KMSnet

# Lo que todo Directivo Hospitalario debe saber sobre la IA generativa

19 de mayo de 2023

**La IA generativa está evolucionando a una velocidad récord, mientras que los directivos aún están aprendiendo el valor agregado y los riesgos de la tecnología. Aquí, ofrecemos algunos de los elementos esenciales de la IA generativa.**

En medio del entusiasmo que rodea a la IA generativa desde el lanzamiento de ChatGPT, Bard, Claude, Midjourney y otras herramientas de creación de contenido, es comprensible que los directores de hospitales, gerentes de clínicas y directivos de salud se pregunten: ¿es esta una exageración tecnológica o una oportunidad revolucionaria? Y si es esto último, ¿cuál es la propuesta de valor para mi negocio?

La versión pública de ChatGPT llegó a 100 millones de usuarios en solo dos meses. Democratizó la IA de una manera nunca antes vista y se convirtió, con mucho, en la aplicación de más rápido crecimiento de la historia. Su accesibilidad lista para usar hace que la IA generativa sea diferente de todas las IA anteriores. Los usuarios no necesitan un título en aprendizaje automático para interactuar o obtener valor de él; casi cualquiera que pueda hacer preguntas puede usarlo. Y, al igual que con otras tecnologías innovadoras como la computadora personal o el iPhone, una plataforma de IA generativa puede dar lugar a muchas aplicaciones para audiencias de cualquier edad o nivel educativo y en cualquier lugar con acceso a Internet.

Todo esto es posible porque los chatbots generativos de IA funcionan con modelos básicos, que son redes neuronales expansivas entrenadas en grandes cantidades de datos no estructurados y sin etiquetar en una variedad de formatos, como texto y audio. Los modelos de base se pueden utilizar para una amplia gama de tareas. Por el contrario, las generaciones anteriores de modelos de IA a menudo eran "estrechas", lo que significa que solo podían realizar una tarea, como predecir la rotación de clientes. Un modelo básico, por ejemplo, puede crear un resumen ejecutivo para un informe técnico de 20.000 palabras sobre computación cuántica, redactar una estrategia de comercialización para un negocio de poda de árboles y proporcionar cinco recetas diferentes para los diez ingredientes en el refrigerador de alguien. La desventaja de tal versatilidad es que, por ahora, la IA generativa a veces puede proporcionar resultados menos precisos.

Con las medidas de protección adecuadas, la IA generativa no solo puede desbloquear nuevos casos de uso para las empresas, sino también acelerar, escalar o mejorar los existentes. Imagine una llamada de ventas de un cliente, por ejemplo. Un modelo de IA especialmente entrenado podría sugerir oportunidades de venta adicionales a un vendedor, pero hasta ahora, generalmente se basaban solo en datos estáticos del cliente obtenidos antes del inicio de la llamada, como la demografía y los patrones de compra. Una herramienta generativa de inteligencia artificial podría sugerir oportunidades de venta adicionales al vendedor en tiempo real según el contenido real de la conversación, a partir de datos internos del cliente, tendencias del mercado externo y datos de personas influyentes en las redes sociales. Al

mismo tiempo, la IA generativa podría ofrecer un primer borrador de un argumento de venta para que el vendedor lo adapte y personalice.

El ejemplo anterior demuestra las implicaciones de la tecnología en un puesto de trabajo. Pero casi todos los trabajadores del conocimiento pueden beneficiarse de asociarse con IA generativa. De hecho, si bien la IA generativa puede eventualmente usarse para automatizar algunas tareas, gran parte de su valor podría derivar de cómo los proveedores de software integran la tecnología en las herramientas cotidianas (por ejemplo, correo electrónico o software de procesamiento de texto) utilizadas por los trabajadores del conocimiento. Estas herramientas mejoradas podrían aumentar sustancialmente la productividad.

Los directores ejecutivos quieren saber si deben actuar ahora y, de ser así, cómo comenzar. Algunos pueden ver la oportunidad de superar a la competencia al volver a imaginar cómo los humanos hacen el trabajo con aplicaciones de IA generativa a su lado. Otros pueden querer tener cuidado, experimentar con algunos casos de uso y aprender más antes de realizar grandes inversiones. Las empresas también tendrán que evaluar si tienen la experiencia técnica, la tecnología y la arquitectura de datos, el modelo operativo y los procesos de gestión de riesgos necesarios que requerirán algunas de las implementaciones más transformadoras de la IA generativa.

El objetivo de este artículo es ayudar a los directivos y gerentes hospitalarios y sus equipos a reflexionar sobre el caso de la creación de valor para la IA generativa y cómo comenzar su viaje. En primer lugar, ofrecemos una introducción a la IA generativa para ayudar a los ejecutivos a comprender mejor el estado de rápida evolución de la IA y las opciones técnicas disponibles. La siguiente sección analiza cómo los hospitales pueden participar en la IA generativa a través de cuatro casos de ejemplo dirigidos a mejorar la eficacia organizacional. Estos casos reflejan lo que estamos viendo entre los primeros usuarios y arrojan luz sobre la variedad de opciones en los requisitos de tecnología, costo y modelo operativo. Finalmente, abordamos el rol vital del director o gerente hospitalario en el posicionamiento de una organización para el éxito con IA generativa.

