnerabilidades).
- ▶ **Testeo A/B:** comparación para evaluar dos versiones del LLM (A y B), o de un LLM frente a un ser humano.
- ▶ **Focus groups:** recabar opiniones de diversos usuarios sobre el comportamiento del LLM, p. ej., en materia de ética, adecuación cultural, discriminación, etc.
- ▶ **Experiencia del usuario (UX tracking):** observar y evaluar las interacciones de los usuarios con el LLM a lo largo del tiempo o en tiempo real.
- ▶ **Simulacros de incidentes:** simular escenarios adversos para probar la respuesta del LLM (p. ej., prueba de estrés, comprobación de copias de seguridad, medición del tiempo de recuperación, etc.).
- ▶ **Mantenimiento de registros:** revisar los diarios y registros del sistema LLM, garantizando el cumplimiento de la normativa y la traza de auditoría.

Benchmarks de evaluación de LLM

La mayoría de los modelos de inteligencia artificial generativa, incluidos los LLM, se someten a pruebas utilizando *benchmarks* públicos que evalúan su desempeño en una variedad de tareas relacionadas con la comprensión y el uso del lenguaje natural. Estas pruebas sirven para medir cómo maneja el LLM tareas específicas y refleja el entendimiento humano. Algunos de estos *benchmarks* incluyen:

- ▶ GLUE/SuperGLUE: evalúa la comprensión del lenguaje a través de tareas que miden la capacidad de un modelo para entender el texto.
- ▶ Eleuther AI Language Model Evaluation Harness: realiza una evaluación “few-shot” de los modelos, es decir, su precisión con muy pocos ejemplos de entrenamiento.
- ▶ ARC (AI2 Reasoning Challenge): pone a prueba la habilidad del modelo para responder preguntas de ciencia que requieren razonamiento.
- ▶ HellaSwag: evalúa el sentido común del modelo a través de tareas que requieren predecir el final coherente de una historia.
- ▶ MMLU (Massive Multitask Language Understanding): prueba la precisión del modelo en una amplia gama de tareas para evaluar su comprensión multitarea.
- ▶ TruthfulQA: desafía al modelo a discernir entre información verdadera y falsa, evaluando su habilidad para manejar datos verídicos.
- ▶ Winogrande: otra herramienta para evaluar el sentido común, similar a HellaSwag pero con diferentes métodos y énfasis.
- ▶ GSM8K: evalúa la capacidad lógico-matemática del modelo a través de problemas de matemáticas diseñados para estudiantes.

¹¹⁸Datta, Dickerson (2023).

¹¹⁹Guzmán (2015).

Nuevas tendencias

El campo de la validación de LLM se encuentra en constante evolución, impulsado por los rápidos avances en el desarrollo de estos modelos y por la creciente conciencia sobre la importancia de garantizar su fiabilidad, equidad y alineación con la ética y la regulación.

A continuación, se presentan algunas de las principales tendencias emergentes en este ámbito:

- **Explicabilidad de los LLM:** a medida que los LLM ganan en complejidad y opacidad, crece la demanda de mecanismos que permitan entender y explicar su funcionamiento interno. Las técnicas de XAI (*eXplainable AI*) como SHAP, LIME o la atribución de importancia a los tokens de entrada están ganando protagonismo en la validación de LLM. Aunque para los modelos tradicionales hay una variedad de técnicas *post-hoc* disponibles para comprender el funcionamiento de los modelos a nivel local y global¹²⁰ (p. ej., Anchors, PDP, ICE), y ha proliferado la definición e implementación de modelos inherentemente interpretables por construcción, la implementación de estos principios para los LLM no está todavía resuelta.
- **Uso de LLM para explicar LLM:** una tendencia emergente consiste en utilizar un LLM para generar explicaciones sobre el comportamiento o las respuestas de otro LLM. En otras palabras, se emplea un modelo del lenguaje para interpretar y comunicar de forma más comprensible el razonamiento subyacente de otro modelo. Para enriquecer estas explicaciones, se están desarrollando herramientas¹²¹ que incorporan además técnicas de análisis *post-hoc*.

► **Técnicas de interpretabilidad *post-hoc*:** estas técnicas se basan en la interpretabilidad de los resultados en la etapa posterior al entrenamiento o *fine-tuning*, y permiten identificar qué partes de la entrada han influido más en la respuesta del modelo (importancia de características), encontrar ejemplos similares en el conjunto de datos de entrenamiento (similitud basada en *embeddings*) o diseñar *prompts* específicos que guíen al modelo hacia explicaciones más informativas (estrategias de *prompting*).

► **Puntuaciones por atribución:** como parte de la interpretabilidad *post-hoc*, se están desarrollando técnicas¹²² que permiten identificar qué partes del texto de entrada tienen mayor influencia en la respuesta generada por un LLM. Ayudan a entender qué palabras o frases son más importantes para el modelo. Existen diferentes métodos para calcular estas puntuaciones:

- Métodos basados en el gradiente: analizan cómo cambian los gradientes (una medida de sensibilidad) para cada palabra al retroceder por la red neuronal.
- Métodos basados en perturbaciones: modifican ligeramente el texto de entrada y observan cómo cambia la respuesta del modelo.
- Interpretación de métricas internas: utilizan métricas calculadas por el propio modelo, como los pesos de atención en los *transformers*, para determinar la importancia de cada palabra.

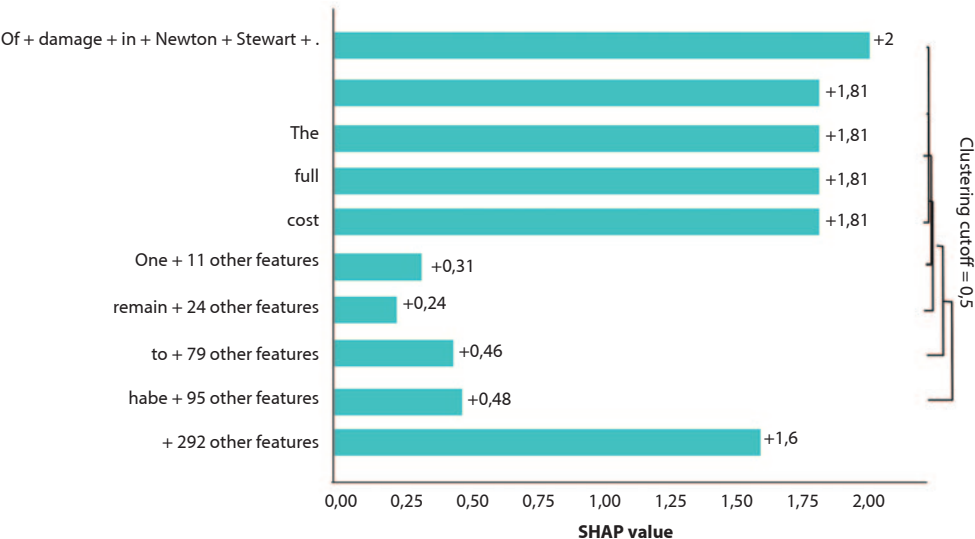
¹²⁰Management Solutions (2023). Explainable Artificial Intelligence.

¹²¹Wang (2024).

¹²²Sarti (2023).

Fig. 12. Implementación de valores de SHAP para resumen de textos.

Resumen de la salida: "Todavía se está evaluando el coste total de los daños en Newton Stewart, una de las zonas más afectadas . La Primera Ministra, Nicola Sturgeon, visitó la zona para inspeccionar los daños. El líder adjunto del Partido Laborista Escocés, Alex Rowley, estuvo el lunes en Hawick para ver la situación de primera mano. Afirmó que era importante aplicar correctamente el plan de protección contra las inundaciones".



Un ejemplo de puntuación por atribución es la aplicación de la técnica SHAP para proporcionar una medida cuantitativa de la importancia de cada palabra para la salida del LLM, lo que facilita su interpretación y comprensión (Fig. 12).

- **Validación continua y monitorización en producción:** más allá de la evaluación puntual antes del despliegue, se extiende la práctica de realizar un seguimiento continuo del comportamiento de los LLM una vez que están en uso, al igual que se hace con modelos tradicionales. Esto permite detectar posibles desviaciones o degradaciones en su rendimiento a lo largo del tiempo, así como identificar sesgos o riesgos no previstos inicialmente.
- **Validación colaborativa y participativa:** se promueve una mayor implicación de diversos stakeholders en el proceso de validación, incluyendo no solo a expertos técnicos sino también a usuarios finales, reguladores, auditorías externas y representantes de la sociedad civil. Esta participación plural permite incorporar diferentes perspectivas y fomenta la transparencia y la responsabilidad.
- **Validación ética y alineada con la regulación:** más allá de las métricas de rendimiento, se otorga cada vez más importancia a evaluar si el comportamiento de los LLM es ético y está alineado con los valores humanos y con la regulación. Esto implica analizar cuestiones como la equidad, la privacidad, la seguridad, la transparencia o el impacto social de estos sistemas.
- **Machine unlearning:** se trata de una técnica emergente¹²³ que permite “desaprender” información conocida de un LLM sin reentrenarlo desde cero. Esto se consigue, por ejemplo, adaptando los hiperparámetros del modelo a los datos que se desea desaprender. Se puede usar el mismo principio para eliminar los sesgos que se hayan identificado. Así, se obtiene un modelo que mantiene su conocimiento general, pero ha eliminado los sesgos problemáticos, mejorando su equidad y alineación ética de forma eficiente y selectiva. Actualmente se están explorando varios métodos de machine unlearning, como el *gradient descent*¹²⁴, el uso de *fine-tuning*¹²⁵ o la modificación selectiva de determinados pesos, capas o neuronas del modelo¹²⁶.

