Respuestas a las 4 preguntas más importantes sobre la seguridad de la IA generativa

Adopte rápidamente la IA generativa y, al mismo tiempo, ayude a garantizar la seguridad, la privacidad y el cumplimiento

Este libro electrónico va dirigido a los líderes empresariales, en particular a los responsables de la toma de decisiones de TI y a los jefes de equipos de seguridad, que tienen previsto o están contemplando cómo integrar de forma segura la IA generativa en sus organizaciones.

Índice

INTRODUCCIÓN

Preparado, listo, genere: adopte la IA generativa de forma rápida y segura

La carrera por la IA generativa ha comenzado. Las empresas se están volcando en reinventar las aplicaciones y las experiencias de los clientes, animadas por posibles mejoras masivas en la productividad y la experiencia.

Si bien la era de la inteligencia artificial (IA) generativa no ha hecho más que empezar, las organizaciones ya perciben beneficios tangibles en prácticamente todas las unidades empresariales. Sin embargo, los profesionales de la seguridad recomiendan cautela. Citan la privacidad de los datos, el sesgo de los modelos, la creación de contenidos dañinos (como los *deepfakes*) y los riesgos de la introducción de datos maliciosos en los modelos como razones para abordar la adopción de la IA generativa con precaución.

Es imperativo que las organizaciones aborden la IA generativa con una estrategia clara sobre cómo proteger sus datos, usuarios y reputación, al tiempo que permiten una rápida adopción y mejoran la experiencia del cliente.

Si bien esto representa un desafío multifacético, las organizaciones deben recordar que aún se aplican las prácticas recomendadas estándares para la inteligencia artificial (IA), el machine learning (ML), la protección de datos y la seguridad de las cargas de trabajo en la nube. De hecho, es posible que su organización esté mejor preparada de lo que cree para proteger la IA generativa.

Establecer ahora las protecciones adecuadas para las cargas de trabajo de la IA generativa ayudará a impulsar la innovación en toda la organización, lo que dará a sus equipos la confianza necesaria para perseguir grandes ideas y la libertad de centrarse en hacer crecer su empresa.

En este libro electrónico, se analizarán cuatro preguntas clave que debe plantearse al iniciar su andadura hacia cargas de trabajo de IA generativa más seguras.

1. **¿Qué necesita proteger?**
2. **¿Cómo abordar los problemas de cumplimiento?**
3. **¿Cómo garantizar que los modelos funcionen según lo previsto?**
4. **¿Por dónde empezar?**