**Una cartilla generativa de IA**

La tecnología de IA generativa avanza rápidamente (Anexo 1). El ciclo de lanzamiento, la cantidad de nuevas empresas y la rápida integración en las aplicaciones de software existentes son notables. En esta sección, analizaremos la amplitud de las aplicaciones de IA generativa y brindaremos una breve explicación de la tecnología, incluido cómo se diferencia de la IA tradicional.

**Más que un chatbot**

La IA generativa se puede utilizar para automatizar, aumentar y acelerar el trabajo. A los efectos de este artículo, nos centramos en las formas en que la IA generativa puede mejorar el trabajo en lugar de cómo puede reemplazar el papel de los humanos.

Si bien los chatbots generadores de texto, como ChatGPT, han recibido una gran atención, la IA generativa puede habilitar capacidades en una amplia gama de contenido, incluidas imágenes, video, audio y código de computadora. Y puede realizar varias funciones en las organizaciones, incluida la clasificación, edición, resumen, respuesta a preguntas y redacción de contenido nuevo. Cada una de estas acciones tiene el potencial de crear valor al cambiar la forma en que

se realiza el trabajo a nivel de actividad en las funciones comerciales y los flujos de trabajo. Los siguientes son algunos ejemplos.

**Clasificar**

- Un analista de detección de fraudes puede ingresar descripciones de transacciones y documentos de clientes en una herramienta de inteligencia artificial generativa y pedirle que identifique transacciones fraudulentas.

- Un gerente de atención al cliente puede usar IA generativa para categorizar archivos de audio de llamadas de clientes según los niveles de satisfacción de la persona que llama.

**Editar**

- Un redactor puede usar IA generativa para corregir la gramática y convertir un artículo para que coincida con la voz de la marca de un cliente.

- Un diseñador gráfico puede eliminar un logotipo obsoleto de una imagen.

**Resumir**

- Un asistente de producción puede crear un video destacado basado en horas de metraje del evento.

- Un analista de negocios puede crear un diagrama de Venn que resuma los puntos clave de la presentación de un ejecutivo.

**Responder preguntas**

- Los empleados de una empresa de fabricación pueden hacer preguntas técnicas a un "experto virtual" basado en IA generativa sobre los procedimientos operativos.

- Un consumidor puede hacer preguntas a un chatbot sobre cómo ensamblar un nuevo mueble.

**Borrador**

- Un desarrollador de software puede solicitar a la IA generativa que cree líneas completas de código o sugerir formas de completar líneas parciales de código existente.

- Un gerente de marketing puede usar IA generativa para redactar varias versiones de mensajes de campaña.

A medida que la tecnología evoluciona y madura, este tipo de IA generativa se puede integrar cada vez más en los flujos de trabajo empresariales para automatizar tareas y realizar directamente acciones específicas (por ejemplo, enviar automáticamente notas de resumen al final de las reuniones). Ya vemos herramientas emergentes en esta área.

## En qué se diferencia la IA generativa de otros tipos de IA

Como sugiere el nombre, la forma principal en que la IA generativa difiere de las formas anteriores de IA o análisis es que puede generar contenido nuevo, a menudo en formas "no estructuradas" (por ejemplo, texto escrito o imágenes) que no se representan de forma natural en tablas con filas y columnas (ver el "Glosario" para obtener una lista de términos asociados con la IA generativa).

---

# Glosario

**La interfaz de programación de aplicaciones (API)** es una forma de acceder mediante programación (generalmente externos) a modelos, conjuntos de datos u otras piezas de software.

**La inteligencia artificial (IA)** es la capacidad del software para realizar tareas que tradicionalmente requieren inteligencia humana.

**El aprendizaje profundo (Deep Learning)** es un subconjunto del aprendizaje automático que utiliza redes neuronales profundas, que son capas de "neuronas" conectadas cuyas conexiones tienen parámetros o pesos que se pueden entrenar. Es especialmente eficaz para aprender de datos no estructurados, como imágenes, texto y audio.

**El ajuste fino** es el proceso de adaptar un modelo básico previamente entrenado para desempeñarse mejor en una tarea específica. Esto implica un período relativamente corto de entrenamiento en un conjunto de datos etiquetados, que es mucho más pequeño que el conjunto de datos en el que se entrenó inicialmente el modelo. Esta capacitación adicional permite que el modelo aprenda y se adapte a los matices, la terminología y los patrones específicos que se encuentran en el conjunto de datos más pequeño.

**Los modelos Foundation (FM)** son modelos de aprendizaje profundo entrenados en grandes cantidades de datos no estructurados y sin etiquetar que se pueden usar para una amplia gama de tareas listas para usar o adaptarse a tareas específicas a través de ajustes finos. Ejemplos de estos modelos son GPT-4, PaLM, DALL·E 2 y Stable Diffusion.

**La IA generativa** es una IA que normalmente se crea utilizando modelos básicos y tiene capacidades que la IA anterior no tenía, como la capacidad de generar contenido. Los modelos básicos también se pueden usar para fines no generativos (por ejemplo, clasificar la opinión del usuario como negativa o positiva en función de las transcripciones de llamadas) al tiempo que ofrecen una mejora significativa con respecto a los modelos anteriores. Para simplificar, cuando nos referimos a la IA generativa en este artículo, incluimos todos los casos de uso del modelo base.