SHAP (SHapley Additive exPlanations) aplicado a un LLM

SHAP es un método de explicabilidad *post-hoc* basado en la teoría de juegos cooperativos. Asigna a cada característica (token) un valor de importancia (valor Shapley) que representa su contribución a la predicción del modelo.

Formalmente, sea $x = (x_1, \dots, x_n)$ una secuencia de tokens de entrada. La predicción del modelo se denota como $f(x)$. El valor Shapley ϕ para el token x_i se define como:

$$\phi_i = \sum_{S \subseteq N_i} \frac{|S|!(n - |S| - 1)!}{n!} [f(S \cup \{i\}) - f(S)]$$

donde N es el conjunto de todos los tokens, S es un subconjunto de tokens, y $f(S)$ es la predicción del modelo para el subconjunto S .

Intuitivamente, el valor Shapley ϕ_i captura el impacto promedio del token x_i en la predicción del modelo, considerando todos los subconjuntos posibles de tokens.

Ejemplo: se considera un LLM entrenado para clasificar correos electrónicos corporativos como “importante” o “no importante”. Dado el vector de tokens de entrada:

$x = [\text{El, informe, financiero, del, Q2, muestra, un, aumento, significativo, en, los, ingresos, y, la, rentabilidad}]$

El modelo clasifica el correo como “importante” con $f(x) = 0.85$.

Aplicando SHAP, se obtienen los siguientes valores Shapley:

- $\phi_1 = 0.01$ (El)
- $\phi_2 = 0.2$ (informe)
- $\phi_3 = 0.15$ (financiero)
- $\phi_4 = 0.02$ (del)
- $\phi_5 = 0.1$ (Q2)
- $\phi_6 = 0.05$ (muestra)
- $\phi_7 = 0.01$ (un)
- $\phi_8 = 0.15$ (aumento)
- $\phi_9 = 0.1$ (significativo)
- $\phi_{10} = 0.01$ (en)
- $\phi_{11} = 0.02$ (los)
- $\phi_{12} = 0.12$ (ingresos)
- $\phi_{13} = 0.01$ (y)
- $\phi_{14} = 0.02$ (la)
- $\phi_{15} = 0.08$ (rentabilidad)

Interpretación: los tokens “informe” (0.2), “financiero” (0.15), “aumento” (0.15) e “ingresos” (0.12) tienen las mayores contribuciones a la clasificación del correo como “importante”. Esto sugiere que el LLM ha aprendido a asociar estos términos con la importancia del mensaje en un contexto empresarial.

¹²³ Liu (2024).

¹²⁴ Jang (2022).

¹²⁵ Yu (2023).

¹²⁶ Wu (2023).

Caso práctico: validación de un chatbot de políticas

“La inteligencia artificial alcanzará niveles humanos hacia 2029”.

Ray Kurzweil¹²⁷

“Creo que tendremos una IA más inteligente que cualquier ser humano probablemente a finales de 2025”.

Perplexity¹²⁸



Para ilustrar la aplicación de las técnicas de validación de un LLM descritas, en esta sección se presenta un caso práctico de validación de un *chatbot* de políticas internas de una compañía.

Definición del caso

La compañía ha desarrollado un *chatbot* basado en un LLM de código abierto para responder preguntas y proporcionar información sobre sus políticas internas. El objetivo principal de este *chatbot* es facilitar el acceso de los empleados a las políticas de la empresa.

El *chatbot* ha sido construido utilizando una infraestructura en la nube y se ha alimentado con todas las políticas de la compañía, que abarcan aproximadamente 1.000 páginas de documentación. Para mejorar su capacidad de respuesta, se han aplicado técnicas de *Retrieval-Augmented Generation* (RAG), que permiten al modelo recuperar información relevante de su base de conocimientos antes de generar una respuesta. Inicialmente, se consideró la posibilidad de aplicar *fine-tuning* al modelo, pero tras las pruebas iniciales, se concluyó que la combinación del LLM base con RAG era suficiente para obtener resultados satisfactorios.

Antes de su implementación definitiva, la compañía ha decidido llevar a cabo un proceso de validación exhaustivo para evaluar la precisión, la seguridad y la adecuación del *chatbot* en el contexto específico de su uso previsto. Este proceso de validación tiene como objetivo identificar posibles áreas de mejora y garantizar que el *chatbot* cumpla con los estándares de calidad y las expectativas de la empresa.

La validación del *chatbot* de políticas se realizará mediante una combinación de métricas cuantitativas y técnicas de evaluación humana, siguiendo el marco de validación multidimensional descrito en la sección anterior. Los resultados de este proceso servirán para tomar decisiones informadas sobre la implementación del *chatbot* y para establecer un plan de mejora continua.

Diseño del enfoque de validación

Para validar de manera integral el *chatbot* de políticas, siguiendo el marco presentado en la sección anterior, se ha diseñado un enfoque de validación a medida que abarca las dimensiones clave del ciclo de vida del modelo: datos, diseño, evaluación, implementación y uso. Este enfoque combina métricas cuantitativas y técnicas de evaluación humana, con el objetivo de obtener una visión completa del desempeño y la adecuación del *chatbot* en el contexto específico de la compañía.

A continuación, se resumen las pruebas y técnicas seleccionadas para cada dimensión:

Datos

- Métricas: se utilizará la escala Flesch-Kincaid para evaluar la legibilidad y complejidad de las políticas que alimentan al *chatbot*.
- Evaluación humana: se revisará una muestra representativa de las políticas para identificar posibles inconsistencias, errores o ambigüedades.

Diseño del modelo

- Métricas: se modificarán elementos concretos del LLM en el código de desarrollo (p. ej., la técnica de RAG y sus hiperparámetros, como el tamaño o la estrategia de "*chunking*"¹²⁹) que pueden modificar su rendimiento ante respuestas, y se compararán los resultados contra el modelo original.

¹²⁷Ray Kurzweil (n. 1948). Director de Ingeniería en Google, científico computacional, inventor y futurista, conocido por la invención del OCR y por sus contribuciones en IA.

¹²⁸Elon Musk (n. 1971), CEO de X, SpaceX, Tesla. Empresario sudafricano-estadounidense, conocido por fundar o cofundar empresas como Tesla, SpaceX y PayPal, dueño de X (anteriormente Twitter), red social que tiene su propio LLM, llamado Grok.

¹²⁹La estrategia de "*chunking*" se refiere al proceso de dividir el texto de entrada a un LLM en unidades más pequeñas y manejables («chunks») durante su uso o implementación.

- Evaluación humana: se realizará una revisión exhaustiva de los componentes del *chatbot*, incluyendo la configuración de RAG, los filtros de entrada y salida, la definición de *prompts* y la optimización de hiperparámetros. Además, se llevarán a cabo pruebas A/B para comparar el desempeño del *chatbot* con otros LLM disponibles en el mercado.

Evaluación del modelo

► Privacidad y seguridad

- Métricas: se aplicarán pruebas de K-anonimato para evaluar la protección de datos personales en las respuestas del *chatbot*, y pruebas de PII (*Personal Identifiable Information*) para identificar atributos sensibles en los datos, utilizando PIIfilter.
- Evaluación humana: se realizarán pruebas de *hacking* ético para identificar posibles vulnerabilidades y se mantendrán registros detallados de las interacciones del *chatbot*.

► Precisión

- Métricas: se utilizarán las métricas *Word Error Rate* (WER) y ROUGE para evaluar la precisión de las respuestas del *chatbot* en comparación con las políticas originales. También se emplearán *benchmarks* específicos del dominio, como un conjunto de preguntas y respuestas diseñado por expertos en políticas de la compañía.
- Evaluación humana: se realizará una revisión caso por caso de una muestra representativa de interacciones del *chatbot* para identificar posibles errores o imprecisiones.

► Consistencia

- Métricas: se aplicarán la similitud coseno y el índice de Jaccard para evaluar la consistencia de las respuestas del *chatbot* ante consultas similares.
- Evaluación humana: se llevarán a cabo pruebas A/B para comparar las respuestas del *chatbot* en diferentes escenarios y se realizará una revisión caso por caso para identificar posibles inconsistencias.

► Robustez

- Métricas: se utilizarán herramientas como TextFooler para generar texto adversario y evaluar la resiliencia del *chatbot* ante información engañosa. Además, se contabilizará el número de rechazos del *chatbot* ante *prompts* malintencionados.
- Evaluación humana: se realizarán pruebas de *hacking* ético y simulacros de incidentes para evaluar la capacidad del *chatbot* para manejar situaciones adversas.