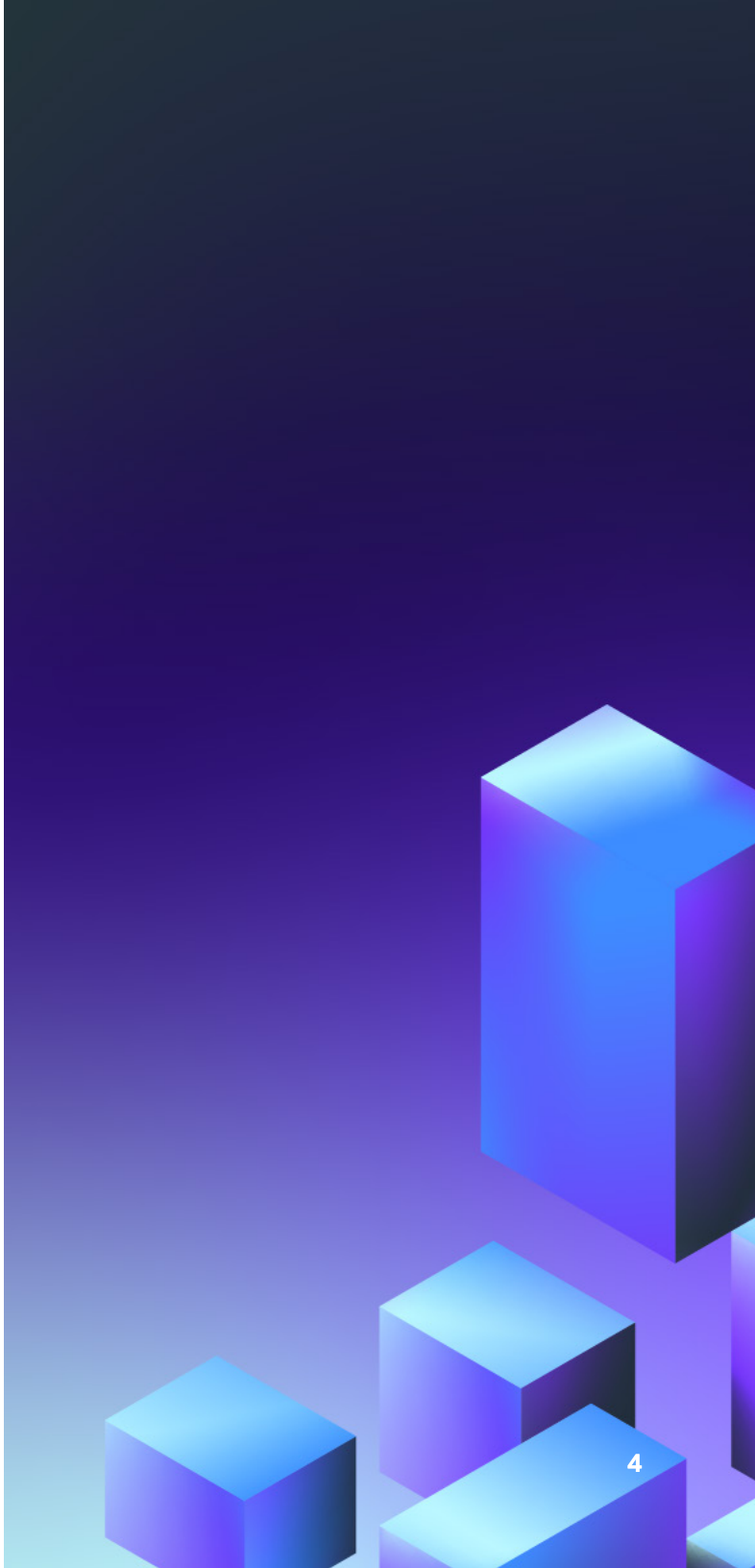
REQUISITOS DE PROTECCIÓN DE DATOS

Pregunta 1:

¿Qué necesita proteger?

Antes de poder desarrollar y desplegar aplicaciones de IA generativa de forma segura, es importante comprender qué es exactamente lo que necesita protección. Resulta útil agrupar estas actividades en tres categorías:

- **Protección de las cargas de trabajo en la nube**
- **Protección de los datos**
- **Protección de las aplicaciones de IA generativa**



Protección de las cargas de trabajo en la nube

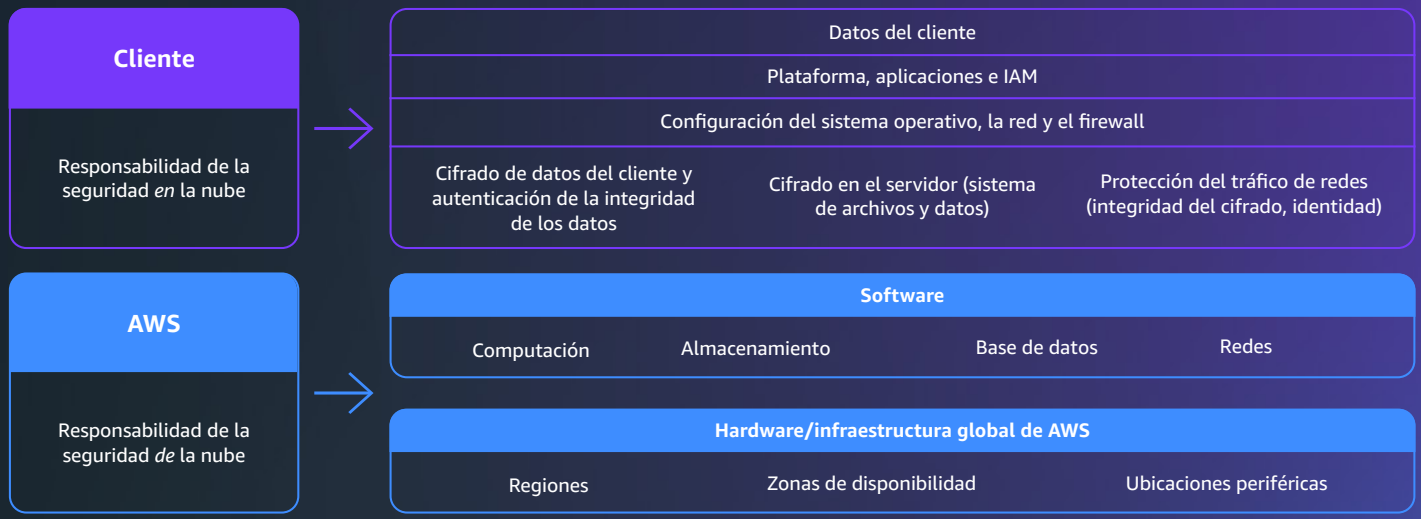
El uso de la IA generativa para alcanzar los objetivos de seguridad y privacidad comienza con la protección de la infraestructura, los servicios y las configuraciones generales de la nube. Para ello, primero tendrá que distinguir sus responsabilidades de seguridad de las responsabilidades de las que se encarga su proveedor de servicios en la nube.

