Una revisión de "*The Machines are Watching: Exploring the Potential of Large Language Models for Detecting Algorithmically Generated Domains*"

Tomás Pelayo-Benedet
Universidad de Zaragoza
España
tpelayo@unizar.es

Ricardo J. Rodríguez
Universidad de Zaragoza
España
ricardo@unizar.es

Carlos H. Gañán
Delft University of Technology
Países Bajos
C.HernandezGanan@tudelft.nl

***Resumen*—Los Dominios Generados Algorítmicamente (AGDs) son parte de muchas campañas de malware modernas, permitiendo a los atacantes establecer canales de mando y control (C&C) resilientes. Si bien las técnicas de aprendizaje automático se emplean cada vez más para detectar AGDs, el potencial de los Modelos de Lenguaje Grandes (LLMs) en este ámbito sigue siendo escasamente explorado. En este trabajo evaluamos la capacidad de nueve LLMs comerciales para identificar AGDs maliciosos sin ajuste de parámetros ni entrenamiento específico, mediante enfoques zero-shot y few-shot con estrategias de prompting. Los resultados muestran que los LLMs alcanzan precisiones de detección entre el 77,3 % y el 89,3 %. Sin embargo, surgen limitaciones significativas frente al tráfico DNS real, donde la degradación del rendimiento ante dominios benignos que guardan similitudes estructurales con dominios maliciosos evidencia el riesgo de falsos positivos en entornos reales.**

***Index Terms*—Modelos de Lenguaje Grandes, Dominios Generados Algorítmicamente, Análisis de Tráfico DNS, Detección de Malware**

Tipo de contribución: *Investigación ya publicada* [6]

I. Introducción

El cibercrimen se ha convertido en una industria altamente rentable que supera en ingresos al tráfico ilegal de drogas a nivel mundial [1]. Una técnica clave empleada por los atacantes son los Algoritmos de Generación de Dominios (DGA), que generan dinámicamente grandes cantidades de dominios posibilitando establecer conexión entre el malware y el centro de mando y control (C&C), dificultando que los defensores los bloqueen o rastreen [2]. Los DGA generan Dominios Generados Algorítmicamente (AGDs) mediante diversas técnicas pseudoaleatorias: métodos aritméticos, basados en hash, de diccionario y de permutación [3].

A pesar de más de 14 años de investigación en detección de AGDs ([4], [5]), no existen trabajos que utilicen directamente LLMs para esta tarea. Por ello, en nuestro trabajo evaluamos la capacidad de los LLMs para identificar AGDs sin ajuste de parámetros ni entrenamiento específico respondiendo a tres preguntas de investigación:

- **RQ1:** *¿Con qué eficacia detectan los LLMs AGDs maliciosos mediante análisis de cadenas de nombres de dominio? ¿Qué impacto tiene proporcionar características lingüísticas específicas de los DGAs?*
- **RQ2:** *¿En qué medida pueden los LLMs distinguir entre diferentes familias de malware, para ello proveyendo a los modelos de ejemplos de AGDs de cada familia?*
- **RQ3:** *¿Cómo se comportan los LLMs al evaluar dominios reales no maliciosos que comparten similitudes estructurales con AGDs maliciosos?*

Este resumen extendido resume el trabajo publicado en [6].

II. Configuración Experimental

Modelos evaluados. En nuestros experimentos hemos utilizado nueve LLMs de cuatro proveedores: GPT-4o y GPT-4o-mini (OpenAI), Claude Sonnet 3.5 y Haiku 3.5 (Anthropic), Gemini 1.5 Pro, Flash y Flash-8B (Google), y Mistral Large y Small (MistralAI).

Conjuntos de datos. Hemos construido tres *datasets*: D_1 con 50.000 dominios: 25.000 AGDs maliciosos de DGArchive [3] y 25.000 benignos de Tranco [7]; D_2 con 50.000 AGDs de 25 familias de malware representando esquemas aritmético, hash y diccionario; y D_3 con 50.000 dominios legítimos extraídos de los registros DNS de la Universidad de Zaragoza durante 347 días.

