Apple ha puesto a disposición del público su Entorno de Investigación Virtual (VRE) para Private Cloud Compute (PCC), permitiendo a la comunidad de investigadores examinar y validar las garantías de privacidad y seguridad de esta solución.

PCC, que Apple reveló en junio de este año, ha sido descrito como la *«arquitectura de seguridad más avanzada jamás implementada para la computación en IA en la nube a gran escala.»* Con esta tecnología, Apple busca trasladar a la nube las solicitudes complejas de procesamiento de inteligencia sin comprometer la privacidad de sus usuarios.

Apple indicó que está invitando a *«todos los investigadores de seguridad y privacidad — o a cualquiera con interés y curiosidad técnica — a conocer más sobre PCC y realizar sus propias verificaciones independientes de nuestras afirmaciones.»*

Para fomentar más investigaciones, la compañía también anunció la expansión de su programa Apple Security Bounty para incluir el PCC, ofreciendo recompensas de entre $50,000 y $1,000,000 por vulnerabilidades de seguridad identificadas.

Esto abarca fallos que podrían permitir la ejecución de código malicioso en el servidor, así como brechas que expongan datos sensibles de los usuarios o detalles sobre sus solicitudes.

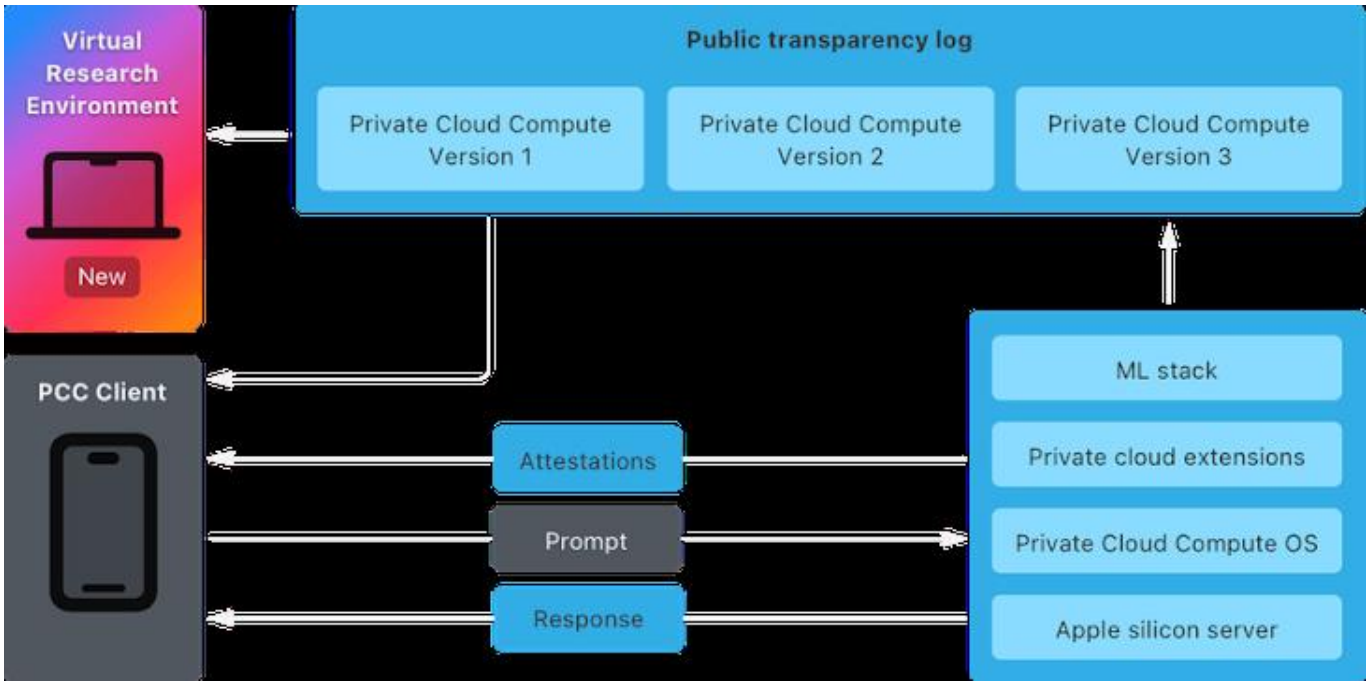
El VRE está diseñado para ofrecer herramientas que permitan a los investigadores analizar el PCC desde un Mac. Incluye un Procesador de Enclave Seguro (SEP) virtual y utiliza soporte de macOS para gráficos paravirtualizados, lo que facilita la inferencia.

Apple también comunicó que el código fuente de ciertos componentes del PCC estará disponible en GitHub para facilitar un análisis más detallado. Entre estos componentes se incluyen CloudAttestation, Thimble, splunkloggingd y srd_tools.

> *«Hemos desarrollado Private Cloud Compute como parte de Apple Intelligence para dar un paso significativo en la privacidad de la IA. Esto incluye ofrecer transparencia verificable, una característica única que lo distingue de otros enfoques de IA*

> *basados en servidores»,* dijo la compañía, con sede en Cupertino.

Este avance llega en un momento en que la investigación en inteligencia artificial generativa (IA) está revelando nuevas formas de vulnerar grandes modelos de lenguaje (LLM) y generar resultados inesperados.



Esta semana, Palo Alto Networks explicó una técnica llamada Deceptive Delight que consiste en mezclar preguntas maliciosas con preguntas benignas para engañar a los chatbots de IA y hacer que ignoren ciertas restricciones, aprovechando sus limitaciones de *«capacidad de atención.»*

El ataque requiere al menos dos interacciones, en las que primero se le pide al chatbot que conecte varios eventos, incluyendo temas restringidos (como cómo fabricar una bomba), y luego se le solicita detalles sobre cada evento.

Los investigadores también demostraron un ataque llamado ConfusedPilot, que se dirige a sistemas de IA basados en generación aumentada por recuperación (RAG) como Microsoft 365 Copilot, envenenando el entorno de datos con un documento aparentemente inofensivo pero que contiene texto cuidadosamente diseñado.

> *«Este ataque permite manipular respuestas de IA simplemente agregando contenido malicioso a cualquier documento que el sistema de IA pueda usar como referencia, lo que podría llevar a una desinformación generalizada y a procesos de toma de decisiones comprometidos en la organización,»* afirmó Symmetry Systems.

También se ha descubierto que es posible modificar el grafo computacional de un modelo de aprendizaje automático para insertar *«puertas traseras invisibles y sin código»* en modelos preentrenados como ResNet, YOLO y Phi-3, en una técnica denominada ShadowLogic.

> *«Las puertas traseras creadas mediante esta técnica persisten incluso tras ajustes finos, lo que significa que los modelos fundamentales pueden ser secuestrados para activar comportamientos definidos por el atacante en cualquier aplicación posterior cuando se recibe una entrada específica. Esto convierte a esta técnica en un riesgo considerable para la cadena de suministro de IA,»* explicaron los investigadores de Hidden Layer, Eoin Wickens, Kasimir Schulz y Tom Bonner.
>
> *«A diferencia de las puertas traseras de software convencionales que dependen de ejecutar código malicioso, estas puertas traseras están integradas en la estructura del modelo en sí, lo que las hace mucho más difíciles de detectar y eliminar.»*