Los clientes de Amazon Web Services (AWS) pueden consultar el **modelo de responsabilidad compartida** para obtener orientación en este ámbito. Explica que, en términos generales, AWS es responsable de operar, administrar y controlar la infraestructura que ejecuta todos los servicios ofrecidos en la nube de AWS, lo que se conoce como «seguridad *de* la nube».

Por otra parte, los clientes de AWS son responsables de administrar el sistema operativo invitado (incluidas las actualizaciones y los parches de seguridad) y cualquier otro software de aplicaciones asociado y de la configuración del firewall del grupo de seguridad que ofrece AWS. El alcance y las tareas concretas en manos de los clientes dependen de los servicios de AWS que decidan utilizar. Esto se llama «seguridad *en* la nube».

Si bien la popularidad de la IA generativa es algo nuevo, las prácticas recomendadas de seguridad tradicionales aún son un punto de partida útil. Esto incluye prácticas básicas de higiene de seguridad para:

- Identity and Access Management (IAM)
- Detección y respuesta
- Protección de la infraestructura
- Protección de datos
- Seguridad de las aplicaciones



Protección de los datos

A continuación, tendrá que ayudar a garantizar la seguridad y la privacidad de los datos que utilizan sus aplicaciones de IA generativa. Esto puede incluir información sujeta a derechos de propiedad, propiedad intelectual (IP) valiosa e información de identificación personal (PII).

Las aplicaciones de IA generativa funcionan con modelos fundacionales (FM), que se entrenan con grandes cantidades de datos. Los FM analizan estos datos para identificar patrones y aprender a generar contenido nuevo y similar. A la hora de crear aplicaciones de IA generativa que cumplan con requisitos empresariales específicos, normalmente tendrá que personalizar un FM existente y entrenarlo con los datos de su organización.

Para ayudar a proteger estos datos, tendrá que tener en cuenta los controles de privacidad de datos y las prácticas recomendadas de la política de IAM.

Cuando personalice un FM, asegúrese de que los equipos empleen una versión del modelo que esté almacenada de forma segura y que no se utilice para mejorar el FM en sí. Al configurar una capacidad dedicada de un solo inquilino en **Amazon Bedrock**, el servicio puede adjuntar sus instancias de inferencia a su **Amazon Virtual Private Cloud (Amazon VPC)** para leer y escribir en **Amazon Simple Storage Service (Amazon S3)**.

Una IAM eficaz ayuda a validar que las personas y máquinas adecuadas tengan acceso a los recursos correctos en las condiciones apropiadas. El **AWS Well-Architected Framework** describe los principios de diseño y las prácticas recomendadas de arquitectura para ayudar a administrar las identidades. Este recurso es una herramienta útil con la que elaborar políticas de IAM y abordar otros problemas de seguridad, como la detección de amenazas y la seguridad de la red.

Protección de las aplicaciones de IA generativa

Con objeto de proteger la IA generativa a nivel de aplicación, debe identificar, clasificar, corregir y mitigar los riesgos de forma continua. Un primer paso es implementar las prácticas recomendadas actuales para mantener seguros los entornos y los datos.

A partir de ahí, debe plantearse cómo trasladar la seguridad a una etapa más temprana del proceso de desarrollo. De este modo, puede agilizar sus esfuerzos y permitir a los equipos de desarrollo innovar más rápido y con mayor libertad, al tiempo que evita convertir la seguridad en un cuello de botella.

A continuación, debe considerar cómo proteger los tres componentes cruciales de toda aplicación de IA: las entradas, las salidas y el propio modelo.

