A Review of: "The Machines are Watching: Exploring the Potential of Large Language Models for Detecting Algorithmically Generated Domains"

Tomás Pelayo-Benedet[†] Ricardo J. Rodríguez[†] Carlos H. Gañán[‡]

[†]Universidad de Zaragoza [‡]Delft University of Technology

May 6, 2026

JNIC'26

Universidad Zaragoza

TU Delft

Introduction

Journal of Information Security and Applications 93 (2025) 104176

Contents lists available at ScienceDirect

Journal of Information Security and Applications

journal homepage: www.elsevier.com/locate/jisa

The machines are watching: Exploring the potential of Large Language Models for detecting Algorithmically Generated Domains

Tomás Pelayo-Benedet [a], Ricardo J. Rodríguez [a,*], Carlos H. Gañán [b]

[a] Dpto. de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, Spain
[b] Delft University of Technology, The Netherlands

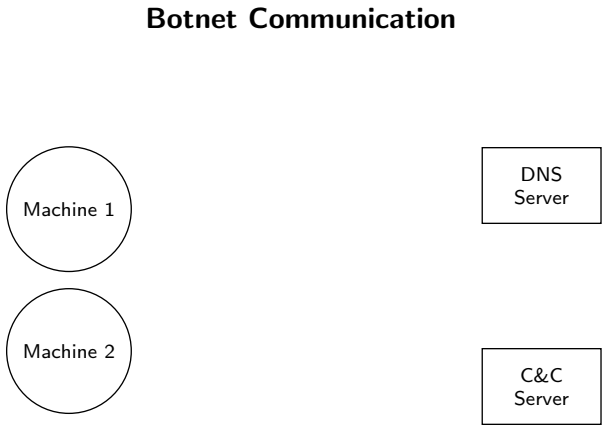
ARTICLE INFO

Keywords:
Large Language Models
Algorithmically Generated Domains
DNS traffic analysis
Malware detection

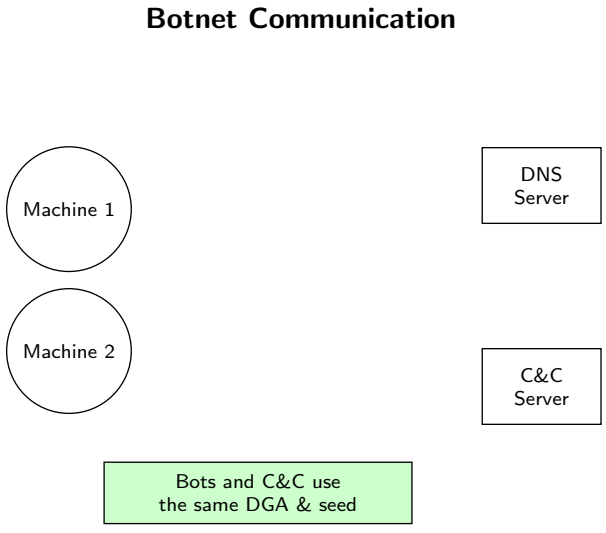
ABSTRACT

Algorithmically Generated Domains (AGDs) are integral to many modern malware campaigns, allowing adversaries to establish resilient command and control channels. While machine learning techniques are increasingly employed to detect AGDs, the potential of Large Language Models (LLMs) in this domain remains largely underexplored. In this paper, we examine the ability of nine commercial LLMs to identify malicious AGDs, without parameter tuning or domain-specific training. We evaluate zero-shot approaches and few-shot learning approaches, using minimal labeled examples and diverse datasets with multiple prompt strategies. Our results show that certain LLMs can achieve detection accuracy between 77.3% and 89.3%. In a 10-shot classification setting, the largest models excel at distinguishing between malware families, particularly those employing hash-based generation schemes, underscoring the promise of LLMs for advanced threat detection. However, significant limitations arise when these models encounter real-world DNS traffic. Performance degradation on benign but structurally suspect domains highlights the risk of false positives in operational environments. This shortcoming has real-world consequences for security practitioners, given the need to avoid erroneous domain blocking that disrupt legitimate services. Our findings underscore the practicality of LLM-driven AGD detection, while emphasizing key areas where future research is needed (such as more robust warning design and model refinement) to ensure reliability in production environments.

