Securing Research Infrastructure for Advanced AI

5 jun 2024

Source:
https://openai.com/index/securing-research-infrastructure-for-advanced-ai/

Estamos compartiendo algunos detalles de alto nivel sobre la arquitectura de seguridad de nuestros supercomputadores de investigación.

OpenAI opera algunos de los supercomputadores de entrenamiento de IA más grandes, lo que nos permite ofrecer modelos que son líderes en la industria tanto en capacidades como en seguridad, mientras avanzamos en las fronteras de la IA. Nuestra misión es asegurar que la IA avanzada beneficie a todos, y la base de este trabajo es la infraestructura que impulsa nuestra investigación.

Para lograr esta misión de manera segura, priorizamos la seguridad de estos sistemas. Aquí, delineamos nuestra arquitectura y operaciones actuales que respaldan el entrenamiento seguro de modelos de frontera a gran escala. Esto incluye medidas diseñadas para proteger los pesos de los modelos sensibles dentro de un entorno seguro para la innovación en IA. Aunque estas características de seguridad evolucionarán con el tiempo, creemos que es valioso proporcionar una instantánea actual de cómo pensamos sobre la seguridad de nuestra infraestructura de investigación. Esperamos que esta información sea útil para otros laboratorios de investigación de IA y profesionales de la seguridad en su enfoque para asegurar sus propios sistemas (y estamos contratando).

Modelo de Amenazas

La infraestructura de investigación presenta un desafío de seguridad único dado el carácter diverso y rápidamente evolutivo de las cargas de trabajo necesarias para la experimentación.

La infraestructura de investigación alberga varios tipos importantes de activos que es esencial proteger. Entre ellos, los pesos de modelos no publicados son de suma importancia, ya que representan propiedad intelectual central y deben ser protegidos contra la divulgación o el compromiso no autorizados.

Con este propósito en mente, OpenAI creó una serie de entornos de investigación dedicados al desarrollo y la seguridad de modelos de frontera. La infraestructura de investigación debe apoyar la protección de los pesos de los modelos, secretos algorítmicos y otros activos sensibles utilizados para desarrollar modelos de frontera protegiéndolos contra la exfiltración y el compromiso no autorizados. Al mismo tiempo, los investigadores deben tener acceso suficiente a los recursos y la infraestructura informática subyacente para ser productivos y eficientes.

Arquitectura

Nuestra arquitectura técnica para la investigación está construida sobre Azure, utilizando Kubernetes para la orquestación. Aprovechamos ambos para implementar una arquitectura de seguridad que permite la investigación mientras se ajusta a nuestro modelo de amenazas.

Securing Research Infrastructure for Advanced AI

5 jun 2024

1. Fundamento de Identidad

Nuestro fundamento de identidad se basa en Azure Entra ID (anteriormente Azure Active Directory). Azure Entra ID se integra con marcos y controles internos de autenticación y autorización. Azure Entra ID permite la verificación basada en riesgos en la creación de sesiones, el uso de tokens de autenticación y la detección de inicios de sesión anómalos. Estas características complementan nuestras herramientas internas de detección para identificar y bloquear posibles amenazas.

2. Arquitectura de Kubernetes

Usamos Kubernetes para orquestar y gestionar las cargas de trabajo en nuestra infraestructura. Las cargas de trabajo de investigación están protegidas por políticas de control de acceso basado en roles (RBAC) de Kubernetes para adherirse a los principios de privilegio mínimo. Las políticas del Controlador de Admisión establecen una línea base de seguridad para las cargas de trabajo, controlando los privilegios de los contenedores y el acceso a la red para reducir los riesgos.

Confiamos en la tecnología VPN moderna para proporcionar redes seguras a nuestros entornos de investigación. Las políticas de red definen cómo las cargas de trabajo se comunican con servicios externos. Adoptamos una política de egress (salida) por defecto denegada y permitimos explícitamente las rutas de comunicación externa autorizadas. Utilizamos extensamente el enrutamiento de red de enlace privado donde se ofrece para eliminar rutas requeridas a Internet y mantener corta esta lista de permitidos.

Para algunas tareas de mayor riesgo usamos gVisor (se abre en una nueva ventana), un runtime de contenedores que proporciona aislamiento adicional. Este enfoque de defensa en profundidad asegura una gestión robusta de la seguridad y eficiente de las cargas de trabajo.

3. Almacenamiento de Datos Sensibles

Los datos sensibles como credenciales, secretos y cuentas de servicio requieren protección adicional. Usamos servicios de gestión de claves para almacenar y gestionar información sensible en nuestra infraestructura de investigación, y control de acceso basado en roles para limitar el acceso a secretos, de modo que solo las cargas de trabajo y usuarios autorizados puedan recuperarlos o modificarlos.

4. Gestión de Identidad y Acceso (IAM) para Investigadores y Desarrolladores

La gestión de acceso es crucial para administrar el acceso de los investigadores y desarrolladores a los sistemas descritos anteriormente. Los objetivos de seguridad con cualquier solución IAM son permitir estrategias de acceso de "privilegio mínimo" por tiempo limitado en todos los recursos, una gestión eficiente y auditabilidad.

Con ese fin, construimos un servicio llamado AccessManager como un mecanismo escalable para gestionar la autorización interna y permitir la autorización de privilegio mínimo. Este servicio federado

Securing Research Infrastructure for Advanced AI

5 jun 2024

decisiones de gestión de acceso a aprobadores según lo definido por las políticas. Esto asegura que las decisiones para otorgar acceso a recursos sensibles, incluidos los pesos de los modelos, sean tomadas por personal autorizado con la supervisión adecuada.

