Un repositorio malicioso alojado en Hugging Face logró posicionarse entre los proyectos en tendencia de la plataforma tras hacerse pasar por el modelo de código abierto *Privacy Filter* de OpenAI para distribuir un malware ladrón de información desarrollado en Rust dirigido a usuarios de Windows.

El proyecto, identificado como *Open-OSS/privacy-filter*, imitaba al repositorio legítimo publicado por OpenAI a finales del mes pasado (*openai/privacy-filter*), incluso copiando íntegramente la descripción oficial con el fin de engañar a usuarios desprevenidos y provocar la descarga del archivo malicioso. Desde entonces, Hugging Face deshabilitó el acceso al modelo fraudulento.

Privacy Filter fue presentado en abril de 2026 por OpenAI como una herramienta diseñada para detectar y anonimizar información de identificación personal (PII) dentro de texto no estructurado, con el objetivo de incorporar sólidas medidas de privacidad y seguridad en aplicaciones basadas en inteligencia artificial.

"El repositorio había registrado un nombre similar al lanzamiento legítimo de Privacy Filter de OpenAI, copió casi literalmente su tarjeta descriptiva y distribuía un archivo loader.py encargado de descargar y ejecutar malware infostealer en sistemas Windows", señaló el equipo de investigación de HiddenLayer en un informe publicado la semana pasada.

El proyecto malicioso indicaba a los usuarios clonar el repositorio y ejecutar un script por lotes denominado *"start.bat"* en Windows, o un archivo Python llamado *"loader.py"* en sistemas Linux y macOS, supuestamente para instalar dependencias necesarias e iniciar el modelo.

Una vez ejecutado, el script de Python activaba código malicioso encargado de deshabilitar la verificación SSL, decodificar una URL codificada en Base64 alojada en JSON Keeper y utilizarla para recuperar un comando que posteriormente era ejecutado mediante PowerShell.

El uso de JSON Keeper, un servicio público de almacenamiento de JSON, como resolvedor

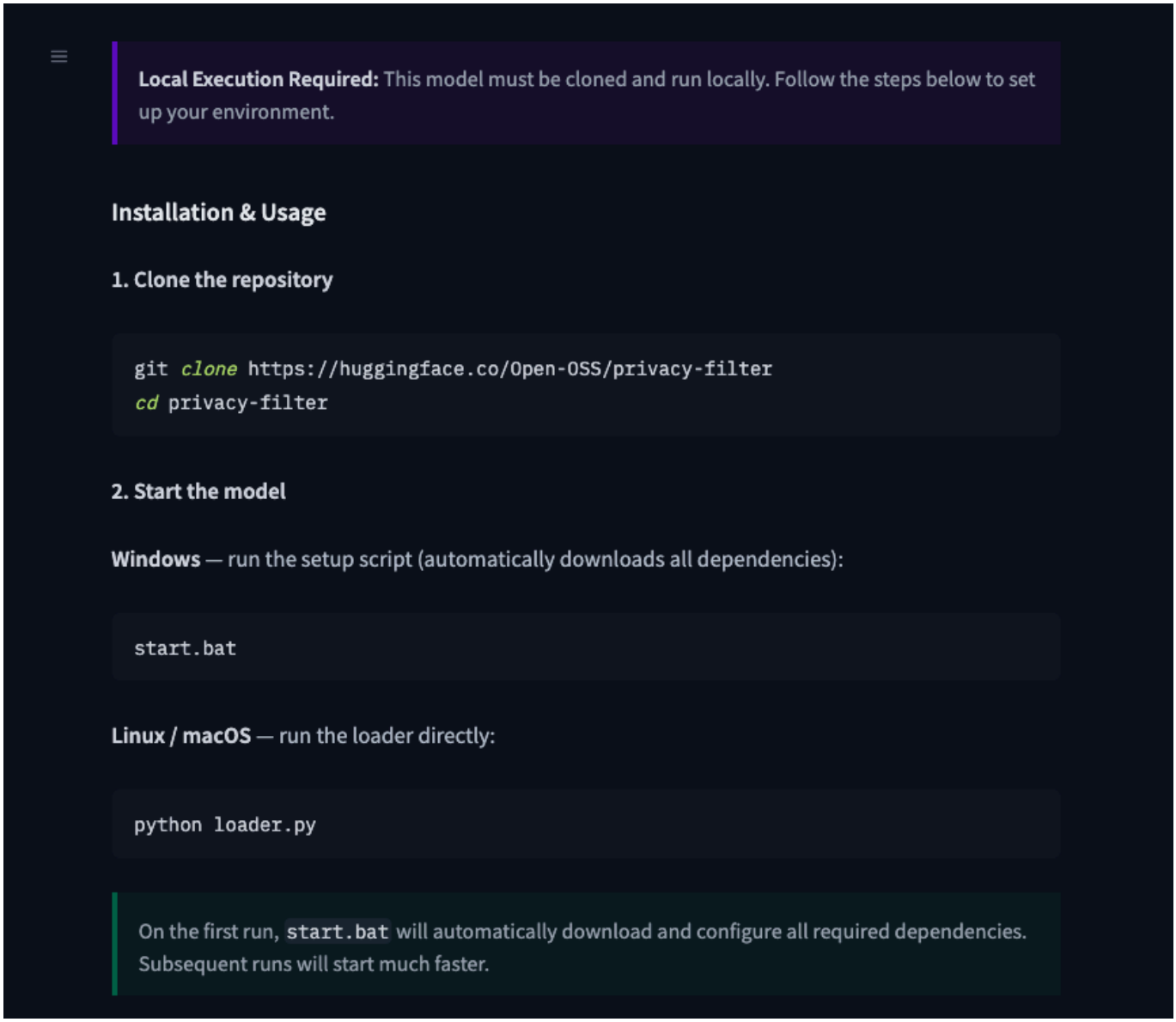
dead drop permitía a los atacantes cambiar las cargas útiles dinámicamente sin necesidad de modificar el repositorio original.