What are Domain Generation Algorithms (DGAs)?

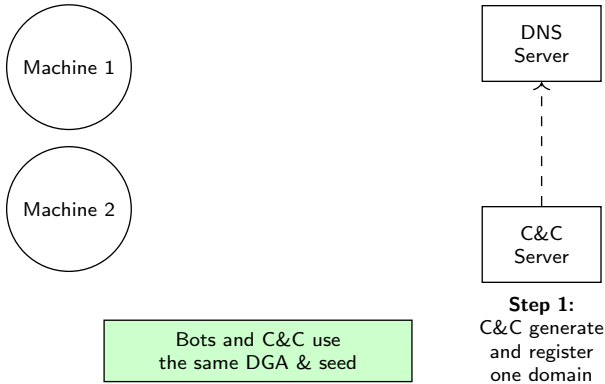


What are Domain Generation Algorithms (DGAs)?

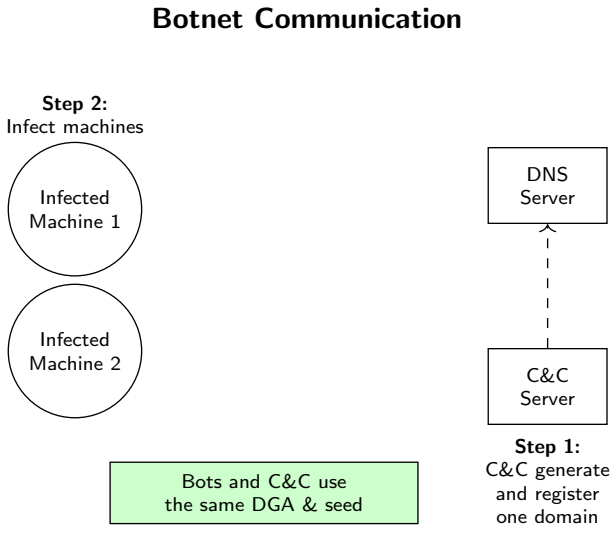


What are Domain Generation Algorithms (DGAs)?

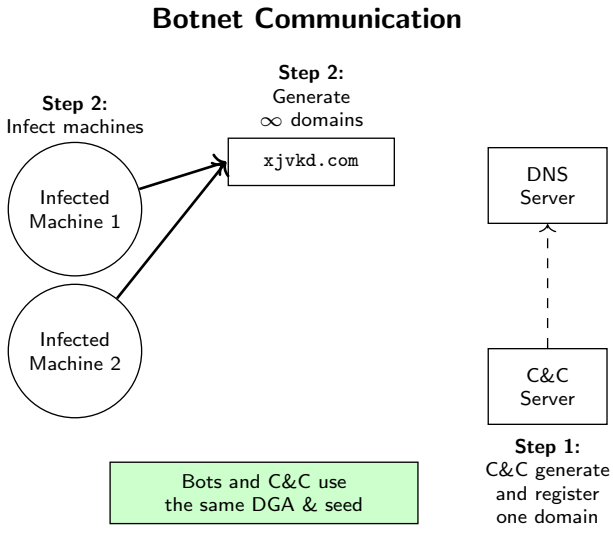
Botnet Communication



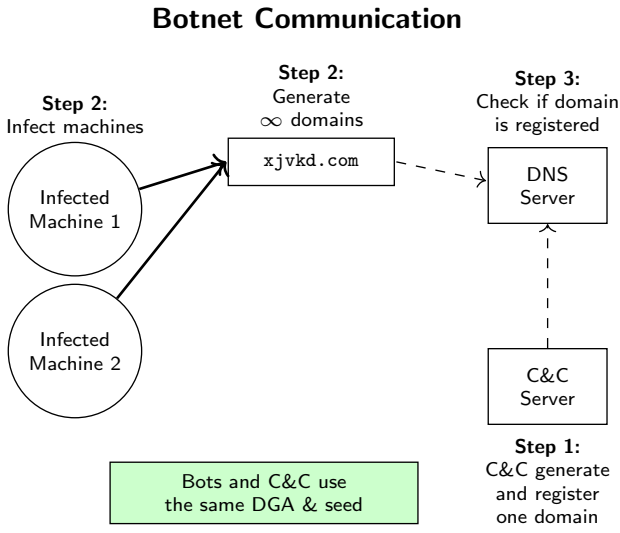
What are Domain Generation Algorithms (DGAs)?



