

Evaluación de los grandes modelos de lenguaje para la detección de dominios generados algorítmicamente

Tomás Pelayo-Benedet¹, Ricardo J. Rodríguez¹, Carlos H. Gañán²

¹Grupo de I+D en Computación Distribuida
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, e-mail: tpelayo@unizar.es
²Delft University of Technology, Países Bajos

Resumen

Los ciberdelincuentes usan algoritmos para generar miles de dominios maliciosos y evadir bloqueos. En este trabajo evaluamos si los grandes modelos de lenguaje (ChatGPT, Claude, Gemini...) pueden detectarlos sin entrenamiento previo. Alcanzan hasta un 89% de acierto, pero generan demasiados falsos positivos para ser desplegados en sistemas reales.

Introducción

El cibercrimen mueve más dinero que el tráfico de drogas a nivel mundial [1]. Una de las técnicas más utilizadas por los atacantes son los algoritmos de generación de dominios (DGA), que crean automáticamente miles de direcciones web para mantener el contacto entre el malware y sus servidores de control, dificultando enormemente su bloqueo [2]. Estos dominios generados algorítmicamente (AGDs) pueden crearse mediante métodos aritméticos, basados en hash, de diccionario o de permutación [3].

Aunque llevan más de una década estudiándose [4, 5], nadie había evaluado hasta ahora si los grandes modelos de lenguaje (como ChatGPT, Claude o Gemini) son capaces de detectarlos sin entrenamiento previo. En este trabajo respondemos a tres preguntas concretas. Primero, ¿con qué eficacia detectan estos modelos dominios maliciosos analizando únicamente el nombre de dominio, y mejora al proporcionarles características lingüísticas específicas? Segundo, ¿son capaces de distinguir entre diferentes familias de malware cuando se les dan ejemplos? Tercero, ¿cómo se comportan cuando clasifican dominios legítimos que se parecen a maliciosos? Para ello, evaluamos nueve modelos bajo distintas estrategias de consulta y medimos su rendimiento tanto en condiciones de laboratorio como frente a tráfico DNS real. Este resumen extendido resume el trabajo publicado en [6].

Configuración Experimental

Para evaluar la capacidad de los grandes modelos de lenguaje en la detección de AGDs, seleccionamos nueve modelos de cuatro proveedores distintos: GPT-4o y GPT-4o-mini (OpenAI), Claude 3.5 Sonnet y Claude 3.5 Haiku (Anthropic), Gemini 1.5 Pro, Flash y Flash-8B (Google), y Mistral Large y Small (MistralAI).

Construimos tres conjuntos de datos para cubrir distintos escenarios. El primero, D1 contiene 50.000 dominios equilibrados entre maliciosos, extraídos de DGArchive [3], y benignos, extraídos de la lista Tranco [7]. El segundo, D2, contiene 50.000 AGDs de 25 familias de malware distintas, representando esquemas aritméticos, basados en hash y de diccionario. El tercero, D3, contiene 50.000 dominios legítimos extraídos del tráfico DNS real de la Universidad de Zaragoza durante casi un año.

Para interactuar con los modelos diseñamos cuatro estrategias de consulta incrementales. P1 es una consulta mínima; P2 añade características léxicas propias de los dominios maliciosos; P3 incorpora ejemplos de las 25 familias de malware; y P4 añade contexto sobre dominios legítimos reales.

Detección de AGDs

Sin entrenamiento previo, los modelos alcanzan exactitudes de entre el 77,3% y el 89,4%, siendo Claude 3.5 Sonnet el mejor, seguido de Mistral Large y Gemini 1.5 Pro. En general, los modelos más grandes obtienen mejores resultados. Las diferencias entre P1 y P2 no superan el punto porcentual en ningún caso, por lo que el prompting avanzado no aporta mejoras significativas para esta tarea.

Una limitación crítica es la elevada tasa de falsos positivos, de entre el 13,2% y el 31,9%. En un entorno con un millón de consultas legítimas diarias,

esto equivaldría a bloquear erróneamente entre 132.000 y 319.000 dominios benignos cada día.