Protección de las entradas

Comience por revisar los datos que ingresan en su sistema de IA. Los usuarios no deben tener acceso directo al FM sin filtro de entrada a fin de reducir el riesgo de ataques a la integridad, como la manipulación, la suplantación de identidad o la inyección de solicitudes. Estas técnicas de ataque eluden los controles o abusan del modelo. Otras estrategias que hay que tener en cuenta para proteger las entradas son la automatización de la calidad de los datos, la supervisión continua y el modelado de amenazas.

Protección de las salidas

Los riesgos para las salidas de las aplicaciones de IA generativa incluyen la divulgación de información, los incidentes de propiedad intelectual y el uso indebido o abuso del modelo, que pueden dañar la reputación de la organización. Al desarrollar su modelo de amenazas, tenga en cuenta la huella de información y el contexto de uso e incluya la detección y la supervisión de comportamientos complejos.

Protección del modelo en sí

Por último, estudie cómo los adversarios podrían intentar eliminar datos del propio modelo o de sus componentes asociados. Los riesgos incluyen tergiversaciones de la vida real o de los datos del modelo y daños a la integridad o disponibilidad del modelo. Modele las amenazas para sus objetivos empresariales e implemente la supervisión de estos escenarios de amenazas.

REQUISITOS DE CONFORMIDAD

Pregunta 2:

¿Cómo abordar los problemas de cumplimiento?

Al mitigar los riesgos de diseñar y desarrollar aplicaciones de IA generativa, su organización puede generar confianza en sus socios y clientes, mantener la reputación de la marca y acatar en todo momento los requisitos de cumplimiento.

La regulación legislativa de las aplicaciones de IA generativa todavía se encuentra en sus primeras etapas y aún no hay consenso sobre las prácticas recomendadas. Por consiguiente, surcar el laberinto de estándares contradictorios y supervisión en diferentes jurisdicciones presenta un desafío complejo y continuo.

Póngase en contacto con sus asesores legales y expertos en privacidad para evaluar los requisitos y las implicaciones de crear su aplicación de IA generativa. Podría tener que examinar sus derechos legales en cuanto al uso de datos y modelos específicos y determinar la aplicabilidad de las leyes en materia de privacidad, biometría, antidiscriminación y otras regulaciones específicas de los casos de uso.

Tenga en cuenta los diferentes requisitos legales en los estados, provincias y países, así como las nuevas regulaciones de IA que se están proponiendo en todo el mundo. Examine estas consideraciones en las etapas operativas y de despliegue futuras.

La colaboración con compañeros, expertos en IA y organizaciones gubernamentales también puede ayudarle a mantener el cumplimiento y, al mismo tiempo, demostrar a los clientes que se toma en serio las normas legales y éticas sobre la IA. Recientemente, Amazon se sumó a la Casa Blanca y a seis empresas líderes en inteligencia artificial **para comprometerse voluntariamente con el desarrollo responsable y seguro de la IA**, lo que demuestra el valor de estos compromisos y sienta las bases para una colaboración futura.

Riesgos inherentes de la inteligencia artificial

Como ocurre con todas las soluciones que utilizan ML, las aplicaciones de IA generativa conllevan riesgos que van más allá de los del software tradicional. Para crear y desplegar aplicaciones de forma segura con la IA generativa, tendrá que abordar y desarrollar estrategias para mitigar estos riesgos, que incluyen:

- Salidas sesgadas, falsas, engañosas, dañinas u ofensivas
- Complejidades y costes a escala
- Conjuntos de datos que crecen demasiado, quedan obsoletos o se alejan del contexto previsto
- Preocupaciones por el aumento de la opacidad y la reproducibilidad
- Normas y procedimientos de prueba poco desarrollados

En la siguiente sección, expondremos las estrategias generales destinadas a reducir algunos de estos riesgos y las prácticas recomendadas para definir las repercusiones profesionales, organizativas y sociales de sus aplicaciones de IA generativa.

