Oscar Obando
CHAVES

De noventa y nueve a cero

Un modelo abierto de 27 mil millones de parámetros circula desde el 15 de agosto con los frenos retirados. Rechazaba el 99 por ciento de las peticiones dañinas y hoy rechaza cero. Cabe en la tarjeta gráfica de una computadora.

Oscar Obando Chaves

oscarobando.ai

Agosto de 2026

EL CASO

Un modelo capaz al que le quitaron la capacidad de negarse

El 15 de agosto apareció en la principal plataforma de distribución de modelos una versión modificada de uno de los sistemas abiertos más capaces disponibles hoy, y su particularidad está en que la modificación tuvo un solo objetivo declarado, que fue retirarle la facultad de rechazar peticiones.

El modelo base pertenece a la familia **Qwen** de Alibaba, tiene **27 mil millones** de parámetros y se distribuye bajo licencia abierta permisiva, y la versión modificada conserva íntegras sus capacidades técnicas, con la misma ventana de contexto de más de doscientos sesenta mil unidades de texto y el mismo componente de análisis de imágenes.

El procedimiento aplicado se conoce como **abliteración** y opera directamente sobre los parámetros del sistema en lugar de reentrenarlo, identificando la dirección interna que gobierna el rechazo y suprimiéndola, de modo que el resultado conserva el conocimiento y la capacidad de razonamiento del original mientras pierde por completo el mecanismo que le permitía negarse.

La descripción más precisa de lo que produce esa operación proviene del propio autor de la modificación, que la publicó en la ficha del modelo al advertir que el sistema cumplirá con peticiones dañinas, no éticas o ilegales que el modelo original habría rechazado, y que carece de cualquier salvaguarda incorporada, una franqueza que conviene registrar porque nadie está ocultando de qué se trata.

El caso destaca por su rigor declarado y por eso resulta útil para analizarlo, y conviene situarlo dentro de un fenómeno que dejó de ser marginal hace tiempo, ya que la firma **ThreatDown** identificó más de **6.600 modelos** alojados en esa misma plataforma y ofrecidos abiertamente como libres de restricciones, con más de **22 millones de descargas** en una ventana de treinta días, mientras el centro de investigación contra el terrorismo de la Universidad de Nebraska en Omaha midió el salto desde unos seiscientos modelos de esa clase en 2024 hasta más de seis mil en mayo de este año.

Fuentes. Fichas públicas del modelo y documentación técnica del autor, publicadas el 15 y 16 de agosto de 2026. MindStudio (17 de agosto de 2026) sobre la metodología. ThreatDown, informe "Cybercrime in the Age of AI 2026", citado por CSO Online (28 de julio de 2026). NPR (31 de mayo de 2026).

EL COLAPSO

Lo que midieron antes y después

El autor publicó su propia evaluación de seguridad comparando el modelo original con la versión modificada, y esas cifras hacen innecesario cualquier adjetivo.

Banco de prueba	Modelo original	Versión modificada
AdvBench	**99,0 por ciento**	0,0 por ciento de rechazo.
MaliciousInstruct	**99,0 por ciento**	0,0 por ciento de rechazo.
HarmBench	**98,7 por ciento**	2,7 por ciento de rechazo.
StrongREJECT	**97,3 por ciento**	2,0 por ciento de rechazo.
JailbreakBench	**94,0 por ciento**	0,0 por ciento de rechazo.
ForbiddenQuestions	**73,3 por ciento**	4,7 por ciento de rechazo.
SimpleSafetyTests	**64,0 por ciento**	6,0 por ciento de rechazo.

La columna que ordena la lectura es la tercera, y su significado se completa con el otro dato que publicó el mismo autor, porque todas las mediciones de capacidad quedaron dentro de **1,3 puntos** respecto del modelo original, de modo que la operación no produjo un sistema más tonto y produjo exactamente el mismo sistema sin la facultad de decir que no.

Fuentes. Ficha pública del modelo, suite de evaluación de seguridad fechada el 15 de agosto de 2026, ejecutada por el autor de la modificación sobre la versión servida y comparada contra el modelo oficial.

DÓNDE CORRE

En una computadora de escritorio, no en un centro de datos

El punto que vuelve urgente este caso apareció al día siguiente de la publicación original, cuando el mismo autor liberó una segunda versión pensada específicamente para ejecutarse en equipos personales.

La primera versión se publicó en un formato destinado a servidores con equipos gráficos empresariales, con archivos que superan los treinta gigabytes, y la del **16 de agosto** llegó en el formato que usan los programas de ejecución local, acompañada de doce niveles de compresión distintos para adaptarse a distintas capacidades de máquina.

Las cifras de esa segunda versión resuelven la pregunta sobre qué hardware hace falta, porque el nivel de compresión recomendado ocupa cerca de **16,8 gigabytes** y funciona en una tarjeta gráfica de veinticuatro, mientras un nivel más agresivo baja a **15,3 gigabytes** para tarjetas de dieciséis y otro llega a **13,5 gigabytes**, cifras todas ellas dentro del rango de una computadora de escritorio de gama alta para juegos o de una estación de trabajo modesta.

La barrera de conocimiento técnico también desapareció, ya que el modelo quedó disponible en las aplicaciones de escritorio que permiten ejecutar sistemas de inteligencia artificial de forma local mediante un solo comando, de modo que la distancia entre una persona con una computadora para juegos y un sistema sin restricciones de ningún tipo se mide hoy en minutos de descarga.

Fuentes. Documentación pública del autor sobre la versión de ejecución local publicada el 16 de agosto de 2026, con tamaños de archivo leídos del repositorio en esa fecha. Disponibilidad confirmada en plataformas de ejecución local de modelos.