► Adaptabilidad

- Métricas: se evaluará el rendimiento del *chatbot* ante nuevas políticas o actualizaciones mediante técnicas de *few-shot learning*. Se evaluará la respuesta del *chatbot* ante idiomas no empleados en las políticas o solicitudes de traducciones a idiomas no incluidos en el RAG (p. ej., polaco).

- Evaluación humana: se realizarán pruebas A/B y revisiones caso por caso para evaluar la capacidad del *chatbot* para adaptarse a nuevos escenarios.

► Explicabilidad

- Métricas: se aplicarán técnicas de explicabilidad, como SHAP, para comprender el proceso de toma de decisiones del *chatbot*. Se evaluará el módulo de interpretabilidad intrínseco del *chatbot*, mediante el cual se provee una explicación del origen de la información en la respuesta al usuario.

- Evaluación humana: se realizará un seguimiento de la experiencia del usuario (UX) y se llevará a cabo un *focus group* para evaluar la percepción de los usuarios sobre la transparencia y la explicabilidad del *chatbot*.

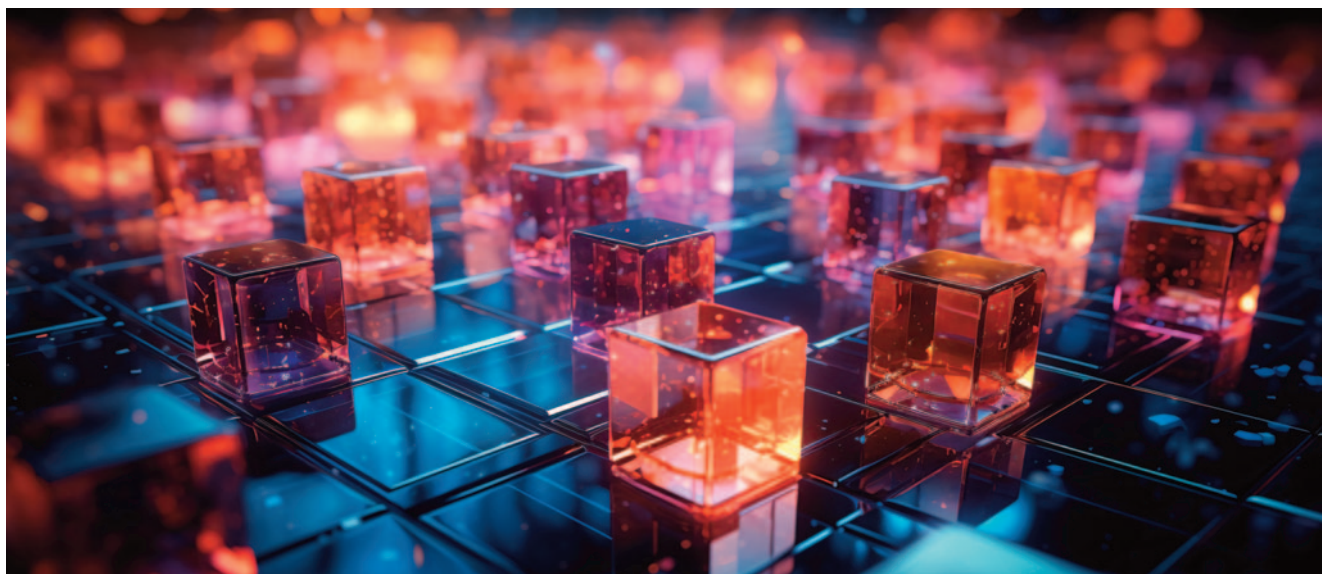
► Sesgos y equidad

- Métricas: se utilizará el conjunto de herramientas AI Fairness 360 para evaluar posibles sesgos demográficos en las respuestas del *chatbot*. También se emplearán *benchmarks* específicos, como el Bias Benchmark for QA (BBQ), para medir la equidad en el contexto de las políticas de la compañía.

- Evaluación humana: se llevarán a cabo pruebas de *hacking* ético y un *focus group* para identificar posibles sesgos o discriminación en las respuestas del *chatbot*.

► Toxicidad

- Métricas: se aplicarán las herramientas Perspective API y Hatebase API para evaluar la presencia de lenguaje tóxico o inapropiado en las respuestas del *chatbot*. Además, se utilizarán *benchmarks* específicos, como RealToxicityPrompts, para medir la toxicidad en el contexto de las políticas de la compañía.
- Evaluación humana: se realizarán pruebas de *hacking* ético para identificar posibles casos de lenguaje ofensivo o inapropiado en las interacciones del *chatbot*.



Implementación y uso

► Escalabilidad

- Métricas: se realizarán pruebas de estrés del sistema utilizando Apache JMeter para evaluar el rendimiento del *chatbot* bajo altas cargas de trabajo.
- Evaluación humana: se llevarán a cabo simulacros para evaluar la capacidad del *chatbot* para manejar un aumento imprevisto en el número de usuarios o consultas.

► Eficiencia

- Métricas: se medirán el tiempo de respuesta (Time-to-First-Byte, TTFB), el uso de recursos (GPU/CPU, memoria) y la latencia para evaluar la eficiencia del *chatbot*.

► Aceptación del usuario

- Métricas: se establecerá una lista de verificación de los requisitos del usuario y se medirá la satisfacción del usuario utilizando indicadores como el Net Promoter Score (NPS) y el Customer Satisfaction Score (CSAT).
- Evaluación humana: se realizará un seguimiento de la experiencia del usuario (UX) para evaluar la aceptación y satisfacción de los usuarios con el *chatbot*.

Este enfoque de validación personalizado permitirá a la compañía obtener una evaluación completa del *chatbot* de políticas, identificando áreas de mejora y garantizando su adecuación para el uso previsto. Los resultados de estas pruebas y evaluaciones servirán como base para la toma de decisiones informadas sobre la implementación y el perfeccionamiento continuo del *chatbot*.

Resultados

Tras aplicar el enfoque de validación personalizado al *chatbot* de políticas, se obtuvieron resultados prometedores que demuestran su adecuación general para el uso previsto en la compañía (Fig. 13). En la mayoría de las dimensiones evaluadas, el *chatbot* alcanzó un desempeño satisfactorio, cumpliendo con los estándares de calidad y las expectativas establecidas.

En cuanto a la calidad de los datos de entrada, se encontró que las políticas que alimentan al *chatbot* tienen, en general, un nivel de legibilidad y complejidad adecuado para su comprensión por parte de los usuarios. Además, la revisión humana no identificó inconsistencias significativas o errores en el contenido de las políticas.

El diseño del modelo también demostró ser apropiado para el caso de uso, con una configuración óptima de los componentes del *chatbot* y un rendimiento superior en comparación con otros LLM disponibles en el mercado.

En términos de evaluación del modelo, el *chatbot* obtuvo resultados positivos en la mayoría de las métricas y pruebas aplicadas. Se destacan la alta precisión de las respuestas, la consistencia en el manejo de consultas similares y la capacidad para adaptarse a nuevos escenarios. Sin embargo, se identificaron algunas áreas de mejora en aspectos como la explicabilidad, la detección de sesgos y la respuesta a preguntas muy específicas donde se requiere un mayor perfeccionamiento del modelo. En el ámbito de ciberseguridad, se requiere un análisis más detallado de las vulnerabilidades específicas de los LLM *open-source* empleados, para mitigar ese riesgo en su puesta en producción.

En cuanto a la implementación y uso, el *chatbot* demostró una buena escalabilidad y eficiencia en el manejo de altas cargas de trabajo. Además, la satisfacción de los usuarios fue alta, lo que indica una buena aceptación de la herramienta en el contexto de la compañía.

Fig. 13. Resumen de resultados de las métricas y técnicas de evaluación humana del chatbot de políticas.

Dimensión	Prueba	Resultado	Interpretación
Datos	Flesch-Kincaid	Legibilidad adecuada (grado 8)	Las políticas son comprensibles para la mayoría de los usuarios
	Revisión humana	Sin inconsistencias significativas	Las políticas son coherentes y no contienen errores importantes
Diseño del modelo	Modelos challenger	Mejoras en parámetros identificadas	Se requiere adaptar los parámetros del RAG al contexto de las políticas (p. ej., chunk size) para mejorar la captura de información en preguntas muy específicas
	Revisión de componentes	Configuración óptima	El diseño del chatbot es apropiado para el caso de uso
	Pruebas A/B	Rendimiento superior a otros LLM	El chatbot supera a otros modelos disponibles en el mercado
Evaluación del modelo	K-anonimato	Protección adecuada de datos personales	El chatbot no revela información sensible en sus respuestas
	Hacking ético	Vulnerabilidades menores identificadas	Se requieren ajustes para fortalecer la seguridad del chatbot
	Word Error Rate (WER)	WER < 5%	Las respuestas del chatbot son altamente precisas
	ROUGE	ROUGE-L > 0.8	Las respuestas del chatbot capturan adecuadamente el contenido de las políticas
	Similitud coseno / Índice de Jaccard	Similitud > 0.9	El chatbot proporciona respuestas consistentes ante consultas similares
	TextFooler	Resiliencia moderada ante texto adversario	El chatbot es moderadamente robusto ante información engañosa
	Few-shot learning	Adaptabilidad satisfactoria	El chatbot puede adaptarse a nuevas políticas o actualizaciones con un entrenamiento mínimo, pero se requiere monitorizar y agregar esos nuevos documentos al RAG periódicamente
	SHAP	Explicabilidad limitada	Se requiere mejorar la capacidad del chatbot para explicar sus decisiones, si bien el componente de RAG se ha construido de manera que el LLM da una respuesta autoexplicativa
	AI Fairness 360 / BBQ	Sesgos demográficos menores identificados	El chatbot presenta algunos sesgos que deben ser mitigados
Implementación y uso	Perspective API / RealToxicityPrompts	Toxicidad baja (< 5%)	Las respuestas del chatbot rara vez contienen lenguaje tóxico o inapropiado
	Apache JMeter	Escalabilidad satisfactoria (hasta 1000 usuarios)	El chatbot puede manejar altas cargas de trabajo sin degradación significativa del rendimiento
	TTFB / Uso de recursos / Latencia	Eficiencia adecuada (TTFB < 1s, uso moderado)	El chatbot responde rápidamente y utiliza los recursos de manera eficiente
	NPS / CSAT	Satisfacción alta (NPS > 60, CSAT > 80%)	Los usuarios están altamente satisfechos con el chatbot y lo recomendarían a otros