**Las unidades de procesamiento de gráficos (GPU**) son chips de computadora que se desarrollaron originalmente para producir gráficos de computadora (como para videojuegos) y

también son útiles para aplicaciones de aprendizaje profundo. Por el contrario, el aprendizaje automático tradicional y otros análisis generalmente se ejecutan en unidades centrales de procesamiento (CPU) , normalmente denominadas "procesador" de una computadora.

**Los modelos de lenguaje extenso (LLM)** constituyen una clase de modelos básicos que pueden procesar cantidades masivas de texto no estructurado y aprender las relaciones entre palabras o partes de palabras, conocidas como tokens. Esto permite que los LLM generen texto en lenguaje natural, realizando tareas como resúmenes o extracción de conocimiento. GPT-4 (que subyace a ChatGPT) y LaMDA (el modelo detrás de Bard) son ejemplos de LLM.

**El aprendizaje automático (Machine Learning) (ML)** es un subconjunto de la IA en el que un modelo adquiere capacidades después de entrenarse o mostrarse en muchos puntos de datos de ejemplo. Los algoritmos de aprendizaje automático detectan patrones y aprenden a hacer predicciones y recomendaciones mediante el procesamiento de datos y experiencias, en lugar de recibir instrucciones de programación explícitas. Los algoritmos también se adaptan y pueden volverse más efectivos en respuesta a nuevos datos y experiencias.

**MLOps** se refiere a los patrones y prácticas de ingeniería para escalar y mantener la IA y el ML. Abarca un conjunto de prácticas que abarcan el ciclo de vida completo de ML (gestión de datos, desarrollo, implementación y operaciones en vivo). Muchas de estas prácticas ahora están habilitadas u optimizadas mediante software de soporte (herramientas que ayudan a estandarizar, optimizar o automatizar tareas).

**La ingeniería Prompt** se refiere al proceso de diseñar, refinar y optimizar las indicaciones de entrada para guiar un modelo generativo de IA hacia la producción de los resultados deseados (es decir, precisos).

**Los datos estructurados** son datos tabulares (por ejemplo, organizados en tablas, bases de datos u hojas de cálculo) que se pueden usar para entrenar algunos modelos de aprendizaje automático de manera efectiva.

**Los datos no estructurados** carecen de un formato o estructura coherente (por ejemplo, texto, imágenes y archivos de audio) y, por lo general, requieren técnicas más avanzadas para extraer información.

---

La tecnología subyacente que permite que la IA generativa funcione es una clase de redes neuronales artificiales llamadas modelos básicos. Las redes neuronales artificiales están inspiradas en los miles de millones de neuronas que están conectadas en el cerebro humano. Están entrenados usando aprendizaje profundo, un término que alude a las muchas capas (profundas) dentro de las redes neuronales. El aprendizaje profundo ha impulsado muchos de los avances recientes en IA.

Sin embargo, algunas características distinguen a los modelos básicos de las generaciones anteriores de modelos de aprendizaje profundo (deep learning). Para empezar, pueden

entrenarse en conjuntos extremadamente grandes y variados de datos no estructurados. Por ejemplo, un tipo de modelo básico denominado modelo de lenguaje extenso se puede entrenar con una gran cantidad de texto que está disponible públicamente en Internet y cubre muchos temas diferentes. Mientras que otros modelos de aprendizaje profundo pueden operar con cantidades considerables de datos no estructurados, generalmente se entrenan en un conjunto de datos más específico. Por ejemplo, un modelo puede ser entrenado en un conjunto específico de imágenes para permitirle reconocer ciertos objetos en fotografías.

De hecho, otros modelos de aprendizaje profundo a menudo solo pueden realizar una de esas tareas. Pueden, por ejemplo, clasificar objetos en una foto o realizar otra función, como hacer una predicción. Por el contrario, un modelo de Foundation puede realizar ambas funciones y generar contenido también. Los modelos de Foundation acumulan estas capacidades mediante el aprendizaje de patrones y relaciones a partir de los amplios datos de entrenamiento que incorporan, lo que, por ejemplo, les permite predecir la siguiente palabra en una oración. Así es como ChatGPT puede responder preguntas sobre temas variados y cómo DALL·E 2 y Stable Diffusion pueden producir imágenes basadas en una descripción.

Dada la versatilidad de un modelo básico, las empresas pueden usar el mismo para implementar múltiples casos de uso comercial, algo que rara vez se logra con los modelos anteriores de aprendizaje profundo. Un modelo básico que haya incorporado información sobre los productos de una empresa podría utilizarse potencialmente tanto para responder a las preguntas de los clientes como para ayudar a los ingenieros a desarrollar versiones actualizadas de los productos. Como resultado, las empresas pueden implementar aplicaciones y obtener sus beneficios mucho más rápido.