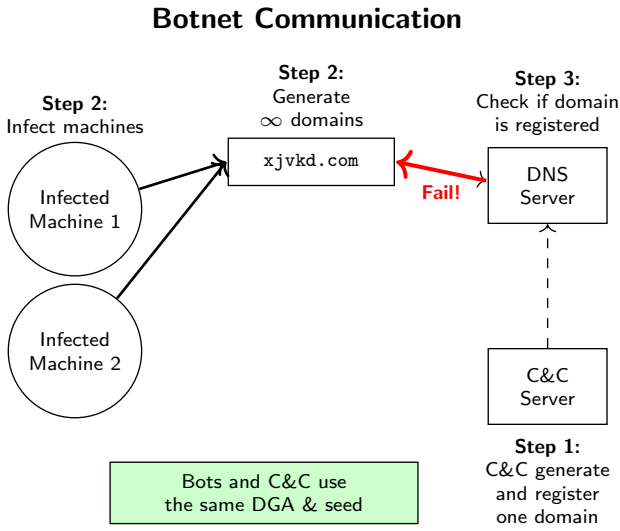
What are Domain Generation Algorithms (DGAs)?



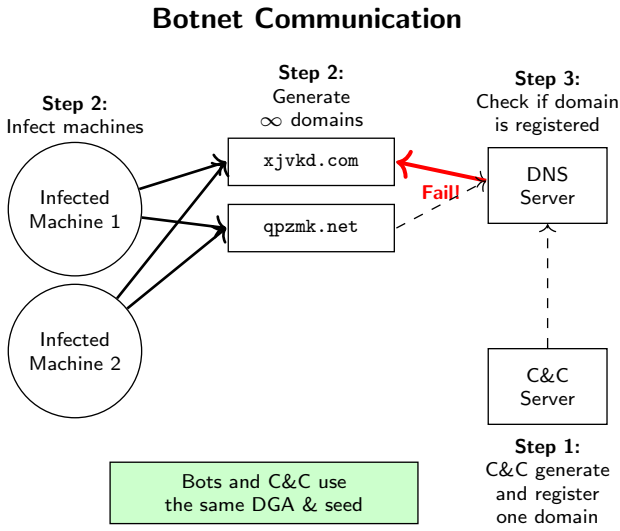
What are Domain Generation Algorithms (DGAs)?



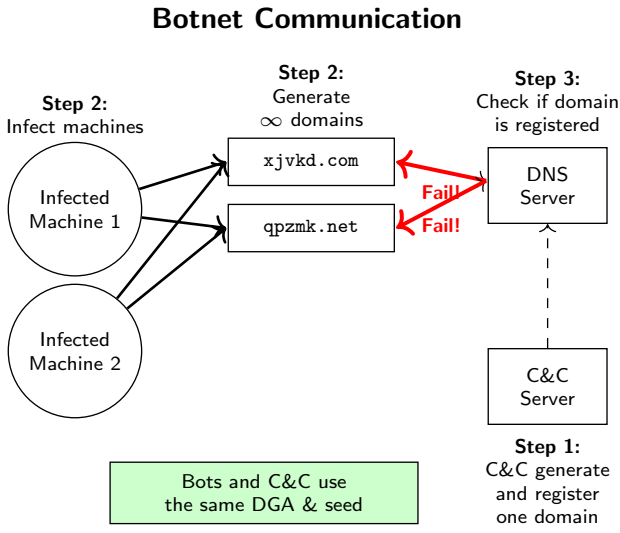
What are Domain Generation Algorithms (DGAs)?



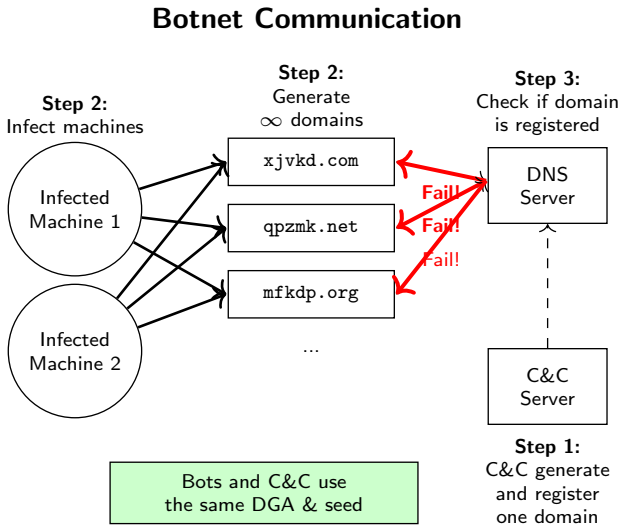
What are Domain Generation Algorithms (DGAs)?