Estrategias de prompting. Para nuestros experimentos hemos diseñado cuatro prompts incrementales: P_1, mínimo; P_2, añade análisis de características léxicas de los dominios maliciosos; P_3, añade ejemplos de 25 familias de malware; y P_4 añade contexto de dominios reales. Hemos utilizado un tamaño de lote de 125 dominios por llamada a la API.

III. Detección de AGDs (RQ1)

Sin ajuste específico, los LLMs alcanzan exactitudes entre 77,3 % y 89,4 %. Claude 3.5 Sonnet obtiene el mejor rendimiento general (Tabla I). Podemos observar una tendencia: los modelos más grandes obtienen mejores resultados.

Una limitación crítica es la elevada tasa de falsos positivos (FPR), entre 13,2 % y 31,9 %. En un entorno con un millón de consultas legítimas diarias, esto supondría bloquear erróneamente entre 132 000 y 319 000 dominios benignos cada día. Como muestra la Tabla I, las diferencias entre P_1 y P_2 son marginales en todos los modelos: las variaciones en exactitud no superan el punto porcentual. Los resultados nos muestran que para la tarea de la clasificación de binaria de AGDs el *prompting* avanzado no aporta mejoras significativas.

IV. Clasificación por Familia de Malware (RQ2)

Para la clasificación multiclase, hemos seguido una estrategia en la que damos 10 ejemplos por familia (P_3), ya que en experimentos preliminares mostró el mejor equilibrio entre

Tabla I
RENDIMIENTO DE LOS LLMS EN DETECCIÓN BINARIA DE AGDS (D_1).

Modelo	P	Acc.	Prec.	Rec.	FPR
GPT-4o	P_1	86,8 %	83,6 %	91,4 %	17,9 %
	P_2	87,0 %	84,6 %	90,5 %	16,5 %
GPT-4o-mini	P_1	77,3 %	73,0 %	86,4 %	31,9 %
	P_2	78,5 %	74,6 %	86,5 %	29,4 %
Claude 3.5 Sonnet	P_1	89,3 %	83,8 %	97,4 %	18,8 %
	P_2	89,4 %	84,2 %	96,8 %	18,2 %
Claude 3.5 Haiku	P_1	85,6 %	84,0 %	87,9 %	16,8 %
	P_2	85,2 %	84,7 %	86,0 %	15,6 %
Gemini 1.5 Pro	P_1	87,7 %	83,8 %	93,5 %	18,1 %
	P_2	87,6 %	84,2 %	92,6 %	17,4 %
Gemini 1.5 Flash	P_1	84,8 %	83,5 %	86,9 %	17,2 %
	P_2	84,9 %	83,6 %	86,8 %	17,1 %
Gemini 1.5 Flash-8B	P_1	81,7 %	78,2 %	87,9 %	24,5 %
	P_2	82,7 %	79,8 %	87,6 %	22,1 %
Mistral Large	P_1	88,7 %	87,3 %	90,6 %	13,2 %
	P_2	88,5 %	87,1 %	90,5 %	13,4 %
Mistral Small	P_1	85,1 %	82,6 %	89,0 %	18,8 %
	P_2	85,5 %	83,7 %	88,1 %	17,1 %

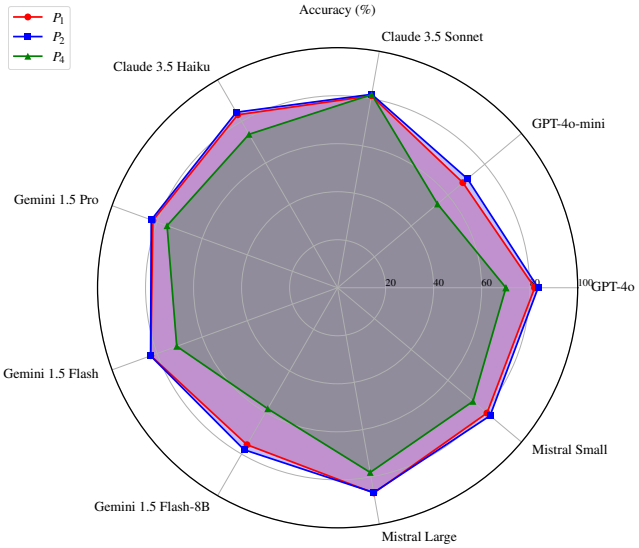

Figura 1. Exactitud de los LLMs sobre P_1, P_2 usando D_1 y P_4, D_3.

exactitud y eficiencia. Los resultados varían significativamente según el esquema de generación (D_2).