VISIBILIDAD DEL COMPORTAMIENTO DEL MODELO

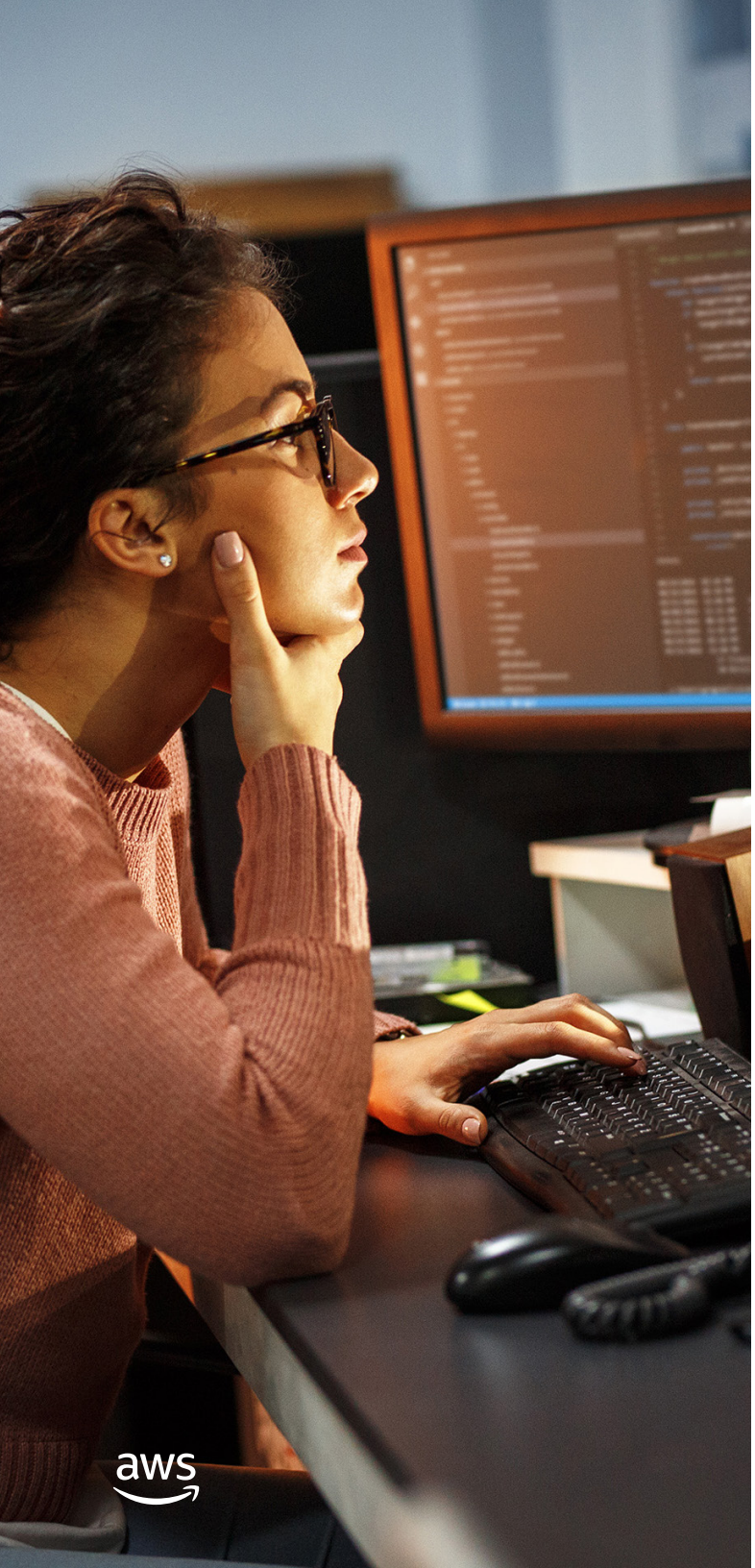
Pregunta 3:

¿Cómo garantizar que los modelos funcionen según lo previsto?

Garantizar el uso responsable de la IA generativa se ha convertido en una tarea empresarial esencial y en un catalizador fundamental de la innovación continua.

Los FM se entrenan con conjuntos de datos masivos y llevan a cabo análisis complejos que les ayudan a comprender cómo generar contenido similar. Si bien muchos FM ofrecen resultados notables, todavía se aplica el antiguo dicho de «entra basura, sale basura» o GIGO. Si un FM recibe datos inexactos, incompletos o sesgados, sus salidas pueden mostrar defectos similares.

Los datos erróneos abren la puerta a oportunidades de uso indebido, acciones maliciosas y otros riesgos. A medida que su aplicación de IA generativa aumenta en usuarios, alcance y función, mayor es la posible repercusión de estos problemas.



