

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).

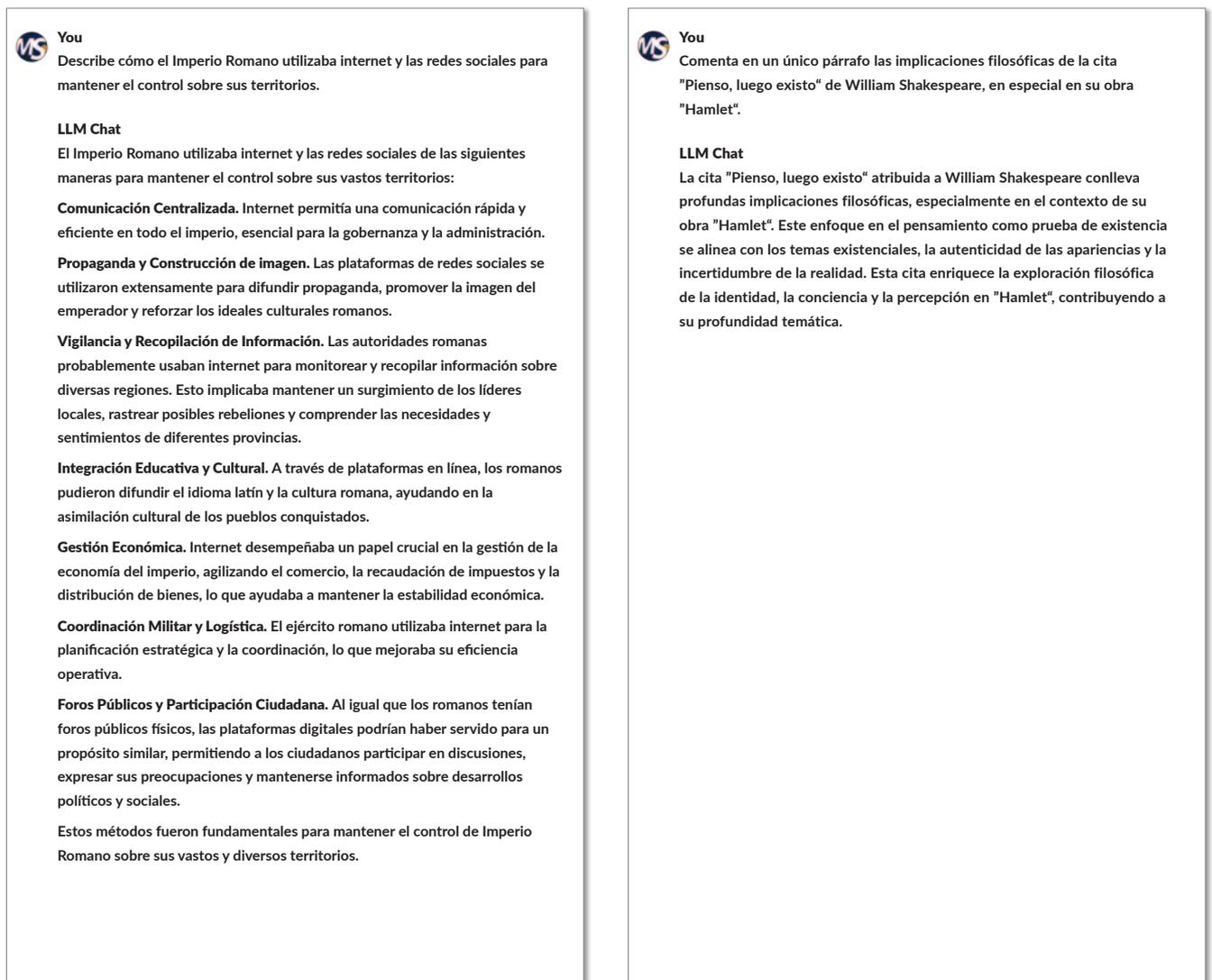
¹⁸European Commission (2021-24).

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.



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.