Clasificación por familias

Para la clasificación por familia de malware, proporcionamos a los modelos 10 ejemplos por familia (P3). Los resultados varían significativamente según el esquema de generación. Los DGAs basados en hash son los más fáciles de clasificar, con Claude 3.5 Sonnet alcanzando una precisión del 100% en esta categoría. Los DGAs aritméticos presentan resultados mixtos, y los de diccionario son los más difíciles, ya que los modelos tienen problemas para identificar dominios maliciosos compuestos por palabras naturales.

Rendimiento con dominios reales

Al evaluar los modelos sobre tráfico DNS real (D3) con P4, todos muestran una degradación sustancial del rendimiento respecto a los dominios benignos de D1. La exactitud disminuye entre un 11% y un 24%, lo que en un entorno con un millón de consultas diarias equivaldría a entre 110.000 y 240.000 falsos positivos adicionales al día, sumados a los ya obtenidos en la sección anterior. Esto evidencia que los LLMs tienen dificultades para distinguir dominios legítimos que comparten similitudes estructurales con los maliciosos.

Conclusiones

Este trabajo presenta la primera evaluación sistemática de los grandes modelos de lenguaje para detectar dominios generados algorítmicamente sin entrenamiento previo. Los resultados muestran que son capaces de distinguir dominios benignos y maliciosos con exactitudes de hasta el 89,4%, pero las altas tasas de falsos positivos y la degradación ante tráfico DNS real los hacen inviables como componente principal en sistemas de filtrado en producción.

Su uso más adecuado, en el estado actual, es como herramienta de apoyo en análisis forense o investigación, no como detector en tiempo real.

AGRADECIMIENTOS

Esta investigación ha sido financiada en parte por la ayuda PID2023-151467OA-I00 (CRAPER), financiada por MICIU/AEI/10.13039/501100011033

y por FEDER/UE; por la ayuda TED2021-131115A-I00 (MIMFA), financiada por MICIU/AEI/10.13039/501100011033 y por la Unión Europea NextGenerationEU/PRTR; por la ayuda Proyecto Estratégico Ciberseguridad EINA UNIZAR, financiada por el Instituto Nacional de Ciberseguridad (INCIBE) y la Unión Europea NextGenerationEU/PRTR; por la ayuda Programa de Proyectos Estratégicos de Grupos de Investigación (grupo de investigación DisCo, ref. T21-23R), financiada por el Dpto. de Universidad, Industria e Innovación del Gobierno de Aragón; y por el proyecto RAPID (Ayuda n.º CS.007) financiado por el Consejo de Investigación de los Países Bajos (NWO). Queremos expresar nuestro agradecimiento al equipo de DGArchive por habernos proporcionado el conjunto de datos actual por adelantado, lo que nos permitió comenzar la experimentación antes.

REFERENCIAS

- [1]. CYBERSECURITY VENTURES. *Cybercrime To Cost The World 10.5 Trillion Annually By 2025*. 2023. Disponible en: <https://cybersecurityventures.com/cybercrime-damages-6-trillion-by-2021/>
- [2]. PORRAS, Phillip A.; SAÏDI, Hassen; YEGNESWARAN, Vinod. A Foray into Confickers's Logic and Rendezvous Points. En: *LEET*. 2009.
- [3]. PLOHMANN, Daniel; YAKDAN, Khaled; KLATT, Michael; BARDER, Johannes; GERHARDS-PADILLA, Elmar. A Comprehensive Measurement Study of Domain Generating Malware. En *25th USENIX Security Symposium*. 2016, pp. 263-278.
- [4]. WOODBRIDGE, Jonathan; ANDERSON, Hyrum S.; AHUJA, Anjum; GRANT, Daniel. Predicting Domain Generation Algorithms with Long Short-Term Memory Networks. *arXiv preprint arXiv:1611.00791*. 2016
- [5]. CEBERE, Bogdan C.; FLUEREN, Jelle L. B.; SEBATIÁN, Salvador; PLOHMANN, Daniel; ROSSOW, Christian. Down to Earth! Guidelines for DGA-based Malware Detection. En: *Proceedings of the 27th RAID*. 2024, pp. 147-165.
- [6]. PELAYO-BENEDET, Tomás; RODRÍGUEZ, Ricardo J.; GAÑÁN, Carlos H. The Machines are Watching: Exploring the Potential of Large Language Models for Detecting Algorithmically Generated Domains. *Journal of Information Security and Applications*, 2025, vol. 93, 104176.
- [7]. TRANCO LIST. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. 2024. Disponible en: <https://tranco-list.eu/>