Estos resultados indican que el *chatbot* de políticas está bien encaminado para su implementación en la compañía, aunque se han identificado algunas áreas específicas que requieren mejoras adicionales. La sección siguiente abordará las principales conclusiones y recomendaciones derivadas de este proceso de validación.

Principales conclusiones

El proceso de validación del *chatbot* de políticas ha demostrado que este sistema basado en LLM puede ser una herramienta valiosa para facilitar el acceso de los empleados a la información relevante de la compañía. Los resultados obtenidos en las diversas pruebas y evaluaciones indican que el *chatbot* cumple, en gran medida, con los requisitos de calidad, seguridad y eficiencia establecidos por la organización.

Entre las fortalezas identificadas, se destacan la precisión y consistencia de las respuestas del *chatbot*, su capacidad para adaptarse a nuevos escenarios y su escalabilidad para manejar altas cargas de trabajo. Además, la satisfacción de los usuarios con la herramienta es alta, lo que sugiere una buena aceptación y adopción por parte de los empleados.

Sin embargo, el proceso de validación también ha revelado algunas áreas de mejora que deben abordarse antes de la implementación definitiva del *chatbot*. En particular, se recomienda:

1. Mejorar la explicabilidad del modelo: es necesario desarrollar técnicas más avanzadas para que el *chatbot* pueda proporcionar explicaciones claras y comprensibles sobre su proceso de toma de decisiones. Esto aumentará la transparencia y la confianza de los usuarios en la herramienta. Si bien el componente de RAG se ha construido de manera que el LLM da una respuesta autoexplicativa y hace referencia a la política

correspondiente, esta explicación no resulta del todo clara para preguntas muy específicas.

2. Mitigar los sesgos identificados: aunque los sesgos detectados son menores, es recomendable aplicar técnicas de *debiasing* para garantizar que las respuestas del *chatbot* sean equitativas y no discriminatorias. Se sugiere una revisión periódica de los sesgos y la implementación de medidas correctivas cuando sea necesario.

3. Fortalecer la seguridad y privacidad: si bien el *chatbot* cumple con los estándares básicos de protección de datos personales, se recomienda realizar pruebas adicionales y recurrentes de *hacking* ético y adoptar medidas de seguridad más robustas para prevenir posibles vulnerabilidades.

4. Establecer un plan de monitoreo y mejora continua: es fundamental definir un proceso de seguimiento y evaluación periódica del desempeño del *chatbot*, con el fin de identificar oportunidades de mejora y garantizar su óptimo funcionamiento a largo plazo. Este plan debe incluir la recopilación de feedback de los usuarios, la actualización regular de las políticas y su inclusión en la base de datos del *chatbot*, el monitoreo para mejorar los parámetros empleados en el RAG y su actualización, y la incorporación de nuevas técnicas y tecnologías cuando estén disponibles.

En conclusión, el *chatbot* de políticas ha demostrado tener potencial para mejorar la eficiencia y la accesibilidad de la información en la compañía. Con la implementación de las mejoras sugeridas y un enfoque de perfeccionamiento continuo, este sistema basado en LLM puede convertirse en una herramienta estratégica para el éxito de la organización. La recomendación final ha sido proceder con la implementación del *chatbot*, teniendo en cuenta las observaciones y recomendaciones derivadas de este proceso de validación.



Conclusiones

“Los LLM son los únicos que pueden escribir una novela, traducirla a diez idiomas, y aun así no entender la trama”.

Perplexity¹³⁰



Los modelos de lenguaje a gran escala (LLM) representan un avance significativo en el campo de la inteligencia artificial y están revolucionando la forma en que interactuamos con la tecnología y aprovechamos el procesamiento del lenguaje natural. Su capacidad para procesar y generar texto coherente y contextualizado abre un amplio abanico de aplicaciones en diversos sectores, desde la creación de contenido y el análisis de sentimientos hasta la automatización de tareas y la mejora de la experiencia del usuario.

Sin embargo, el desarrollo y despliegue de LLM también conlleva una serie de retos y consideraciones importantes. La presencia de sesgos y alucinaciones en sus resultados, la falta de transparencia y explicabilidad de sus decisiones, los desafíos de privacidad y seguridad de la información, y el alto consumo de recursos computacionales son algunos de los principales desafíos que deben abordarse para garantizar un uso responsable y ético de estos sistemas.

Para hacer frente a estos retos, es fundamental establecer un marco robusto de gobierno de la IA, especialmente en el ámbito de la IA generativa y los LLM. Este marco debe abarcar todos los aspectos clave, incluyendo la estrategia, el apetito al riesgo, el gobierno, la organización, el marco de control (políticas y procedimientos), los datos, los sistemas y el *reporting*. Solo a través de un enfoque integral y bien estructurado se podrá garantizar un desarrollo y uso responsable de estas tecnologías.

Dentro de este marco de gobierno, la validación juega un papel crucial. Es conveniente adoptar un enfoque multidimensional que abarque todas las etapas del ciclo de vida de los LLM, desde la calidad de los datos de entrada y la solidez del diseño del modelo hasta la evaluación exhaustiva de sus resultados y la implementación y uso adecuados. Este proceso de validación debe combinar métricas cuantitativas estandarizadas con técnicas de evaluación humana adaptadas al contexto específico de cada caso de uso.

Además, resulta necesario mantenerse al tanto de las últimas tendencias y avances en el campo de la validación de LLM, como el desarrollo de técnicas de explicabilidad más avanzadas, el uso de LLM para explicar el comportamiento de otros LLM, la validación continua y monitorización en producción, y la alineación con los principios éticos y los requisitos regulatorios.

El caso práctico presentado en este *white paper* ilustra cómo la aplicación de un marco de validación personalizado puede ayudar a las organizaciones a identificar fortalezas y áreas de mejora en sus sistemas basados en LLM, y a tomar decisiones informadas sobre su implementación y perfeccionamiento continuo.

En definitiva, los LLM tienen un potencial muy significativo para transformar la forma en que las empresas y la sociedad en general se benefician de la inteligencia artificial. Sin embargo, para aprovechar todo su potencial de manera segura y responsable, es imprescindible establecer un marco sólido de gobierno de la IA que aborde los retos asociados a su desarrollo y despliegue, y que incluya un enfoque riguroso y multidimensional de validación. Solo así podrá garantizarse que estos sistemas sean fiables, equitativos y estén alineados con los valores y objetivos de las organizaciones y de la sociedad.

¹³⁰Perplexity AI es un motor de conversación y búsqueda de investigación basado en LLM fundado en 2022 por Andy Konwinski, Denis Yarats, Johnny Ho y Aravind Srinivas (ex OpenAI), que responde a las consultas mediante texto predictivo en lenguaje natural.

Glosario



AGI (Artificial General Intelligence): inteligencia artificial hipotética futura que igualaría o superaría la inteligencia humana en cualquier dominio intelectual, siendo capaz de realizar cualquier tarea intelectual que un ser humano puede hacer.

Alucinaciones: generación de información o contenido por parte de un LLM que parece plausible pero que no se basa en hechos reales o en el conocimiento adquirido durante el entrenamiento, llevando a inexactitudes o invenciones en las respuestas del modelo.

CNN (Convolutional Neural Network): tipo de red neuronal especializada en procesar datos con una topología de cuadrícula, como imágenes o series temporales. Las CNN utilizan capas de convolución para extraer automáticamente características locales y abstractas de los datos, y son ampliamente utilizadas en tareas de visión por computador y procesamiento de señales.

Cuantización: técnica utilizada para reducir el tamaño y acelerar la inferencia de los LLM, que consiste en reducir la precisión numérica de los pesos del modelo, pasando de números en coma flotante a representaciones de menor precisión, como enteros o números en coma fija.

Datos de entrenamiento: conjunto de ejemplos utilizados para entrenar un modelo de aprendizaje automático, que incluyen las entradas (*features*) y, en el caso del aprendizaje supervisado, las etiquetas o respuestas esperadas. La calidad y diversidad de estos datos es crucial para el rendimiento y la generalización del modelo.