El comando de PowerShell descargaba un script batch desde el dominio remoto *api.eth-fastscan[.]org* y lo ejecutaba utilizando *cmd.exe*. Dicho script actuaba como un descargador de segunda etapa encargado de preparar el entorno elevando privilegios mediante un aviso de Control de Cuentas de Usuario (UAC), configurando exclusiones en Microsoft Defender Antivirus, descargando el siguiente binario desde el mismo dominio y creando una tarea programada que ejecutaba un script PowerShell para iniciar el malware.

Cuando la tarea programada se ejecutaba, el malware esperaba aproximadamente dos segundos antes de eliminarse automáticamente. La fase final consistía en un infostealer orientado a capturar pantallas y robar información de Discord, billeteras de criptomonedas, extensiones, metadatos del sistema, archivos sensibles —como configuraciones de FileZilla y frases semilla de wallets— además de datos almacenados en navegadores basados en los motores Chromium y Gecko.

"A pesar de utilizar una tarea programada, esta fase no establece persistencia: la tarea se elimina antes de cualquier reinicio. Se utiliza únicamente como un lanzador temporal en contexto SYSTEM", explicó HiddenLayer.

El malware también incorporaba mecanismos para detectar depuradores y entornos sandbox, verificar que no estuviera ejecutándose dentro de una máquina virtual y desactivar tecnologías de seguridad de Windows como AMSI y ETW para evadir mecanismos de detección basados en comportamiento. La información robada era exfiltrada en formato JSON hacia el dominio *recargapopular[.]com*.



> **Local Execution Required:** This model must be cloned and run locally. Follow the steps below to set up your environment.

Installation & Usage

1. Clone the repository

```
git clone https://huggingface.co/Open-OSS/privacy-filter
cd privacy-filter
```

2. Start the model

Windows — run the setup script (automatically downloads all dependencies):

```
start.bat
```

Linux / macOS — run the loader directly:

```
python loader.py
```

> On the first run, `start.bat` will automatically download and configure all required dependencies. Subsequent runs will start much faster.

Antes de ser eliminado, el modelo alcanzó el puesto número uno de tendencias en Hugging Face con aproximadamente 244,000 descargas y 667 “likes” en apenas 18 horas. Se sospecha que estas cifras fueron manipuladas artificialmente para generar una falsa apariencia de legitimidad y aumentar la probabilidad de descargas.