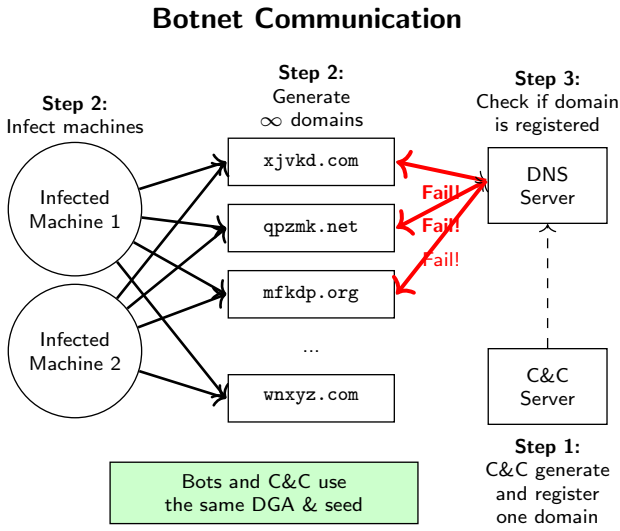
What are Domain Generation Algorithms (DGAs)?



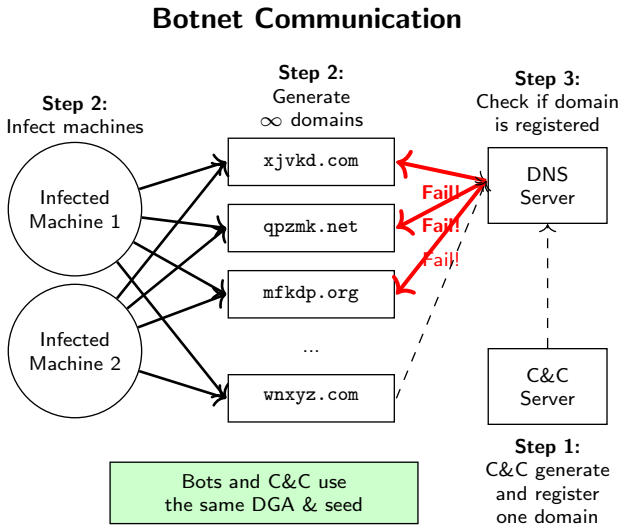
What are Domain Generation Algorithms (DGAs)?



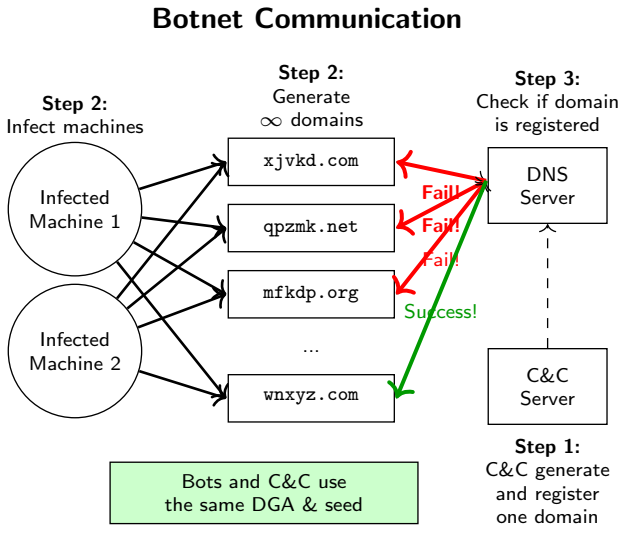
What are Domain Generation Algorithms (DGAs)?