Sin embargo, debido a la forma en que funcionan los modelos de Foundation actuales, no se adaptan naturalmente a todas las aplicaciones. Por ejemplo, los modelos de lenguaje extenso pueden ser propensos a la "alucinación" o responder preguntas con afirmaciones plausibles pero falsas (consulte la barra lateral "Uso responsable de la IA generativa"). Además, no siempre se proporciona el razonamiento subyacente o las fuentes de una respuesta. Esto significa que las empresas deben tener cuidado de integrar IA generativa sin supervisión humana en aplicaciones donde los errores pueden causar daño o donde se necesita explicación (AI explicable). Actualmente, la IA generativa tampoco es adecuada para analizar directamente grandes cantidades de datos tabulares o resolver problemas avanzados de optimización numérica. Los investigadores están trabajando arduamente para abordar estas limitaciones.

## Uso responsable de la IA generativa

La IA generativa plantea una variedad de riesgos. Los gerentes querrán diseñar sus equipos y procesos para mitigar esos riesgos desde el principio, no solo para cumplir con los requisitos normativos en rápida evolución, sino también para proteger su negocio y ganarse la confianza digital de los consumidores.

Imparcialidad: los modelos pueden generar sesgos algorítmicos debido a datos de entrenamiento imperfectos o decisiones tomadas por los ingenieros que desarrollan los modelos.

Propiedad intelectual (PI): los datos de capacitación y los resultados del modelo pueden generar riesgos significativos de PI, incluida la infracción de materiales protegidos por derechos de autor, marcas registradas, patentados o legalmente protegidos. Incluso cuando se utiliza la

herramienta de IA generativa de un proveedor, las organizaciones deberán comprender qué datos se incluyeron en la capacitación y cómo se utilizan en los resultados de la herramienta.

Privacidad: podrían surgir problemas de privacidad si los usuarios ingresan información que luego termina en los resultados del modelo en una forma que hace que las personas sean identificables. La IA generativa también podría usarse para crear y difundir contenido malicioso, como desinformación, falsificaciones profundas e incitación al odio.

Seguridad: los malos actores pueden utilizar la IA generativa para acelerar la sofisticación y la velocidad de los ataques cibernéticos. También se puede manipular para proporcionar resultados maliciosos. Por ejemplo, a través de una técnica llamada inyección rápida, un tercero le da a un modelo nuevas instrucciones que engañan al modelo para que entregue una salida no deseada por el productor del modelo y el usuario final.

Explicabilidad: la IA generativa se basa en redes neuronales con miles de millones de parámetros, lo que desafía nuestra capacidad para explicar cómo se produce una respuesta dada.

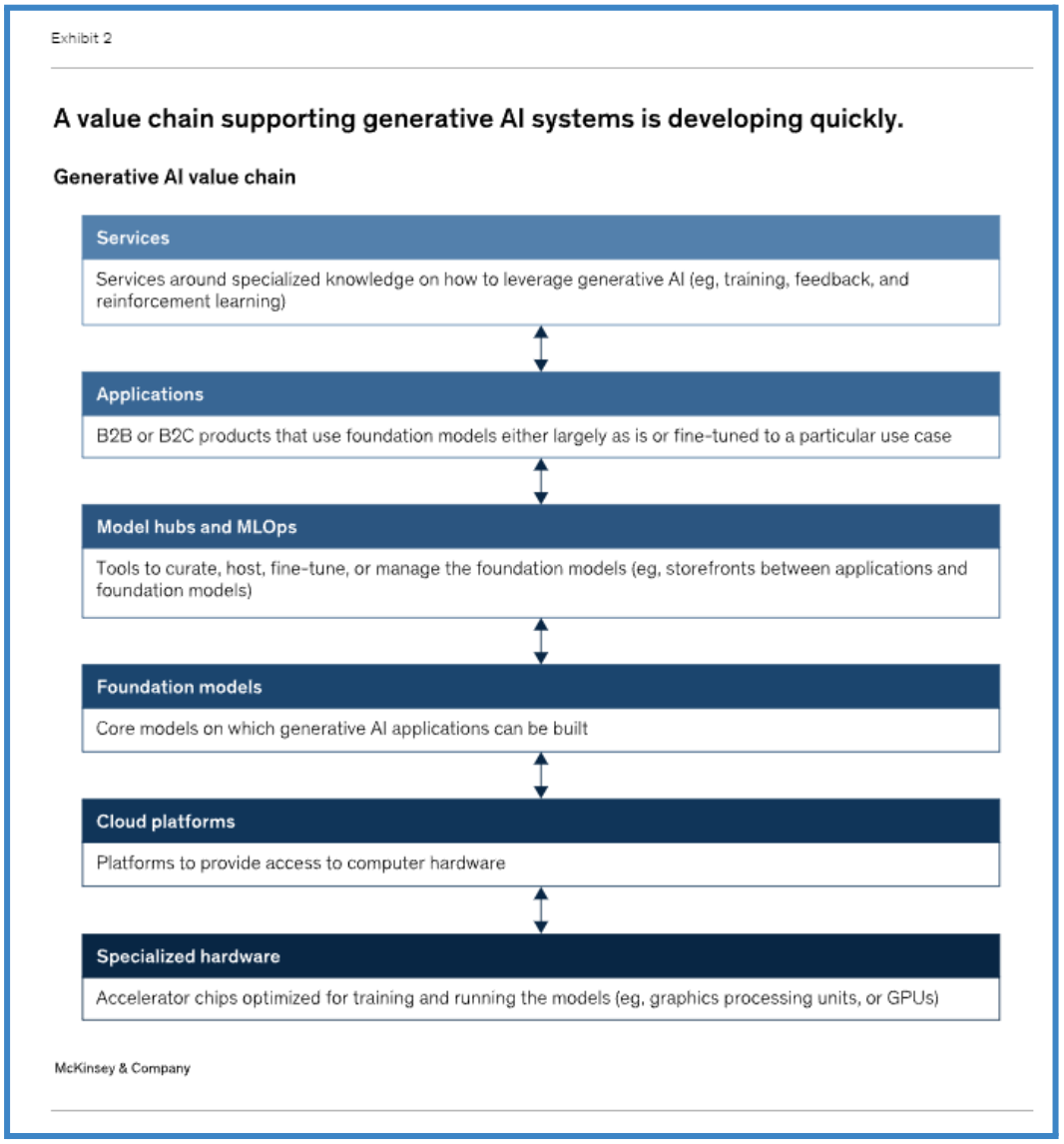
Confiabilidad: los modelos pueden producir diferentes respuestas a las mismas indicaciones, lo que impide que el usuario evalúe la precisión y confiabilidad de los resultados.