Efecto Eliza: fenómeno psicológico por el cual los usuarios tienden a atribuir capacidades cognitivas y emocionales similares a las humanas a los sistemas de conversación basados en IA, a pesar de que estos sistemas no poseen una comprensión real del lenguaje ni inteligencia general.

Embeddings: representaciones densas y continuas de elementos discretos (como palabras, frases o documentos) en un espacio vectorial de alta dimensión, donde elementos similares tienen representaciones cercanas. Se utilizan en los LLM para capturar relaciones semánticas y sintácticas entre los elementos del lenguaje.

Ética de la IA: disciplina que estudia los principios morales, valores y directrices que deben guiar el desarrollo, despliegue y uso de los sistemas de inteligencia artificial, con el objetivo de garantizar que sean beneficiosos, justos, transparentes y alineados con los valores humanos.

Evaluación humana: proceso de revisión y valoración cualitativa del comportamiento y resultados de un sistema de IA por parte de expertos y usuarios, que complementa las métricas cuantitativas y permite detectar errores, sesgos o comportamientos indeseados que podrían pasar desapercibidos en una evaluación puramente automática.

Explicabilidad (XAI, eXplainable AI): propiedad de un modelo de IA que se refiere a su capacidad para proporcionar explicaciones comprensibles para los humanos sobre su funcionamiento interno, el razonamiento detrás de sus predicciones y los factores que influyen en sus decisiones.

Few-shot learning: capacidad de un modelo de aprendizaje automático, especialmente los LLM, para aprender a realizar una nueva tarea a partir de pocos ejemplos (desde uno hasta unas decenas), aprovechando el conocimiento previo adquirido durante el preentrenamiento en grandes cantidades de datos.

Fine-tuning: técnica para adaptar un modelo de lenguaje preentrenado a una tarea específica, mediante el entrenamiento adicional con un conjunto de datos más pequeño y especializado en esa tarea. Permite aprovechar el conocimiento general del modelo y ajustarlo para obtener un alto rendimiento en aplicaciones concretas.



Hacking ético: práctica de probar y desafiar un sistema de IA de manera controlada y con permiso, con el objetivo de identificar vulnerabilidades, fallos, sesgos o comportamientos no deseados, para posteriormente corregirlos y mejorar la seguridad y robustez del sistema.

Instruction tuning: técnica de ajuste fino para LLM que consiste en proporcionar al modelo instrucciones, preguntas y ejemplos de respuestas esperadas, con el objetivo de alinear su comportamiento con las expectativas y preferencias de los usuarios en un dominio específico.

Inteligencia Artificial (IA): campo de la informática y la ingeniería que se dedica al desarrollo de sistemas capaces de realizar tareas que normalmente requieren inteligencia humana, como el aprendizaje, el razonamiento, la percepción, la interacción en lenguaje natural y la resolución de problemas.

Inteligencia artificial generativa (GenAI): subcampo de la IA que se enfoca en la creación de modelos y algoritmos capaces de generar contenido nuevo y original, como texto, imágenes, vídeo, audio, código fuente o diseños 3D, aprendiendo patrones y características a partir de un conjunto de datos de entrenamiento.

Large Language Models (LLM): modelos de aprendizaje profundo especializados en el procesamiento y generación de lenguaje natural, entrenados en enormes cantidades de texto y con un gran número de parámetros (desde millones hasta billones), capaces de realizar diversas tareas lingüísticas con un alto nivel de comprensión y coherencia.

LLMOps (Large Language Model Operations): conjunto de prácticas, herramientas y procesos para gestionar de manera eficiente y escalable el ciclo de vida completo de los LLM en entornos de producción, abarcando el entrenamiento, despliegue, monitorización, actualización y gobierno de estos modelos.

Machine learning: rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a los sistemas aprender y mejorar automáticamente a través de la experiencia, sin ser programados explícitamente para ello.

Machine unlearning: conjunto de técnicas para eliminar o "desaprender" de manera selectiva cierta información o sesgos indeseados de un modelo de aprendizaje automático ya entrenado, sin necesidad de reentrenarlo desde cero, permitiendo cumplir con requisitos de privacidad o corregir comportamientos no deseados.

Métricas cuantitativas: medidas numéricas estandarizadas utilizadas para evaluar de manera objetiva y consistente el rendimiento de un modelo de IA en tareas específicas, como la precisión, la exhaustividad, la exactitud o la eficiencia.

Modelo generativo: tipo de modelo de aprendizaje automático diseñado para aprender la distribución de probabilidad subyacente a un conjunto de datos y generar nuevas muestras que sean similares a los datos de entrenamiento, pudiendo crear contenido nuevo y realista.

Preentrenamiento: etapa inicial del entrenamiento de un LLM en la que se utiliza un gran corpus de texto no estructurado y sin etiquetar para que el modelo aprenda representaciones generales y patrones del lenguaje, adquiriendo un conocimiento amplio y robusto que luego puede ser adaptado a tareas específicas mediante *fine-tuning*.

Privacidad diferencial: técnica criptográfica utilizada para compartir información agregada sobre un conjunto de datos, mientras se protege la privacidad de los individuos presentes en esos datos, introduciendo un ruido aleatorio que dificulta la identificación de entradas individuales a partir de los resultados del análisis.

Prompt engineering: disciplina que se enfoca en diseñar, optimizar y adaptar los *prompts* (entradas de texto) para obtener los mejores resultados posibles de los LLM en tareas específicas, aprovechando técnicas como la inclusión de ejemplos, la especificación de formatos o la orientación paso a paso.

Pruebas A/B: método experimental utilizado para comparar el rendimiento de dos versiones diferentes de un sistema de IA (A y B) o entre un sistema de IA y un enfoque alternativo (como un humano o un modelo base), con el objetivo de determinar cuál funciona mejor según métricas predefinidas.

Regulación de la IA: conjunto de leyes, normativas, estándares y directrices establecidos por gobiernos y organizaciones para garantizar que el desarrollo, despliegue y uso de los sistemas de inteligencia artificial se realice de manera responsable, segura, ética y alineada con los valores y derechos fundamentales de la sociedad.

Retrieval-Augmented Generation (RAG): técnica utilizada en los LLM que consiste en recuperar información relevante de una base de conocimientos externa antes de generar una respuesta, combinando así la capacidad de acceso a información estructurada con la generación de lenguaje natural coherente y fluido.

RNN (Recurrent Neural Network): tipo de red neuronal diseñada para procesar secuencias de datos, como texto o series temporales. A diferencia de las redes neuronales *feedforward*, las RNN tienen conexiones recurrentes que les permiten mantener un estado interno y capturar dependencias temporales. Variantes como LSTM y GRU han sido ampliamente utilizadas en tareas de procesamiento del lenguaje natural antes del auge de los *transformers*.

Seguridad (AI safety): disciplina que se enfoca en identificar, prevenir y mitigar los riesgos potenciales asociados con el desarrollo y uso de sistemas de IA avanzados, tanto a corto como a largo plazo, incluyendo riesgos de seguridad, sesgos, errores, mal uso o consecuencias no deseadas.

Sesgo: tendencia sistemática de un modelo de aprendizaje automático a producir resultados que favorecen o perjudican injustamente a ciertos grupos o individuos, debido a características sensibles como el género, la etnia, la edad o la orientación sexual, y que suele ser resultado de sesgos presentes en los datos de entrenamiento o de decisiones subóptimas durante el desarrollo del modelo.