Fomentar una IA responsable

Comprometerse con una estrategia de IA responsable le ayudará a abordar estos riesgos. Las dimensiones de la IA responsable incluyen la explicabilidad, la equidad, la gobernanza, la privacidad, la seguridad, la solidez y la transparencia. También supone comprender las formas en que la aplicación ve, trata y afecta a las diferentes culturas y grupos demográficos.

Lo mejor es comenzar a considerar la IA responsable desde el principio de su andadura hacia la IA generativa y, luego, no dejar nunca de lado este aspecto durante todo el ciclo de vida de la aplicación como una parte clave de su visión. Empiece con acciones relativamente pequeñas y sencillas. A continuación, escale la forma en que la IA responsable afecta al diseño, el desarrollo y las operaciones a lo largo del tiempo.

En la redacción de políticas responsables de IA y gobernanza, tenga en cuenta cómo afectará su aplicación de IA generativa a sus usuarios, clientes y empleados, así como a la sociedad. Asegúrese de tratar la equidad algorítmica, la representación diversa e inclusiva y la detección de sesgos.

Combatir la toxicidad

La toxicidad en los modelos de lenguaje de gran tamaño (LLM) se refiere a la generación de texto grosero, irrespetuoso o irrazonable. Existen muchas estrategias para ayudar a prevenir la toxicidad y garantizar la imparcialidad en las aplicaciones de IA generativa. Por ejemplo, es posible identificar y eliminar el lenguaje ofensivo o las frases sesgadas de los datos de entrenamiento. También puede realizar pruebas de imparcialidad más específicas centradas en el caso de uso concreto de la aplicación, el público objetivo o las indicaciones y consultas que es más probable que reciba.

Asimismo, puede entrenar modelos de barrera de protección en conjuntos de datos anotados que identifiquen diferentes tipos y grados de toxicidad. Todo ello ayuda al FM a aprender a detectar y filtrar contenido no deseado en los datos de entrenamiento, las indicaciones de entrada y las salidas generadas de forma automática.

Protección de la privacidad

Puede adoptar diversas medidas para ayudar a prevenir la exposición no deseada de información confidencial, secretos comerciales y propiedad intelectual al trabajar con aplicaciones de IA generativa.

La eliminación de modelos es un método que contribuye a resolver los problemas de privacidad. Se trata de eliminar los datos utilizados incorrectamente tan pronto como se hayan identificado, lo que propiciará la eliminación de los efectos de esos datos en cualquier componente del FM.

Otro enfoque es la partición, en la que los datos de entrenamiento se dividen en porciones más pequeñas en las que se entrenan submodelos independientes, que finalmente se combinan para formar el FM general. Esta práctica puede facilitar mucho la reparación de los FM que tienen o corren el riesgo de exponer información privada. En lugar de volver a entrenar todo el modelo, solo tiene que eliminar los datos no deseados o mal utilizados de la partición y, a continuación, volver a entrenar ese submodelo.

El filtrado y el bloqueo también pueden ser enfoques eficaces. Estos métodos comparan explícitamente la información protegida con el contenido generado antes de que el usuario lo vea. Si los dos son demasiado similares, el contenido se suprime o reemplaza para evitar la exposición. Otro elemento que puede resultar útil es limitar el número de veces que aparece un contenido concreto en los datos de entrenamiento.

Mejorar la explicabilidad y la auditabilidad

Para respaldar aún más la IA responsable, considere la necesidad de explicar la metodología y los factores clave que influyen en los resultados de la aplicación. La auditabilidad es otro componente importante de la IA responsable. Implemente mecanismos que le permitan rastrear y revisar el desarrollo y el funcionamiento de la aplicación de IA generativa. Esto le ayudará a rastrear las causas fundamentales de cualquier problema y a satisfacer los requisitos de gobernanza.

Considere la posibilidad de documentar las decisiones y las aportaciones de diseño relevantes a lo largo del ciclo de vida del desarrollo. Establecer un registro rastreable puede ayudar a los equipos internos o externos a evaluar el desarrollo y el funcionamiento de la aplicación de IA generativa.