Las políticas de AccessManager pueden definirse para ser estrictas o flexibles, adaptadas al recurso en cuestión. Solicitar y ser concedido acceso a recursos sensibles, como el almacenamiento en el entorno de investigación que contiene pesos de modelos, requiere la aprobación de múltiples partes. Para recursos sensibles, las concesiones de autorización de AccessManager se configuran para expirar después de un período de tiempo especificado, lo que significa que los privilegios se reducen a un estado no privilegiado si no se renuevan. Al implementar estos controles, reducimos el riesgo de acceso interno no autorizado y compromiso de cuentas de empleados.

Integrarnos GPT-4 en AccessManager para facilitar la asignación de roles de mínimo privilegio. Los usuarios pueden buscar recursos dentro de AccessManager, y el servicio utilizará nuestros modelos para sugerir roles que puedan otorgar acceso a ese recurso. Conectar a los usuarios con roles más específicos combate la dependencia de roles amplios, genéricos y sobrepermisivos. Humanos en el proceso mitigan el riesgo de que el modelo proponga el rol incorrecto, tanto en la solicitud inicial de rol como en un paso de aprobación de múltiples partes si la política para el rol especificado lo requiere.

5. Seguridad CI/CD

Nuestros equipos de infraestructura utilizan pipelines de Integración Continua y Entrega Continua (CI/CD) para construir y probar nuestra infraestructura de investigación. Hemos invertido en asegurar nuestros pipelines CI/CD de infraestructura para hacerlos más resilientes contra posibles amenazas mientras mantenemos la integridad de nuestros procesos de desarrollo y despliegue y la velocidad para nuestros investigadores e ingenieros.

Restringimos la capacidad de crear, acceder y activar pipelines relacionados con la infraestructura para evitar el acceso a secretos disponibles para el servicio CI/CD. El acceso a los trabajadores de CI/CD está restringido de manera similar. Fusionar código a la rama de despliegue requiere la aprobación de múltiples partes, añadiendo una capa adicional de supervisión y seguridad. Usamos paradigmas de infraestructura como código (IaC) para configurar infraestructura a escala de manera consistente, repetible y segura. La configuración esperada es reforzada por CI en cada cambio a nuestra infraestructura, generalmente múltiples veces al día.

6. Flexibilidad

Al mismo tiempo, la investigación requiere empujar la frontera. Esto puede requerir iteraciones rápidas en nuestra infraestructura para apoyar requisitos funcionales y restricciones cambiantes. Esta flexibilidad es esencial para lograr tanto requisitos de seguridad como funcionales, y en algunos casos es vital permitir excepciones con controles compensatorios adecuados para lograr esos objetivos.

Protección de Pesos de Modelos

Securing Research Infrastructure for Advanced AI

5 jun 2024

Proteger los pesos de los modelos contra la exfiltración del entorno de investigación requiere un enfoque de defensa en profundidad que abarque múltiples capas de seguridad. Estos controles específicos están diseñados para proteger nuestros activos de investigación contra el acceso y el robo no autorizados, mientras se asegura que permanezcan accesibles para fines de investigación y desarrollo. Estas medidas pueden incluir:

- Autorización: Las concesiones de acceso a cuentas de almacenamiento de investigación que contienen pesos de modelos sensibles requieren la aprobación de múltiples partes.
- Acceso: Los recursos de almacenamiento para los pesos de los modelos de investigación están vinculados privadamente en el entorno de OpenAI para reducir la exposición a Internet y requieren autenticación y autorización a través de Azure para el acceso.
- Controles de Salida: El entorno de investigación de OpenAI utiliza controles de red que permiten el tráfico de salida solo a objetivos específicos de Internet predefinidos. El tráfico de red hacia hosts no en la lista de permitidos es denegado.
- Detección: OpenAI mantiene un mosaico de controles de detección para respaldar esta arquitectura. Los detalles de estos controles se retienen intencionadamente.

Auditoría y Pruebas

OpenAI utiliza equipos rojos internos y externos para simular adversarios y probar nuestros controles de seguridad para el entorno de investigación. Hemos sometido nuestro entorno de investigación a pruebas de penetración por una consultora de seguridad de terceros líder, y nuestro equipo rojo interno realiza evaluaciones profundas contra nuestras prioridades.

Estamos explorando regímenes de cumplimiento para nuestro entorno de investigación. Dado que proteger los pesos de los modelos es un problema de seguridad específico, establecer un marco de cumplimiento para cubrir este desafío requerirá algo de personalización. En este momento, estamos evaluando estándares de seguridad existentes más controles personalizados específicos para proteger la tecnología de IA. Esto puede crecer para incluir estándares de seguridad y regulatorios específicos de IA que aborden los desafíos únicos de asegurar sistemas de IA, como los esfuerzos emergentes de la Iniciativa de Seguridad de IA de la Cloud Security Alliance (se abre en una nueva ventana) o las actualizaciones de IA del NIST SP 800-218.

Investigación y Desarrollo en Controles Futuros

Asegurar sistemas de IA cada vez más avanzados requerirá innovación y adaptación continuas. Estamos a la vanguardia del desarrollo de nuevos controles de seguridad, como se detalla en nuestra publicación de blog “Reimaginando la Infraestructura Segura para la IA Avanzada”. Nuestro compromiso con la investigación y el desarrollo asegura que nos mantengamos por delante de las amenazas emergentes y continuemos mejorando la seguridad de nuestra infraestructura de IA.