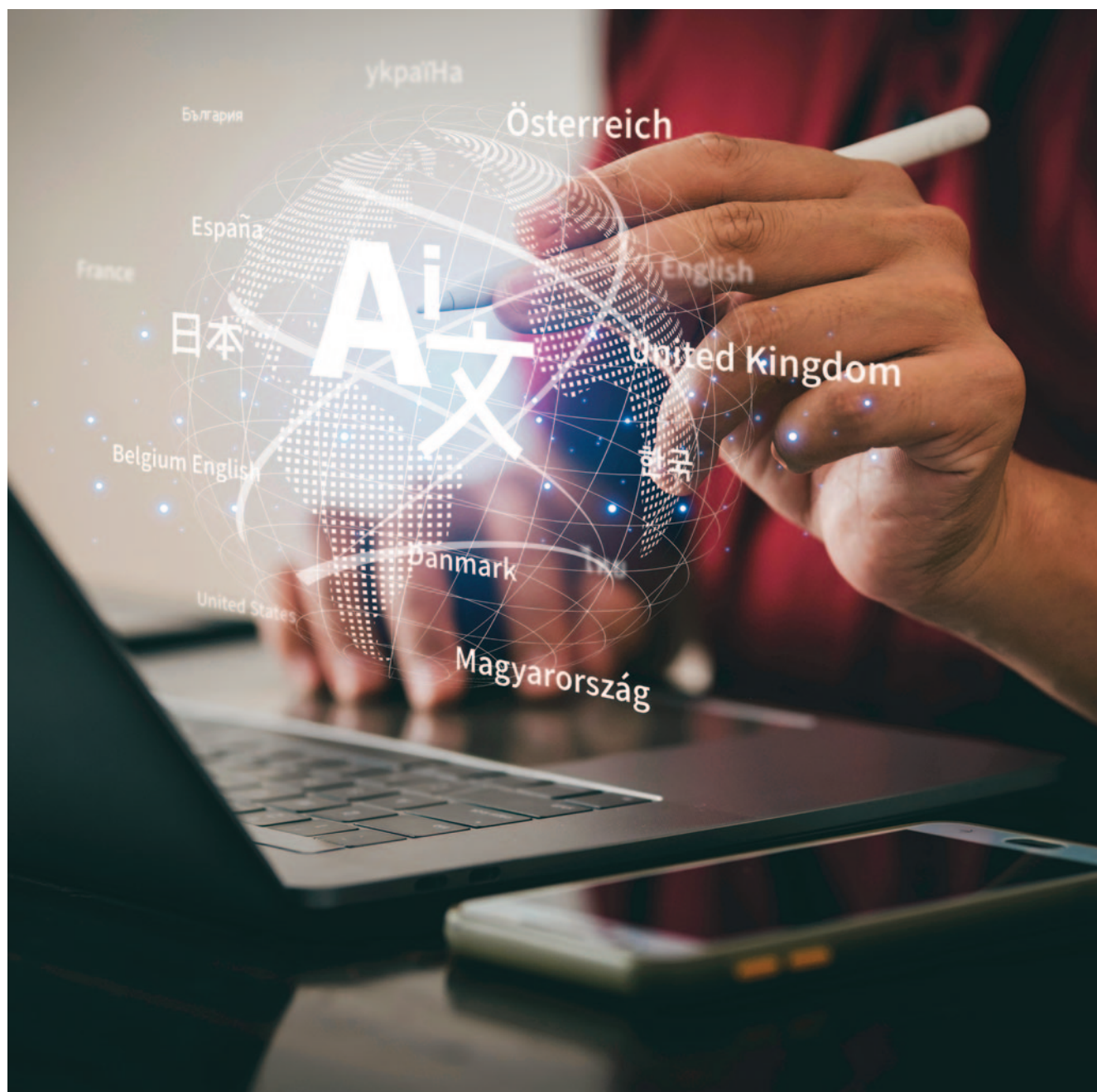
Token: unidad discreta en la que se divide un texto para su procesamiento por parte de un modelo de lenguaje. Los tokens pueden ser palabras, subpalabras o caracteres, y constituyen la entrada básica para el entrenamiento y la inferencia de los LLM.

Tokenización: proceso de convertir un texto en una secuencia de tokens. La elección de la estrategia de *tokenización* tiene un impacto significativo en el rendimiento y la eficiencia del modelo.

Transformers: arquitectura de red neuronal profunda que utiliza mecanismos de atención para procesar y generar secuencias de forma paralela, en lugar de secuencialmente como las RNNs. Permite capturar dependencias a largo plazo y contextuales, siendo la arquitectura dominante para los LLM y estableciendo el estado del arte en diversas tareas de procesamiento del lenguaje natural.

Validación: proceso integral y multidisciplinar para evaluar un sistema de IA, especialmente LLM, en términos de rendimiento, robustez, seguridad, equidad, explicabilidad y alineación con los requisitos y valores éticos y sociales, combinando métricas cuantitativas y evaluación cualitativa por parte de expertos y usuarios.

Bibliografía



Abhyankar, R. et al. (2024). APIServe: Efficient API Support for Large-Language Model Inferencing. <https://arxiv.org/abs/2402.01869>. arXiv:2402.01869v1

Alabdulmohsin, I. et al. (2024). CLIP the Bias: How Useful is Balancing Data in Multimodal Learning? <https://arxiv.org/html/2403.04547v1>. arXiv:2403.04547v1

Banerjee, I., et al. (2023). MLOps with enhanced performance control and observability. <https://arxiv.org/abs/2302.01061>. arXiv:2302.01061v1

Bengio, Y. et al. (2003). A Neural Probabilistic Language Model. <https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf>

Bréal, M. (1883). Les lois intellectuelles du langage fragment de sémantique. *Annuaire de l'Association pour l'encouragement des études grecques en France*. Vol. 17 (1883), pp. 132-142. <https://www.jstor.org/stable/44253893>

Cambon, A. et al. (2023). Early LLM-based Tools for Enterprise Information Workers Likely Provide Meaningful Boosts to Productivity. A first update from Microsoft's research initiative on AI and Productivity.

Chen, D. et al. (2023). Data-Juicer: A One-Stop Data Processing System for Large Language Models. <https://arxiv.org/abs/2309.02033>. arXiv:2309.02033v3

Chen, Y. et al. (2023). LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models. <https://arxiv.org/abs/2309.12307>. arXiv:2309.12307v3

Chiang, C. et al. (2023). Can Large Language Models Be an Alternative to Human Evaluations? <https://arxiv.org/abs/2305.01937>. arXiv:2305.01937v1

Chu, T., Song, Z., Yang, C. (2023). How to Protect Copyright Data in Optimization of Large Language Models? <https://arxiv.org/abs/2308.12247>. arXiv:2308.12247v1

CIO (2023). Chief AI Officer: What it takes to land the C-suite's hottest new job. <https://www.cio.com/article/657977/chief-ai-officer-what-it-takes-to-land-the-c-suites-hottest-new-job.html>

Cui, Q. et al. (2022). Contrastive Vision-Language Pre-training with Limited Resources. <https://arxiv.org/abs/2112.09331>. arXiv:2112.09331v3

CommetML. <https://www.comet.com/site/>.

Datta, T. et al. (2023). Who's Thinking? A Push for Human-Centered Evaluation of LLMs using the XAI Playbook. <https://arxiv.org/abs/2303.06223>. arXiv:2303.06223v1

Dettmers, T. et al. (2023). QLoRA: Efficient Finetuning of Quantized LLMs <https://arxiv.org/abs/2305.14314>. arXiv:2305.14314v1

Devlin, J. et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. <https://arxiv.org/abs/1810.04805>. arXiv:1810.04805v2

Duan, J. et al. (2023). Shifting attention to relevance: towards the uncertainty estimation of large language models. <https://arxiv.org/abs/2307.01379>. arXiv:2307.01379v2

Dun, C. et al. (2024). Sweeping Heterogeneity with Smart MoPs: Mixture of Prompts for LLM Task Adaptation. <https://arxiv.org/abs/2310.02842>. arXiv:2310.02842v2

Elazar, Y. et al. (2021). Measuring and Improving Consistency in Pretrained Language Models. <https://aclanthology.org/2021.tacl-1.60/>.

Euronews (2023). 2023 was the year AI went mainstream. It was also the year we started to panic about it. <https://www.euronews.com/next/2023/12/27/2023-was-the-year-ai-went-mainstream-it-was-also-the-year-we-started-to-panic-about-it>

European Parliament (2024). Artificial Intelligence Act / European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)). <https://artificialintelligenceact.eu/>; <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>

European Commission (2024). Knowledge Center on Interpretation. <https://knowledge-centre-interpretation.education.ec.europa.eu/en/news/what-large-language-model>

Fisher, M., Campagna, G., Choi, E., Lam, M. S., Freund, S. N., Yahav, E., (2021). DIY Assistant: A Multi-modal End-User Programmable Virtual Assistant. <https://dl.acm.org/doi/10.1145/3453483.3454046>.

Gartner (2023). What is generative AI? <https://www.gartner.com/en/topics/generative-ai>

Google DeepMind (2023). Levels of AGI: Operationalizing Progress on the Path to AGI. Meredith Ringel Morris; Jascha Sohl-Dickstein; Noah Fiedel; Tris Warkentin; Allan Dafoe; Aleksandra Faust; Clement Farabet; and Shane Legg. arXiv:2311.02462v1

Google + Implement (2023). The economic opportunity of generative AI in D9+. An Implement Consulting Group study commissioned by Google.

Gozalo-Brizuela, R., y Garrido-Merchán, E.C. (2023). A survey of Generative AI Applications. <https://ar5iv.labs.arxiv.org/html/2306.02781>

Guo, Z. et al. (2023). Evaluating Large Language Models: A Comprehensive Survey. <https://arxiv.org/pdf/2310.19736.pdf>. arXiv:2310.19736v3

Guzman, F. et al. (2015). How do Humans Evaluate Machine Translation. <https://aclanthology.org/W15-3059.pdf>.

Fu, HY. et al. (2023). Estimating Large Language Model Capabilities without Labeled Test Data. <https://arxiv.org/abs/2305.14802>. arXiv:2305.14802v2

Fu, X. et al (2024). Tiny Titans: Can Smaller Large Language Models Punch Above Their Weight in the Real World for Meeting Summarization? <https://arxiv.org/abs/2402.00841>. arXiv:2402.00841

Goyal, S. et al (2024). LLMGuard: Guarding Against Unsafe LLM Behavior. <https://arxiv.org/abs/2403.00826>. arXiv:2403.00826v1

Hendrycks, D. et al (2021). Measuring Massive Multitask Language Understanding. <https://arxiv.org/abs/2009.03300>. arXiv:2009.03300v3

Huang, L. et al. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. <https://arxiv.org/abs/2311.05232>. arXiv:2311.05232v1

Hugging Face Datasets (2024). CodeParrot. <https://huggingface.co/codeparrot>.