Mantener la responsabilidad

Por último, piense en cómo ayudará a garantizar el cumplimiento continuo de sus políticas de IA responsables. Asegúrese de aplicar las lecciones que aprenda y la experiencia que adquiera para desarrollar prácticas de seguridad y privacidad. Instruya periódicamente a todos los empleados de la organización sobre sus obligaciones en lo que se refiere a la seguridad y la protección de las prácticas de IA generativa. Fomente una cultura de IA responsable, utilice las herramientas adecuadas que contribuyan a supervisar el rendimiento del modelo e informar de los riesgos, y permita que los equipos inspeccionen el modelo y sus componentes cuando sea necesario. Pruebe, pruebe y, en caso de duda, vuelva a probar.

Pregunta 4:

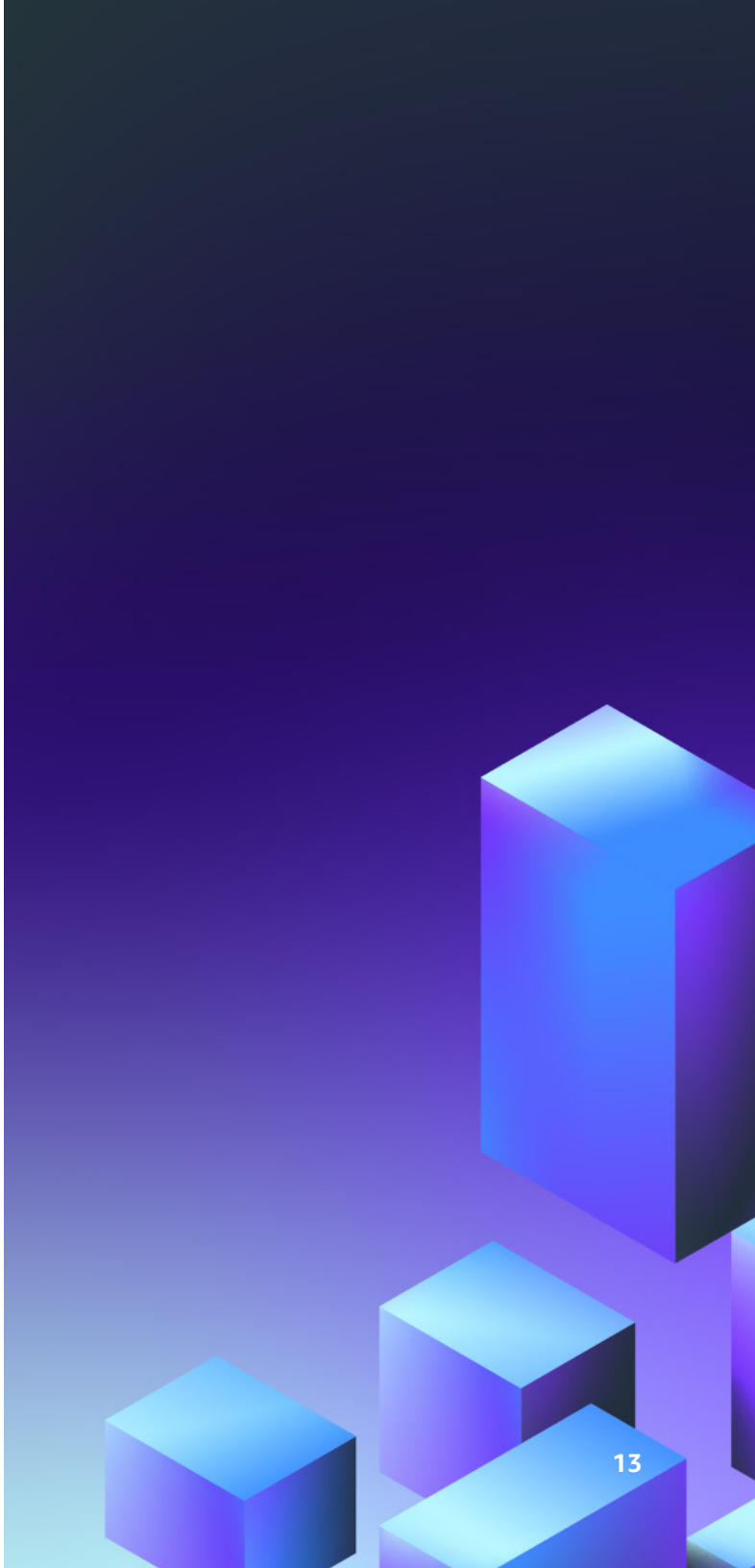
¿Por dónde empezar?