Impacto organizacional: la IA generativa puede afectar significativamente a la fuerza laboral, y el impacto en grupos específicos y comunidades locales podría ser desproporcionadamente negativo.

Impacto social y ambiental: el desarrollo y la capacitación de modelos básicos pueden tener consecuencias sociales y ambientales perjudiciales, incluido un aumento de las emisiones de carbono (por ejemplo, la capacitación de un modelo de lenguaje grande puede emitir alrededor de 315 toneladas de dióxido de carbono).

## El ecosistema emergente de IA generativa

Si bien los modelos básicos funcionan como el "cerebro" de la IA generativa, está surgiendo una cadena de valor completa para respaldar la capacitación y el uso de esta tecnología (Exhib




Acá opera KMSnet

El hardware especializado proporciona la amplia potencia informática necesaria para entrenar los modelos. Las plataformas en la nube ofrecen la posibilidad de aprovechar este hardware. Los proveedores de centros de modelos y MLOps[1] ofrecen las herramientas, tecnologías y prácticas que una organización necesita para adaptar un modelo base e implementarlo dentro de sus aplicaciones de usuario final. Muchas empresas están ingresando al mercado para ofrecer aplicaciones construidas sobre modelos básicos que les permitan realizar una tarea específica, como ayudar a los clientes de una empresa con problemas de servicio.

[1] MLOPs: Machine Learning Operations. Es un paradigma que tiene como objetivo implementar y mantener modelos de aprendizaje automático en producción de manera confiable y eficiente.

El desarrollo de los primeros modelos básicos requería altos niveles de inversión, dados los importantes recursos computacionales requeridos para entrenarlos y el esfuerzo humano requerido para refinarlos. Como resultado, fueron desarrollados principalmente por unos pocos gigantes tecnológicos, nuevas empresas respaldadas por una inversión significativa y algunos colectivos de investigación de código abierto (por ejemplo, BigScience). Sin embargo, se está trabajando tanto en modelos más pequeños que pueden ofrecer resultados efectivos para algunas tareas como en capacitación que es más eficiente. Esto eventualmente podría abrir el mercado a más participantes. Algunas empresas emergentes ya han logrado desarrollar sus propios modelos; por ejemplo, Cohere, Anthropic y AI21 Labs construyen y entrenan sus propios modelos de lenguaje grande.

## Poner a trabajar la IA generativa

Los directores ejecutivos y gerentes deberían considerar que la exploración de la IA generativa es un deber, no un tal vez. La IA generativa puede crear valor en una amplia gama de casos de uso. Los requisitos económicos y técnicos para comenzar no son prohibitivos, mientras que la desventaja de la inacción podría ser quedarse rápidamente atrás de los competidores. Cada CEO debe trabajar con el equipo ejecutivo para reflexionar sobre dónde y cómo jugar. Algunos directores ejecutivos pueden decidir que la IA generativa presenta una oportunidad transformadora para sus empresas, ofreciendo la oportunidad de reinventar todo, desde investigación y desarrollo hasta marketing y ventas y operaciones con clientes. Otros pueden optar por comenzar de a poco y escalar más adelante. Una vez que se toma la decisión, existen vías técnicas que los expertos en IA pueden seguir para ejecutar la estrategia, según el caso de uso.

Gran parte del uso (aunque no necesariamente todo el valor) de la IA generativa en una organización provendrá de los trabajadores que empleen funciones integradas en el software que ya tienen. Los sistemas de correo electrónico brindarán una opción para escribir los primeros borradores de los mensajes. Las aplicaciones de productividad crearán el primer borrador de una presentación basada en una descripción. El software financiero generará una descripción en prosa de las características notables en un informe financiero. Los sistemas de gestión de relaciones con los clientes sugerirán formas de interactuar con los clientes. Estas características podrían acelerar la productividad de todos los trabajadores del conocimiento.