IAPP (2024). Global AI Law and Policy Tracker. <https://iapp.org/resources/article/global-ai-legislation-tracker/>

iDanae 2T23 (2023): Large Language Models: una nueva era en la inteligencia artificial. Cátedra iDanae. Newsletter trimestral 2T23. <http://www.idanae-stem.com/>

iDanae 1T24 (2024): Hacia una inteligencia artificial sostenible. Cátedra iDanae. Newsletter trimestral 1T24. <http://www.idanae-stem.com/>

Imperial, JM., et al. (2023). Flesch or Fumble? Evaluating Readability Standard Alignment of Instruction-Tuned Language Models. <https://arxiv.org/abs/2309.05454>. arXiv:2309.05454v2

IndesIA (2024). Barómetro de adopción de la inteligencia artificial en las pymes españolas. <https://www.indesia.org/wp-content/uploads/2024/04/IndesIA.-Barometro-de-adopcion-de-la-inteligencia-artificial-en-las-pymes-espanolas-Edicion-2024.pdf>

Jang et al. (2022). Knowledge unlearning for mitigating privacy risks in language models. <https://arxiv.org/abs/2210.01504>. arXiv:2210.01504.

Jia, C. et al (2021). Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. <https://arxiv.org/abs/2102.05918>. arXiv:2102.05918v2

Kahng, M. et al. (2024). LLM Comparator: Visual Analytics for Side-by-Side Evaluation of Large Language Models. <https://arxiv.org/abs/2402.10524>. arXiv:2402.10524v1

Kuchnik, M. et al. (2023). Validating Large Language Models with Realm. <https://arxiv.org/abs/2211.15458>. arXiv:2211.15458v2

Kudo, T., Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. <https://arxiv.org/abs/1808.06226>. arXiv:1808.06226v1

Lam, M. (2018). <https://profiles.stanford.edu/monica-lam?tab=publications>. Keeping the Internet Open with an Open-Source Virtual Assistant.

Lee, C. et al (2024). OrchestraLLM: Efficient Orchestration of Language Models for Dialogue State Tracking. <https://arxiv.org/html/2311.09758v2>. arXiv:2311.09758v2

Lee, J. et al. (2022). Seq2Seq-SC: End-to-End Semantic Communication Systems with Pre-trained Language Model. <https://arxiv.org/abs/2210.15237>. arXiv:2210.15237v2

Lester, B. et al. (2021). The Power of Scale for Parameter-Efficient Prompt Tuning. <https://arxiv.org/abs/2104.08691>. arXiv:2104.08691v2

Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. <https://arxiv.org/abs/2005.11401>

Li, H. et al. (2024). Digger: Detecting Copyright Content Misusage in Large Language Model Training. <https://arxiv.org/abs/2401.00676>. arXiv:2401.00676v1

Li, S. et al (2024). Evaluating Quantized Large Language Models. <https://arxiv.org/abs/2402.18158>. arXiv:2402.18158v1

Li, Y. et al (2023). A Survey on Fairness in Large Language Models. <https://arxiv.org/abs/2308.10149>. arXiv:2308.10149.

Liang, P. et al. (2023). Holistic Evaluation of Language Models. <https://arxiv.org/abs/2211.09110>. arXiv:2211.09110v2

Liu, T. et al (2022). Autoregressive Structured Prediction with Language Models. <https://arxiv.org/abs/2210.14698>. arXiv:2210.14698v2

Liu, Y. et al (2024). Datasets for Large Language Models: A Comprehensive Survey. <https://arxiv.org/abs/2402.18041>. arXiv:2402.18041v1

Liu, Y. et al (2023). Robustness Over Time: Understanding Adversarial Examples' Effectiveness on Longitudinal Versions of Large Language Models. <https://arxiv.org/pdf/2308.07847.pdf>. arXiv:2308.07847v1

Luo, Y. et al. (2023). An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning. <https://arxiv.org/pdf/2308.08747.pdf>. arXiv:2308.08747v3

Management Solutions (2023). Explainable Artificial Intelligence (XAI): desafíos en la interpretabilidad de los modelos. <https://www.managementsolutions.com/en/microsites/whitepapers/explainable-artificial-intelligence>

Management Solutions (2022). AutoML, hacia la automatización de los modelos. <https://www.managementsolutions.com/es/publicaciones-y-eventos/informes-sectoriales/white-papers/auto-machine-learning-hacia-la-automatizacion-de-los-modelos>

Management Solutions (2014). Model Risk Management: Quantitative and Qualitative Aspects. Model Risk Management: Quantitative and qualitative aspects | Management Solutions

Meeus, M. et al. (2024). Copyright Traps for Large Language Models. <https://arxiv.org/abs/2402.09363>. arXiv:2402.09363v1

Mehta, S.V. et al. (2023). An Empirical Investigation of the Role of Pre-training in Lifelong Learning. <https://arxiv.org/abs/2112.09153>. arXiv:2112.09153v2

Mikolov, T. et al. (2013). Efficient Estimation of Word Representations in Vector Space. <https://arxiv.org/abs/1301.3781>. arXiv:1301.3781v3.

Minaee, S. et al. (2024). Large Language Models: A Survey. <https://arxiv.org/abs/2402.06196>. arXiv:2402.06196v2

MindsDB (2024). A Comparative Analysis of Leading Large Language Models. <https://mindsdb.com/blog/navigating-the-llm-landscape-a-comparative-analysis-of-leading-large-language-models>

Mökander, J. et al. (2023). Auditing large language models: a three-layered approach. arXiv:2302.08500v2

Nasr, M., et al. (2023). <https://arxiv.org/pdf/2311.17035.pdf>. arXiv:2311.17035v1

Neelakantan, A. et al. (2022). Text and Code Embeddings by Contrastive Pre-Training. <https://arxiv.org/abs/2201.10005>. arXiv:2201.10005v1

NIST (2023). AI Risk Management Framework | NIST. <https://www.nist.gov/itl/ai-risk-management-framework>

Oneto, L., Chiappa, S. (2020). Fairness in Machine Learning. 2012.15816.pdf (arxiv.org) arXiv:2012.15816v1

OpenAI (2024). Prompt engineering. <https://platform.openai.com/docs/guides/prompt-engineering>

Ovadia, O. et al (2024). Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. <https://arxiv.org/pdf/2312.05934.pdf>. arXiv:2312.05934v3

Pankajakshan, R. et al (2024). Mapping LLM Security Landscapes: A Comprehensive Stakeholder Risk Assessment Proposal. <https://arxiv.org/html/2403.13309v1>. arXiv:2403.13309v1.

Parikh, A. P., et al. (2016). A Decomposable Attention Model for Natural Language Inference. <https://arxiv.org/abs/1606.01933>. arXiv:1606.01933v2

Penedo, G. et al (2023). The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. <https://arxiv.org/abs/2306.01116>. arXiv:2306.01116v1

Pew Research Center (2023). Experts Predict the Best and Worst Changes in Digital Life by 2035.

Project Gutenberg (2024). <https://www.gutenberg.org/>.

Rae, JW, et al (2021). Scaling Language Models: Methods, Analysis & Insights from Training Gopher. <https://arxiv.org/abs/2112.11446>. arXiv:2112.11446

Rafailov, R. et al (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. <https://arxiv.org/abs/2305.18290>. arXiv:2305.18290v2

Rejeleene, R.; Xu, X.; Talburt, J.; (2024). Towards Trustable Language Models: Investigating Information Quality of Large Language Models. <https://arxiv.org/abs/2401.13086>. arXiv:2401.13086v1

Risk.net. (2024). The bank quant who wants to stop gen AI hallucinating. <https://www.risk.net/risk-management/7959062/the-bank-quant-who-wants-to-stop-gen-ai-hallucinating>.

Sachdeva, N., et al (2024). How to Train Data-Efficient LLMs. <https://arxiv.org/html/2402.09668v1>. arXiv:2402.09668v1

Samsi, S., et al (2023). From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference. <https://arxiv.org/pdf/2310.03003.pdf>. arXiv:2310.03003v1

Sarti, G. et al (2023). Inseq: An Interpretability Toolkit for Sequence Generation Models. [2302.13942] Inseq: An Interpretability Toolkit for Sequence Generation Models (arxiv.org). arXiv:2302.13942v3

Searle, J. (1980). Minds, Brains, and Programs. The Behavioral and Brain Sciences, vol. 3. Cambridge University Press. <https://web.archive.org/web/20010221025515/http://www.bbsonline.org/Preprints/OldArchive/bbs.searle2.html>

Shaikh, O. et al. (2022). On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning. <https://arxiv.org/abs/2212.08061>. arXiv:2212.08061v2

SHAP documentation. <https://shap.readthedocs.io/>

Shaw, P. et al (2018). Self-Attention with Relative Position Representations. <https://arxiv.org/abs/1803.02155>. arXiv:1803.02155v2

Sherstinsky, A. (2018). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. <https://arxiv.org/abs/1808.03314>. arXiv:1808.03314v10

Shi, W. et al (2024). Detecting pretraining data from large language models. <https://arxiv.org/abs/2310.16789>. arXiv:2310.16789v3

Singh, C. et al (2024). Rethinking Interpretability in the Era of Large Language Models. <https://arxiv.org/abs/2402.01761>. arXiv:2402.01761v1

Sinha, K. et al (2021). Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little. <https://arxiv.org/abs/2104.06644>. arXiv:2104.06644v2

Soskek (2019). BookCorpus. <https://github.com/soskek/bookcorpus>.