La protección de las aplicaciones de IA generativa no es una tarea sencilla, y no existe un conjunto universal de acciones que pueda emprender para lograrlo. Sin embargo, cuando se colabora con el proveedor adecuado y se despliegan las herramientas correctas, el camino hacia el éxito se vuelve mucho más claro.

El uso de **Amazon Bedrock**, por ejemplo, puede simplificar y acelerar drásticamente el proceso de desarrollo de aplicaciones de IA generativa seguras. Amazon Bedrock es un servicio totalmente administrado que facilita FM de Amazon y de las principales startups de IA a través de una API.

Al personalizar un modelo con Amazon Bedrock, el servicio puede ajustarlo para una tarea en particular sin que su equipo tenga que anotar grandes volúmenes de datos. Después, Amazon Bedrock hace una copia independiente del FM base a la que solo usted puede acceder y entrena esta copia privada del modelo. No se utiliza ninguno de sus datos para entrenar los modelos base originales, lo que ayuda a mantener la privacidad y la seguridad de sus datos de propiedad.

También puede configurar los ajustes de **Amazon VPC** para acceder a las API de Amazon Bedrock y proporcionar a su modelo datos de ajuste preciso de forma segura. Los datos siempre están cifrados, tanto en tránsito como en reposo, gracias a claves administradas por el servicio. Además, con **AWS PrivateLink**, tiene la posibilidad de transferir los datos de la nube de AWS a Amazon Bedrock exclusivamente a través de la red de AWS, nunca a través de la Internet pública.



Mejora de la privacidad con AWS

Tanto si crea aplicaciones de IA generativa con Amazon Bedrock, otro servicio (como **Amazon SageMaker**) o sus propias herramientas, cuando ejecuta y administra sus aplicaciones en AWS, obtiene protecciones y controles de privacidad líderes del sector.

AWS admite 143 estándares de seguridad y certificaciones de cumplimiento, lo que ayuda a satisfacer los requisitos de nuestros clientes en todo el mundo. Todos los datos se pueden cifrar en reposo con sus propias claves de **AWS Key Management Service** (Amazon KMS), lo que proporciona un control y una visibilidad totales de cómo se almacenan sus datos y FM y cómo se accede a ellos.

CONCLUSIÓN

Próximos pasos

AWS se compromete a ayudarle a crear aplicaciones de IA generativa que hagan crecer su empresa y, al mismo tiempo, lo ayuden a satisfacer sus objetivos de seguridad, privacidad y cumplimiento.

Creemos firmemente que las aplicaciones de IA generativa se pueden diseñar, desarrollar y operar de forma segura. También reconocemos la validez de las preocupaciones de seguridad y privacidad relacionadas con estas tecnologías. La **IA generativa plantea nuevos desafíos** a la hora de definir, medir y mitigar los problemas relacionados con la privacidad de los datos, la propiedad intelectual, la supervisión legislativa, la igualdad y la transparencia.

Con la introducción de nuevos productos, la creciente complejidad y escala de las soluciones, los nuevos parámetros de entrenamiento y los conjuntos de datos en constante crecimiento, la seguridad de la IA generativa será aún más esencial en los días venideros. Si desarrolla ahora una estrategia de seguridad eficaz e integral para las cargas de trabajo de la IA generativa, podrá maximizar su ventaja competitiva y estar preparado para el futuro que se acerca a gran velocidad.

La buena noticia es que los controles básicos necesarios para diseñar, desarrollar y ejecutar aplicaciones de IA generativa de forma segura llevan años en vigor y están alineados con los principios fiables y comprobados de seguridad en la nube, como los que se encuentran en el **AWS Well-Architected Framework**.

Al explorar las prácticas descritas en este libro electrónico, ya ha dado el primer paso para proteger sus cargas de trabajo de la IA generativa.

Ahora, dé el siguiente paso con AWS. Podemos proporcionarle la información detallada y la orientación específica que necesita para mantenerse al día sobre los temas emergentes, analizar sus desafíos particulares y sacar partido de todas las ventajas de la IA generativa, al tiempo que protegemos sus datos, sus clientes y su empresa.

Más información sobre la IA generativa en AWS ›

Comience rápidamente con Amazon Bedrock ›

Cree y personalice FM en Amazon SageMaker ›

Aumente su seguridad en la nube con AWS ›

Lleve la IA responsable de la teoría a la práctica ›