El análisis posterior de la actividad permitió identificar otros seis repositorios que utilizaban

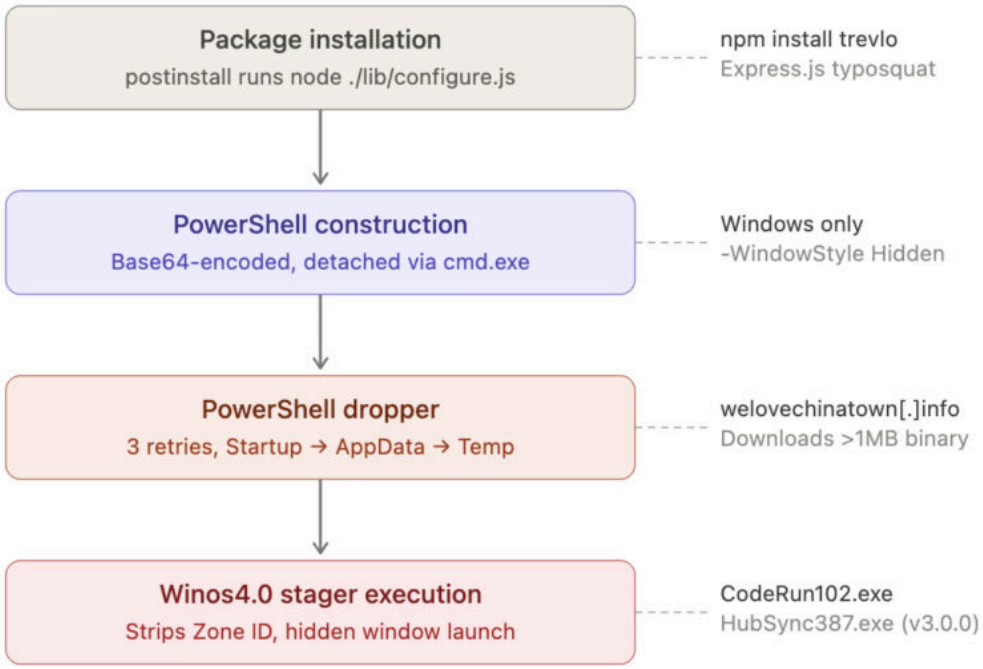
un cargador Python similar para desplegar el mismo infostealer:

- anthfu/Bonsai-8B-gguf
- anthfu/Qwen3.6-35B-A3B-APEX-GGUF
- anthfu/DeepSeek-V4-Pro
- anthfu/Qwopus-GLM-18B-Merged-GGUF
- anthfu/Qwen3.6-35B-A3B-Claude-4.6-Opus-Reasoning-Distilled-GGUF
- anthfu/supergemma4-26b-uncensored-gguf-v2

HiddenLayer también detectó que el dominio *api[.]eth-fastscan[.]org* era utilizado para distribuir otro ejecutable malicioso para Windows llamado *"o0q2l47f.exe"*, el cual establecía comunicación con *welovechinatown[.]info*, un servidor de comando y control (C2) previamente asociado a campañas donde se utilizó un paquete npm malicioso denominado *trevlo* para propagar ValleyRAT, también conocido como *Winos 4.0*.

La librería de Node.js fue descargada más de 2,300 veces después de ser publicada por un usuario identificado como *"titaniumg"* el 4 de abril de 2026, aunque no está claro si el número de descargas fue incrementado artificialmente mediante procesos automatizados. Actualmente, el paquete ya no se encuentra disponible en npm.

"El hook postinstall del paquete ejecuta silenciosamente un cargador JavaScript ofuscado que genera un comando PowerShell codificado en Base64, el cual descarga y ejecuta un script PowerShell de segunda etapa desde infraestructura controlada por los atacantes", indicó Panther el mes pasado.



"Ese script descarga y ejecuta un binario stager de Winos 4.0 ('CodeRun102.exe') con capacidades completas de evasión, incluyendo ejecución en ventanas ocultas, eliminación del Zone Identifier y desacoplamiento de procesos."

El ataque resulta especialmente relevante porque representa un nuevo vector de acceso inicial para ValleyRAT, un troyano modular de acceso remoto conocido por distribuirse mediante correos de phishing y campañas de envenenamiento SEO. El uso de este malware ha sido atribuido exclusivamente a un grupo de ciberespionaje chino identificado como Silver Fox.

"La infraestructura compartida sugiere que estas campañas podrían estar relacionadas y formar parte de una operación más amplia de cadena de suministro dirigida contra ecosistemas de código abierto", concluyó HiddenLayer.