Su, J., et al (2021). Roformer: Enhanced transformer with rotary position embedding. <https://arxiv.org/abs/2104.09864>. arXiv:2104.09864.

Sutskever, I. et al (2014). Sequence to Sequence Learning with Neural Networks. <https://arxiv.org/abs/1409.3215>. arXiv:1409.3215v3

The Next Web (2023). When will AGI arrive? Here's what our tech lords predict. <https://thenextweb.com/news/when-will-agi-arrive-tech-experts-predict-artificial-general-intelligence>

Tian, Y. et al (2024). TinyLLM: Learning a Small Student from Multiple Large Language Models. <https://arxiv.org/abs/2402.04616>. arXiv:2402.04616

Tirumala, K. et al. (2023). D4: Improving LLM Pretraining via Document De-Duplication and Diversification. <https://arxiv.org/abs/2308.12284>. arXiv:2308.12284v1

UK Government (2023). The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023. <https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023>

Vartziotis, T. et al (2024). Learn to Code Sustainably: An Empirical Study on LLM-based Green Code Generation. <https://arxiv.org/html/2403.03344v1>. arXiv:2403.03344v1.

Vaswani, A. et al. (2017). Attention Is All You Need. <https://arxiv.org/abs/1706.03762>

Wan, Z. et al (2024). Efficient Large Language Models: A Survey. <https://arxiv.org/pdf/2312.03863.pdf>. arXiv:2312.03863v3

Wang, Q. et al (2024). LLMCheckup: Conversational Examination of Large Language Models via Interpretability Tools. [2401.12576] LLMCheckup: Conversational Examination of Large Language Models via Interpretability Tools (arxiv.org). arXiv:2401.12576v1

Wang, Y. et al (2024). Two-stage LLM Fine-tuning with Less Specialization and More Generalization. <https://arxiv.org/html/2211.00635v3>. arXiv:2211.00635v3

Wei, J. et al (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. <https://arxiv.org/abs/2201.11903>. arXiv:2201.11903v6

Wenzek, G., et al (2019). CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. <https://arxiv.org/abs/1911.00359>. arXiv:1911.00359v2

Wettig, A. et al. (2024). QuRating: Selecting High-Quality Data for Training Language Models. <https://arxiv.org/abs/2402.09739>. arXiv:2402.09739v1

Weights & Biases: The AI Developer Platform (wandb.ai). <https://wandb.ai/site>

Wikipedia (2024). Dumps. <https://dumps.wikimedia.org/zhwiki/latest/>.

Wired (2023). OpenAI's CEO Says the Age of Giant AI Models Is Already Over. <https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/>

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. <https://dl.acm.org/doi/10.1145/365153.365168>

White House (2022). Blueprint for an AI Bill Of Rights. Making Automated Systems Work for the American People. <https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf>

White House (2023). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>

Wu, X. et al. (2023). Depn: Detecting and editing privacy neurons in pretrained language models. <https://arxiv.org/abs/2310.20138>. arXiv:2310.20138.

Xin Zhao, W., et al. (2023). A Survey of Large Language Models. <https://arxiv.org/abs/2303.18223>. arXiv:2303.18223v13

Xu, L. et al. (2023). Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment. <https://arxiv.org/pdf/2312.12148.pdf>. arXiv:2312.12148v1

Xu, Y. et al. (2021). Non-Autoregressive Text Generation with Pre-trained Language Models. <https://aclanthology.org/2021.eacl-main.18/>

Xu, Z. et al. (2024). Hallucination is Inevitable: An Innate Limitation of Large Language Models. <https://arxiv.org/abs/2401.11817>. arXiv:2401.11817v1

Yang, J. et al. (2024). Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. <https://arxiv.org/abs/2304.13712>. arXiv:2304.13712v2

Yidiz, C. et al (2024). Investigating Continual Pretraining in Large Language Models: Insights and Implications. <https://arxiv.org/html/2402.17400v1>. arXiv:2402.17400v1

Yu, C. et al. (2023). Unlearning bias in language models by partitioning gradients. <https://aclanthology.org/2023.findings-acl.375.pdf>.

Yogarajan, V., et al (2023). Tackling Bias in Pre-trained Language Models: Current Trends and Under-represented Societies. <https://arxiv.org/pdf/2312.01509.pdf>. arXiv:2312.01509v1

Zaharia, M. et al (2018). Accelerating the Machine Learning Lifecycle with MLflow. https://people.eecs.berkeley.edu/~matei/papers/2018/ieee_mlflow.pdf.

Zeng, Y., et al (2023). CLIP2: Contrastive Language-Image-Point Pretraining from Real-World Point Cloud Data. <https://arxiv.org/abs/2303.12417>. arXiv:2303.12417v2

Zhang, B. et al (2024). When Scaling Meets LLM Finetuning: The Effect of Data, Model and Finetuning Method. <https://arxiv.org/abs/2402.17193>. arXiv:2402.17193v1

Zhang, L. et al (2024). Enhancing Large Language Model Performance To Answer Questions and Extract Information More Accurately. <https://arxiv.org/html/2402.01722v1>. arXiv:2402.01722v1.

Zhang, S. et al (2023). Instruction Tuning for Large Language Models: A Survey. https://www.researchgate.net/publication/373263398_Instruction_Tuning_for_Large_Language_Models_A_Survey.

Zhang, Y. et al (2024). Bias Mitigation in Fine-tuning Pre-trained Models for Enhanced Fairness and Efficiency. <https://arxiv.org/html/2403.00625v1>. arXiv:2403.00625v1

Zhao, B., et al (2023). Tuning LayerNorm in Attention: Towards Efficient Multi-Modal LLM Finetuning. <https://arxiv.org/abs/2312.11420>. arXiv:2312.11420v1

Zhou, C. et al (2023). LIMA: Less Is More for Alignment. <https://arxiv.org/abs/2305.11206>. arXiv:2305.11206v1

Zhou, N., et al (2021). Bias, Fairness, and Accountability with AI and ML Algorithms. <https://arxiv.org/abs/2105.06558>. arXiv:2105.06558v1



Nuestro objetivo es superar las expectativas de nuestros clientes convirtiéndonos en socios de confianza

Management Solutions es una firma internacional de servicios de consultoría centrada en el asesoramiento de negocio, finanzas, riesgos, organización y procesos, tanto en sus componentes funcionales como en la implantación de sus tecnologías relacionadas.

Con un equipo multidisciplinar (funcionales, matemáticos, técnicos, etc.) de cerca de 4.000 profesionales, Management Solutions desarrolla su actividad a través de 46 oficinas (20 en Europa, 22 en América, 2 en Asia, 1 en África y 1 en Oceanía).

Para dar cobertura a las necesidades de sus clientes, Management Solutions tiene estructuradas sus prácticas por industrias (Entidades Financieras, Energía, Telecomunicaciones y Otras industrias) y por líneas de actividad que agrupan una amplia gama de competencias: Estrategia, Gestión Comercial y Marketing, Gestión y Control de Riesgos, Información de Gestión y Financiera, Transformación: Organización y Procesos, y Nuevas Tecnologías.

El área de I+D da servicio a los profesionales de Management Solutions y a sus clientes en aspectos cuantitativos necesarios para acometer los proyectos con rigor y excelencia, a través de la aplicación de las mejores prácticas y de la prospección continua de las últimas tendencias en inteligencia artificial, *data science*, *machine learning*, modelización y *big data*.

Javier Calvo Martín

Socio de Management Solutions

javier.calvo.martin@managementsolutions.com

Manuel Ángel Guzmán Caba

Socio de Management Solutions

manuel.guzman@managementsolutions.com

Luz Ferrero Peña

Gerente de Management Solutions

luz.ferrero@msgermany.com.de

Segismundo Jiménez Láinez

Gerente de Management Solutions

segismundo.jimenez@msspain.com

Management Solutions, servicios profesionales de consultoría

Management Solutions es una firma internacional de consultoría centrada en el asesoramiento de negocio, finanzas, riesgos, organización, tecnología y procesos,

Para más información visita www.managementsolutions.com

Síguenos en:     

© **Management Solutions. 2024**
Todos los derechos reservados

www.managementsolutions.com

Madrid Barcelona Bilbao Coruña Málaga London Frankfurt Düsseldorf Paris Amsterdam Copenhagen Oslo Stockholm Warszawa Wrocław Zürich Milano
Roma Bologna Lisboa Beijing Istanbul Johannesburgo Sydney Toronto New York New Jersey Boston Pittsburgh Atlanta Birmingham Houston Miami
SJ de Puerto Rico San José Ciudad de México Monterrey Querétaro Medellín Bogotá Quito São Paulo Rio de Janeiro Lima Santiago de Chile Buenos Aires