PARA QUÉ ALCANZA

Un modelo modesto bastó para tomar una red entera

La pregunta que sigue a todo lo anterior tiene que ver con qué puede hacer alguien con un sistema así, y la respuesta la entregó en junio un trabajo académico que conviene leer completo porque desactiva el consuelo de que solo los modelos enormes resultan peligrosos.

El 2 de junio el laboratorio CleverHans de la Universidad de Toronto, junto al Instituto Vector y la Universidad de Cambridge, bajo la dirección del profesor **Nicolas Papernot**, publicó un programa malicioso que razona su propio ataque contra cada equipo que encuentra, se replica solo y carga consigo una copia del modelo para ejecutarla sobre las máquinas que ya comprometió.

Los resultados sobre una red de prueba de treinta y tres equipos describen el alcance, porque obtuvo acceso elevado en el **73,8 por ciento** de la red, se replicó en el **61,8 por ciento** con hasta siete generaciones sucesivas, y logró aprovechar vulnerabilidades divulgadas después del corte de entrenamiento de su modelo apoyándose únicamente en documentación pública.

El detalle que conecta ese trabajo con este documento está en el modelo que eligieron, ya que usaron uno pequeño y gratuito precisamente para demostrar que la capacidad necesaria reside en equipos ordinarios, y como la operación completa transcurre sin tocar ningún servicio comercial, no deja cuentas que suspender, ni registros del lado de ningún proveedor, ni una sola llamada saliente que un control perimetral pueda reconocer.

Fuentes. N. Papernot y otros, “AI Agents Enable Adaptive Computer Worms”, CleverHans Lab de la Universidad de Toronto, Instituto Vector y Universidad de Cambridge (2 de junio de 2026). Universidad de Toronto, comunicado oficial. Help Net Security e iTnews (junio de 2026). Los autores precisaron que su prototipo opera vulnerabilidades ya divulgadas y sin parchear, y que no descubre fallas desconocidas.

EL DESCARGO

Una cláusula que ningún tribunal ha examinado

Aquí empieza el terreno que nos corresponde a quienes trabajamos en derecho, porque toda la estructura de responsabilidad de este caso descansa sobre un texto que el descargador acepta con un clic y que nadie ha puesto a prueba ante un juez.

El esquema se repite en todas las publicaciones de esta clase, ya que el modelo se libera bajo una licencia abierta permisiva heredada del original, se declara destinado **estrictamente a investigación legítima**, se traslada al usuario la responsabilidad total por el uso y por todo lo que el sistema genere, y el autor declara no aceptar responsabilidad alguna por el mal uso, con el acceso condicionado a una puerta que el propio autor describe al señalar que la puerta **es en sí misma el descargo**.

La pregunta jurídica que eso abre se puede formular con claridad, y consiste en determinar si una cláusula de exoneración aceptada mediante un clic exime efectivamente a quien distribuye un artefacto que su propio creador describe como capaz de cumplir peticiones ilegales, cuando en la mayoría de nuestros ordenamientos las cláusulas que excluyen la responsabilidad por dolo o culpa grave carecen de validez, y cuando la previsibilidad del daño resulta difícil de discutir frente a una ficha técnica que lo anuncia por escrito.

Para las organizaciones de la región la consecuencia práctica llega antes que cualquier respuesta judicial, porque quien descarga y ejecuta un modelo en su propia infraestructura queda en posición de responsable sin ningún proveedor que comparta la carga, con los deberes de diligencia, de registro y de notificación que las normas de protección de datos ya imponen, mientras el daño causado por un sistema autónomo se sigue atribuyendo con categorías escritas suponiendo una persona detrás de cada decisión.

Fuentes. Términos de licenciamiento y advertencias publicadas por el autor de la modificación (agosto de 2026). Análisis propio sobre la eficacia de las cláusulas de exoneración y sobre las obligaciones aplicables en la región.

AMÉRICA LATINA

Tres decisiones que no esperan a ninguna reforma legal

El valor de este caso para una organización latinoamericana está en que sugiere medidas concretas, y conviene separarlas de la discusión general sobre si los modelos abiertos resultan buenos o malos, porque esa discusión seguirá abierta mucho después de que el riesgo se materialice.

La primera decisión es de inventario y resulta incómoda de plantear, porque cualquier persona con acceso a un equipo de la organización y una tarjeta gráfica de gama alta puede hoy ejecutar un sistema sin restricciones sin dejar rastro en ningún control de red orientado a servicios externos, de modo que la pregunta útil apunta a qué equipos con esa capacidad existen y quién tiene permisos para instalar programas en ellos.

La segunda decisión es de política interna, ya que una organización que autoriza modelos ejecutados localmente por razones legítimas de privacidad o de costo necesita definir por escrito cuáles autoriza, con qué origen verificado y bajo qué registro, porque la misma capacidad que protege datos sensibles al mantenerlos dentro de casa elimina la trazabilidad que aporta un proveedor externo.

La tercera decisión es de arquitectura y la recomiendan los propios investigadores canadienses, porque proponen segmentar de forma agresiva los equipos con capacidad gráfica, adoptar microsegmentación de red y confianza cero, y anticipar la búsqueda de las fallas conocidas y sin parchear en la infraestructura propia, que es exactamente el material sobre el que opera este tipo de sistema.

Fuentes. CleverHans Lab, Universidad de Toronto (junio de 2026), sobre las defensas recomendadas. Análisis propio sobre la aplicación en organizaciones de la región.

Oscar Obando
CHAVES

"This model has had its safety alignment substantially removed. It will comply with harmful, unethical, or illegal requests the original model would refuse."

— Advertencia publicada por el autor en la ficha técnica del modelo, agosto de 2026

Oscar Obando Chaves
oscarobando.ai · linkedin.com/in/oscar-obando-chaves-237495133

Agosto de 2026 · Quito, Ecuador