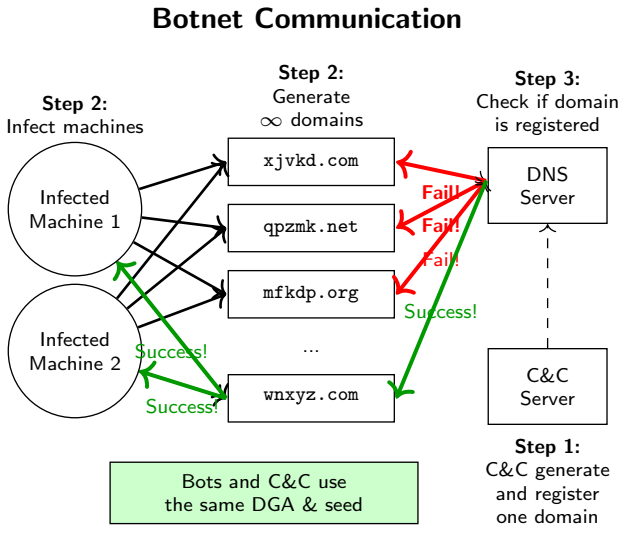
What are Domain Generation Algorithms (DGAs)?



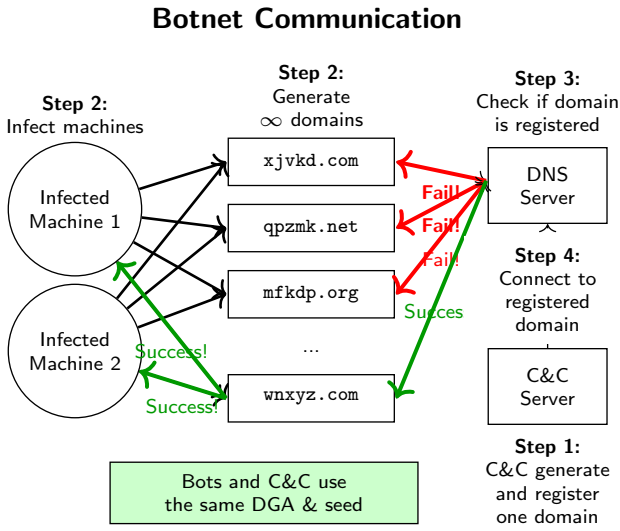
What are Domain Generation Algorithms (DGAs)?



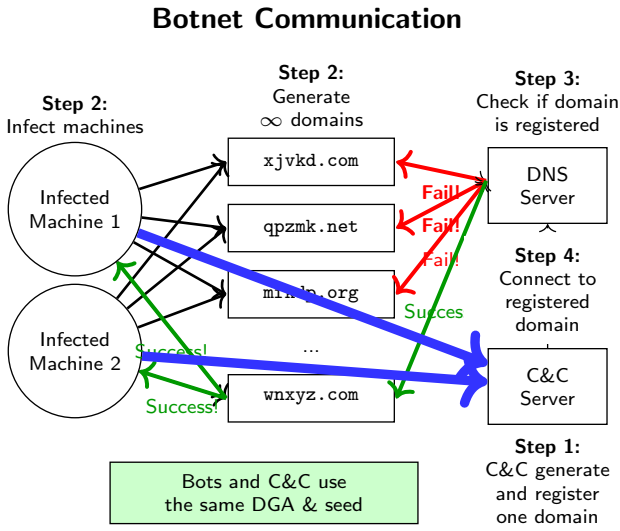
What are Domain Generation Algorithms (DGAs)?



What are Domain Generation Algorithms (DGAs)?



What are Domain Generation Algorithms (DGAs)?