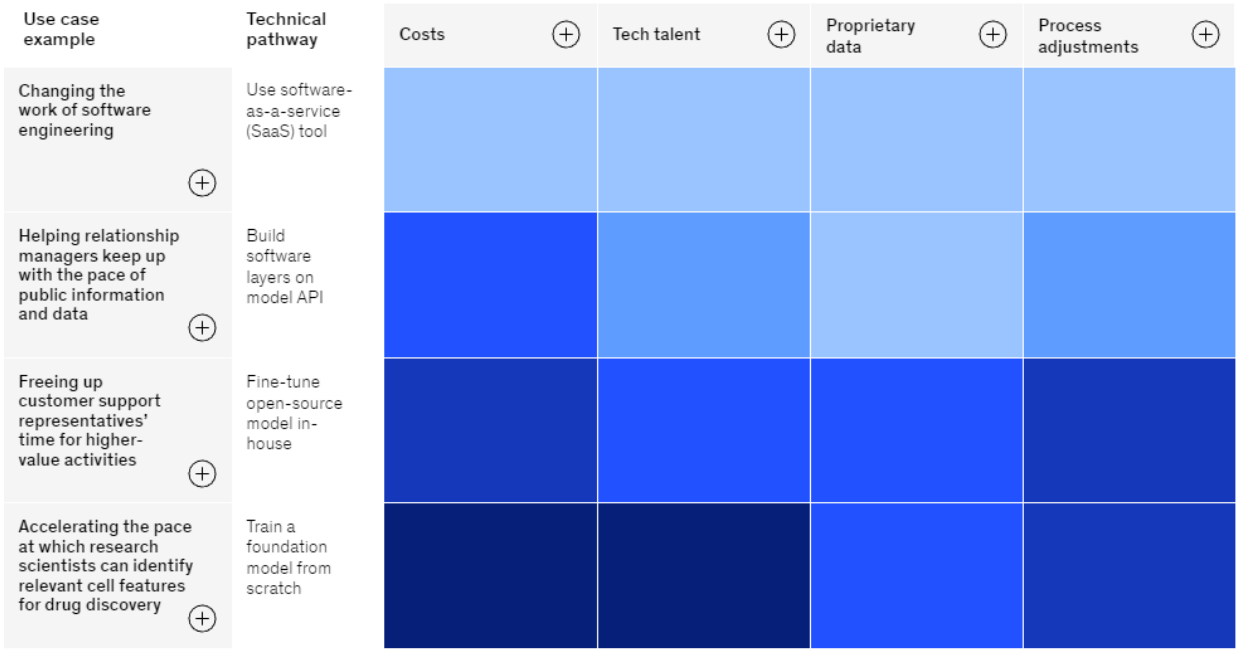
Pero la IA generativa también puede ser más transformadora en ciertos casos de uso. A continuación, analizamos cuatro ejemplos de cómo las empresas de diferentes industrias están utilizando la IA generativa en la actualidad para remodelar la forma en que se realiza el trabajo dentro de su organización. Los ejemplos van desde aquellos que requieren recursos mínimos hasta empresas que requieren muchos recursos. (Para una comparación rápida de estos ejemplos y más detalles técnicos, consulte el Exhibit 3).

Exhibit 3

**The organizational requirements for generative AI range from low to high, depending on the use case.**

Click a row or column header for more

Low High




McKinsey & Company

### 1) Cambiando el trabajo de ingeniería de software

El primer ejemplo es un caso de complejidad relativamente baja con beneficios de productividad inmediatos porque utiliza una solución de IA generativa lista para usar y no requiere personalización interna.

La mayor parte del trabajo de un ingeniero de software es escribir código. Es un proceso laborioso que requiere una extensa prueba y error e investigación en documentación privada y pública. En esta empresa, la escasez de ingenieros de software capacitados ha provocado una gran acumulación de solicitudes de funciones y corrección de errores.

Para mejorar la productividad de los ingenieros, la empresa está implementando un producto de finalización de código basado en IA que se integra con el software que usan los ingenieros para codificar. Esto permite a los ingenieros escribir descripciones de código en lenguaje natural, mientras que la IA sugiere varias variantes de bloques de código que satisfarán la descripción. Los ingenieros pueden seleccionar una de las propuestas de la IA, realizar los ajustes necesarios y hacer clic en ella para insertar el código.

Nuestra investigación ha demostrado que tales herramientas pueden acelerar la generación de código de un desarrollador hasta en un 50 por ciento. También puede ayudar en la depuración, lo que puede mejorar la calidad del producto desarrollado. Pero hoy, la IA generativa no puede reemplazar a los ingenieros de software calificados. De hecho, los ingenieros con más experiencia parecen obtener los mayores beneficios de productividad de las herramientas, mientras que los desarrolladores sin experiencia ven resultados menos impresionantes y, a veces, negativos. Un riesgo conocido es que el código generado por IA puede contener vulnerabilidades u otros errores, por lo que los ingenieros de software deben participar para garantizar la calidad y la seguridad del código (consulte la sección final de este artículo para conocer las formas de mitigar los riesgos).

El costo de esta herramienta de codificación de IA generativa lista para usar es relativamente bajo, y el tiempo de comercialización es corto porque el producto está disponible y no requiere un desarrollo interno significativo. El costo varía según el proveedor de software, pero las suscripciones de tarifa fija oscilan entre $10 y $30 por usuario por mes. Al elegir una herramienta, es importante discutir los problemas de licencia y propiedad intelectual con el proveedor para asegurarse de que el código generado no genere infracciones.