Los DGAs basados en hash obtienen clasificación casi perfecta en todos los modelos: Claude 3.5 Sonnet alcanza precisión del 100 % en esta categoría. Los DGAs aritméticos presentan resultados mixtos: familias como `metastealer`, `rovnix` y `zloader` mantienen tasas de deteción altas, pero `conficker` exhibe tasas más bajas a pesar de su mecanismo de generación más simple. Los DGAs de diccionario son los más difíciles de detectar, los modelos encuentran problemas a la hora de detectar dominios maliciosos compuestos por palabras naturales. Claude 3.5 Sonnet lidera con un F1-score global del 93,7 %, seguido de Gemini 1.5 Pro (79,6 %), Gemini 1.5 Flash-8B (56,8 %) y Mistral Large (41,4 %). El resto de modelos no son capaces de realizar esta tarea de clasificación multiclase ya que no superan el 40 % en F1-score.

V. RENDIMIENTO CON DOMINIOS REALES (RQ3)

Al evaluar sobre D_3 (tráfico DNS real) utilizando P_4, todos los modelos muestran una degradación sustancial del rendimiento, tal y como se aprecia en la Figura 1. Los valores de exactitud con P4 disminuyen entre un 11 % y un 24 % respecto a los dominios benignos de D_1 con P_1 y P_2. En un despliegue con un millón de consultas diarias, esto equivaldría a entre 110 000 y 240 000 falsos positivos adicionales al día, sumados a los falsos positivos que ya se obtenían.

VI. CONCLUSIONES

Este trabajo presenta la primera evaluación sistemática del uso de LLMs para detectar AGDs maliciosos sin ajuste de parámetros. Los resultados muestran que los LLMs son capaces de distinguir dominios benignos y maliciosos con exactitudes de hasta el 89,4 %. Sin embargo, las altas tasas de falsos positivos, la degradación ante tráfico DNS real y las restricciones de latencia y coste representan obstáculos fundamentales para su despliegue en producción.

Los LLMs resultan más adecuados, en su estado actual, para un análisis posterior como la investigación forense que como componentes primarios de detección en sistemas de filtrado DNS en tiempo real. El trabajo futuro explorará arquitecturas en dos etapas, ajuste fino de modelos de código abierto sobre conjuntos de datos de DGA, y técnicas de compresión que permitan inferencia a velocidad de línea.

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. [En línea; https://cybersecurityventures.com/cybercrime-damages-6-trillion-by-2021/].

[2] Porras, Phillip A., Hassen Saïdi, y Vinod Yegneswaran: “A Foray into Conficker’s Logic and Rendezvous Points.” En LEET 9: 7. 2009.

[3] Plohmann, D., Yakdan, K., Klatt, M., Bader, J., y Gerhards-Padilla, E.: “A comprehensive measurement study of domain generating malware”. En 25º *USENIX Security Symposium* (pp. 263-278). 2026.

[4] Woodbridge, J., Anderson, H. S., Ahuja, A., y Grant, D. *et al.*: “Predicting Domain Generation Algorithms with Long Short-Term Memory Networks”. *arXiv preprint arXiv:1611.00791*. 2016.

[5] Cebere, B. C., Flueren, J. L. B., Sebastián, S., Plohmann, D., y Rossow, C.: “Down to Earth! Guidelines for DGA-based Malware Detection”, en *Proc. 27th RAID*, pp. 147–165, 2024.

[6] T. Pelayo-Benedet, R. J. Rodríguez, y C. H. Gañán: “The Machines are Watching: Exploring the Potential of Large Language Models for Detecting Algorithmically Generated Domains”, *Journal of Information Security and Applications, 93, 104176.* 2025.

[7] Tranco List. [En línea; https://tranco-list.eu/]. Accedido en agosto de 2024.