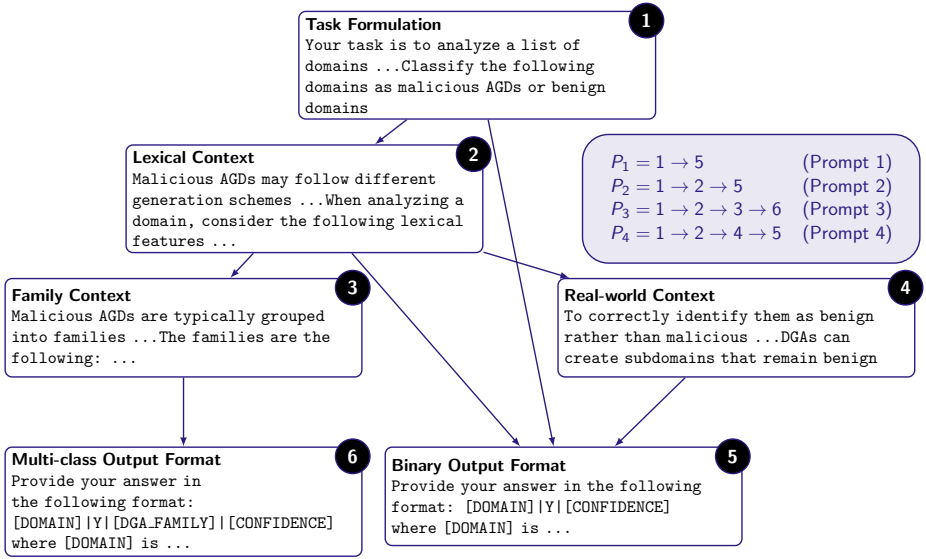
Methodology

- **Evaluated models:** 9 commercial LLMs



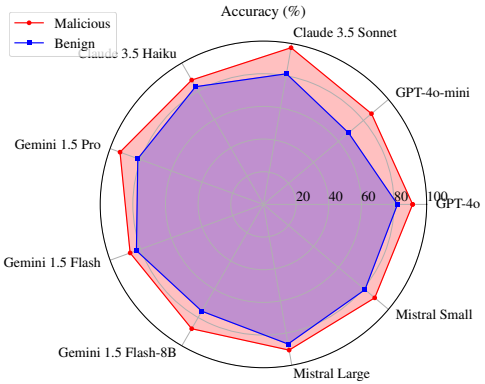

- **Datasets:**
  - **D1:** 25k malicious AGDs and 25k benign domains
  - **D2:** 50k AGDs from 25 malware families (multi-class evaluation)
  - **D3:** 50k real-world domains (DNS logs from Universidad de Zaragoza)
- **Data availability:** Dataset available upon request

Methodology: Prompts

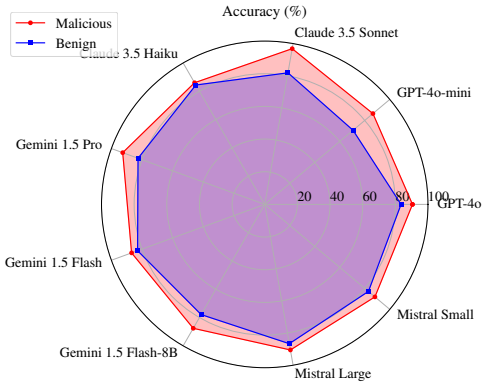


Detection Evaluation (D1)

- LLMs achieve accuracy rates between **77.3% and 89.3%** with no prior training
- High false positive rates remain a key limitation across all models



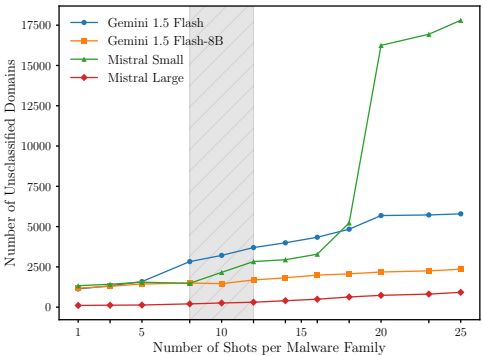
(a) Prompt strategy P_1



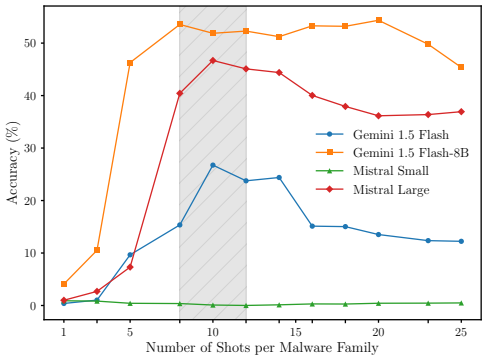
(b) Prompt strategy P_2

Malware Family Classification (D2)