El soporte de la nueva herramienta es un pequeño equipo multifuncional centrado en seleccionar el proveedor de software y monitorear el rendimiento, lo que debe incluir la verificación de problemas de propiedad intelectual y seguridad. La implementación sólo requiere cambios en el flujo de trabajo y en las políticas. Debido a que la herramienta es puramente software como servicio (SaaS), los costos adicionales de computación y almacenamiento son mínimos o inexistentes.

## 2) Ayudar a los gerentes de relaciones a mantenerse al día con la información y los datos públicos

Las empresas pueden decidir crear sus propias aplicaciones de IA generativa, aprovechando los modelos básicos (a través de API o modelos abiertos), en lugar de utilizar una herramienta estándar. Esto requiere un aumento en la inversión con respecto al ejemplo anterior, pero facilita un enfoque más personalizado para satisfacer el contexto y las necesidades específicas de la empresa.

En este ejemplo, un gran banco corporativo quiere usar IA generativa para mejorar la productividad de los gerentes de relaciones (RM). Los RM pasan un tiempo considerable revisando documentos grandes, como informes anuales y transcripciones de llamadas de ganancias, para mantenerse informados sobre la situación y las prioridades de un cliente. Esto permite que el RM ofrezca servicios adecuados a las necesidades particulares del cliente.

El banco decidió crear una solución que acceda a un modelo base a través de una API. La solución escanea documentos y puede proporcionar rápidamente respuestas sintetizadas a las preguntas planteadas por los RM. Se crean capas adicionales alrededor del modelo base para optimizar la experiencia del usuario, integrar la herramienta con los sistemas de la empresa y aplicar controles de riesgo y cumplimiento. En particular, los resultados del modelo deben verificarse, de la misma manera que una organización verificaría los resultados de un analista junior, porque se sabe que algunos modelos de lenguaje grandes tienen alucinaciones. Los RM también están capacitados para hacer preguntas de una manera que proporcione las respuestas más precisas de la solución (lo que se denomina ingeniería rápida), y se implementan procesos para agilizar la validación de los resultados de la herramienta y las fuentes de información.

En este caso, la IA generativa puede acelerar el proceso de análisis de un RM (de días a horas), mejorar la satisfacción laboral y, potencialmente, capturar información que el RM podría haber pasado por alto.

El costo de desarrollo proviene principalmente de la construcción de la interfaz de usuario y las integraciones, que requieren tiempo de un científico de datos, un ingeniero de aprendizaje automático o un ingeniero de datos, un diseñador y un desarrollador front-end. Los gastos continuos incluyen el mantenimiento del software y el costo de usar las API. Los costos dependen de la elección del modelo y de las tarifas del proveedor externo, el tamaño del equipo y el tiempo hasta el producto mínimo viable.

### 3) Liberar a los representantes de atención al cliente para actividades de mayor valor.

El siguiente nivel de sofisticación es el ajuste fino de un modelo básico. En este ejemplo, una empresa utiliza un modelo básico optimizado para conversaciones y lo ajusta en sus propios chats de clientes de alta calidad y preguntas y respuestas específicas del sector. La empresa opera en un sector con terminología especializada (por ejemplo, derecho, medicina, bienes raíces y finanzas). El servicio al cliente rápido es un diferenciador competitivo.

Los representantes de atención al cliente de esta empresa manejan cientos de consultas al día. Los tiempos de respuesta a veces eran demasiado altos, lo que provocaba la insatisfacción de los usuarios. La empresa decidió introducir un bot de servicio al cliente de inteligencia artificial generativa para manejar la mayoría de las solicitudes de los clientes. El objetivo era dar una respuesta rápida en un tono que coincidiera con la marca de la empresa y las preferencias del cliente. Parte del proceso de ajuste y prueba del modelo base incluye garantizar que las respuestas estén alineadas con el lenguaje específico del dominio, la promesa de la marca y el tono establecido para la empresa; se requiere un monitoreo continuo para verificar el desempeño del sistema en múltiples dimensiones, incluida la satisfacción del cliente.

La empresa creó una hoja de ruta del producto que consta de varias oleadas para minimizar los posibles errores del modelo. En la primera ola, el chatbot se puso a prueba internamente. Los empleados pudieron dar respuestas positivas o negativas a las sugerencias del modelo, y el modelo pudo aprender de estas entradas. Como siguiente paso, el modelo "escuchó" las conversaciones de atención al cliente y ofreció sugerencias. Una vez que la tecnología se probó lo suficiente, comenzó la segunda ola y el modelo se cambió hacia casos de uso orientados al cliente con un ser humano en el circuito. Eventualmente, cuando los líderes confían completamente en la tecnología, se puede automatizar en gran medida.