- 10-shot learning provides the optimal balance between accuracy and performance



(a) Few-shot limitations



(b) Few-shot performance

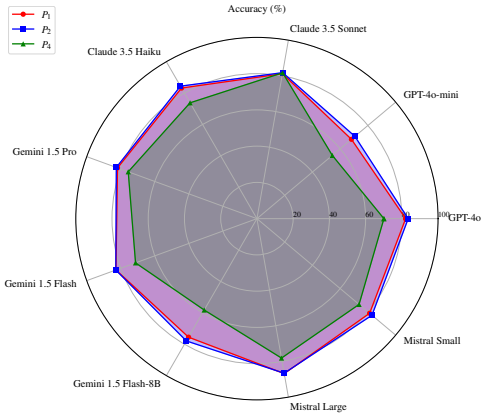
Malware Family Classification (D2)

- Excellent accuracy on *hash-based* schemes; struggles with *dictionary-based* DGAs

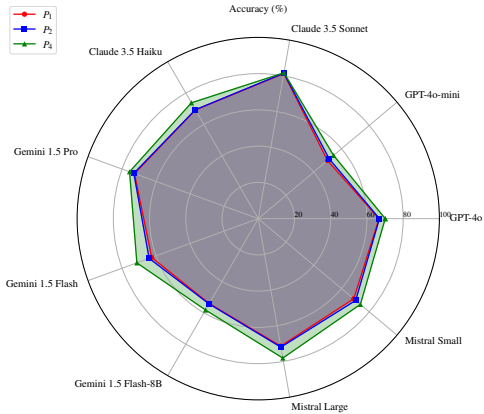
Family	Claude Sonnet 3.5			Gemini 1.5 Pro			Gemini 1.5 Flash-8B			Mistral Large			Type
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	
banjori	99.9	99.6	99.7	96.9	98.9	97.9	83.0	63.9	72.2	96.2	63.0	76.1	Arithmetic
conficker	96.9	97.6	97.2	85.9	69.1	76.6	54.1	31.1	39.5	76.5	14.1	23.8	
...													
virut	99.0	100.0	99.5	94.8	100.0	97.3	95.1	99.8	97.4	79.8	99.5	88.5	
zloader	100.0	100.0	100.0	97.1	97.3	97.2	70.9	93.4	80.6	74.7	44.6	55.8	
gozi	97.2	99.9	98.5	95.8	97.2	96.5	55.9	90.3	69.1	90.1	87.5	88.7	Dictionary
matsnu	94.4	99.6	96.9	80.5	92.6	86.1	75.2	95.1	84.0	7.9	63.0	14.0	
nymaim2	96.1	99.6	97.8	83.8	97.3	90.0	46.5	42.6	44.4	9.1	22.9	13.0	
suppobox	99.6	88.8	93.9	98.6	96.5	97.6	98.5	39.4	56.3	86.8	96.2	91.3	
darkwatchman	100.0	100.0	100.0	100.0	100.0	100.0	89.3	99.9	94.3	83.7	99.8	91.1	Hash
dyre	99.0	100.0	99.9	99.8	100.0	99.9	97.8	99.8	98.8	99.0	88.9	93.7	
grandoreiro	100.0	100.0	100.0	100.0	100.0	100.0	99.9	100.0	99.9	100.0	99.9	99.9	
monerominer	100.0	100.0	100.0	99.8	100.0	99.9	99.2	99.9	99.6	83.2	99.6	90.7	
pandabanker	100.0	100.0	100.0	100.0	99.8	99.9	98.8	97.6	98.2	97.8	60.7	74.9	
tinynuke	100.0	100.0	100.0	99.8	99.3	99.6	96.0	67.7	79.4	97.1	83.9	90.0	
wd	100.0	100.0	100.0	99.3	99.9	99.6	97.9	99.6	98.7	95.7	99.0	97.3	

Real-world Environment Evaluation (D3)

- Struggle to distinguish benign domains that share structural similarities with AGDs



(a) Benign domain accuracy: D_1 vs D_3



(b) Effect of prompt engineering on D_3

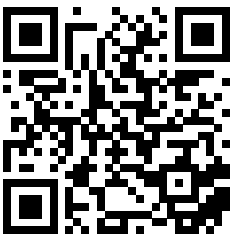
Conclusions

- LLMs show significant potential for detecting threats without fine-tunning
- High false positive rates on legitimate domains remain

Thank you!

Questions?

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



May 6, 2026

JNIC 2026