En este caso, la IA generativa liberó a los representantes de servicio para que se concentran en consultas de clientes complejas y de mayor valor, mejoró la eficiencia y la satisfacción laboral de los representantes y aumentó los estándares de servicio y la satisfacción del cliente. El bot tiene acceso a todos los datos internos del cliente y puede "recordar" conversaciones anteriores (incluidas las llamadas telefónicas), lo que representa un cambio radical con respecto a los chatbots de clientes actuales.
Para capturar los beneficios, este caso de uso requirió inversiones materiales en software, infraestructura en la nube y talento tecnológico, así como mayores grados de coordinación interna en riesgo y operaciones. En general, ajustar los modelos básicos cuesta de dos a tres veces más que crear una o más capas de software sobre una API. Los costos de talento y de terceros para la computación en la nube (si se ajusta un modelo auto hospedado) o para la API (si se ajusta a través de una API de terceros) representan el aumento de los costos. Para implementar la solución, la empresa necesitó la ayuda de expertos en DataOps y MLOps, así como el aporte de otras funciones, como especialistas en administración de productos, diseño, legal y servicio al cliente.

## 4) Acelerando el descubrimiento de drogas.

Los casos de uso de IA generativa más complejos y personalizados surgen cuando no hay disponibles modelos básicos adecuados y la empresa necesita construir uno desde cero. Esta situación puede surgir en sectores especializados o al trabajar con conjuntos de datos únicos que son significativamente diferentes de los datos utilizados para entrenar modelos básicos existentes, como lo demuestra este ejemplo farmacéutico. Entrenar un modelo de base desde cero presenta importantes desafíos técnicos, de ingeniería y de recursos. El retorno adicional de la inversión por usar un modelo de mayor rendimiento debería superar los costos financieros y de capital humano.

En este ejemplo, los científicos de investigación en el descubrimiento de fármacos en una empresa farmacéutica tenían que decidir qué experimentos ejecutar a continuación, basándose en imágenes de microscopía. Tenían un conjunto de datos de millones de estas imágenes, que contenían una gran cantidad de información visual sobre las características de las células que son relevantes para el descubrimiento de fármacos pero difíciles de interpretar para un ser humano. Las imágenes se utilizaron para evaluar posibles candidatos terapéuticos.

La empresa decidió crear una herramienta que ayudaría a los científicos a comprender la relación entre la química de los fármacos y los resultados microscópicos registrados para acelerar los esfuerzos de I+D. Dado que estos modelos multimodales aún están en pañales, la empresa decidió entrenar los suyos propios. Para construir el modelo, los miembros del equipo emplearon imágenes del mundo real que se usan para entrenar modelos básicos basados en imágenes y su gran conjunto de datos de imágenes de microscopía interna.

El modelo entrenado agregó valor al predecir qué fármacos candidatos podrían conducir a resultados favorables y al mejorar la capacidad de identificar con precisión las características celulares relevantes para el descubrimiento de fármacos. Esto puede conducir a procesos de descubrimiento de fármacos más eficientes y efectivos, no solo mejorando el tiempo de obtención de valor, sino también reduciendo la cantidad de análisis inexactos, engañosos o fallidos.

En general, entrenar un modelo desde cero cuesta entre diez y veinte veces más que crear software en torno a una API modelo. Los equipos más grandes (incluidos, por ejemplo, expertos en aprendizaje automático con doctorado) y un mayor gasto en computación y almacenamiento explican las diferencias en el costo. El costo proyectado de entrenar un modelo base varía ampliamente según el nivel de rendimiento del modelo deseado y la complejidad del modelado. Esos factores influyen en el tamaño requerido del conjunto de datos, la composición del equipo y los recursos informáticos. En este caso de uso, el equipo de ingeniería y los gastos continuos de la nube representaron la mayoría de los costos.

La empresa descubrió que se necesitan actualizaciones importantes en su infraestructura y procesos tecnológicos, incluido el acceso a muchas instancias de GPU para entrenar el modelo, herramientas para distribuir la capacitación en muchos sistemas y MLOps de mejores prácticas para limitar el costo y la duración del proyecto. Además, se requirió un trabajo sustancial de procesamiento de datos para la recopilación, la integración (garantizar que los archivos de diferentes conjuntos de datos tengan el mismo formato y resolución) y la limpieza (filtrar datos de baja calidad, eliminar duplicados y garantizar que la distribución esté en línea con el objetivo previsto). usar). Dado que el modelo básico se entrenó desde cero, se necesitaban pruebas rigurosas del modelo final para garantizar que la salida fuera precisa y segura de usar.

## Lecciones que los gerentes y directores ejecutivos pueden sacar de estos ejemplos

Los casos de uso descritos aquí ofrecen importantes conclusiones para los gerentes y directores ejecutivos a medida que se embarcan en el viaje de la IA generativa:

- Ya existen casos de uso transformadores que ofrecen beneficios prácticos para los trabajos y el lugar de trabajo. Las empresas de todos los sectores, desde el farmacéutico hasta la banca y el comercio minorista, están presentando una variedad de casos de uso para capturar el potencial de creación de valor. Las organizaciones pueden comenzar pequeñas o grandes, según sus aspiraciones.

- Los costos de buscar IA generativa varían ampliamente, según el caso de uso y los datos necesarios para el software, la infraestructura de la nube, la experiencia técnica y la mitigación de riesgos. Las empresas deben tener en cuenta los problemas de riesgo, independientemente del caso de uso, y algunas requerirán más recursos que otras.

- Si bien es bueno comenzar rápido, construir antes un caso de negocios básico ayudará a las empresas a navegar mejor en sus viajes generativos de IA.