Exploring the Zero-Shot Potential of Large Language Models for Detecting Algorithmically Generated Domains

Tomás Pelayo-Benedet[1], Ricardo J. Rodríguez[1], Carlos H. Gañán[2]

[1]Dpto. de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, Spain
[2]Delft University of Technology, the Netherlands

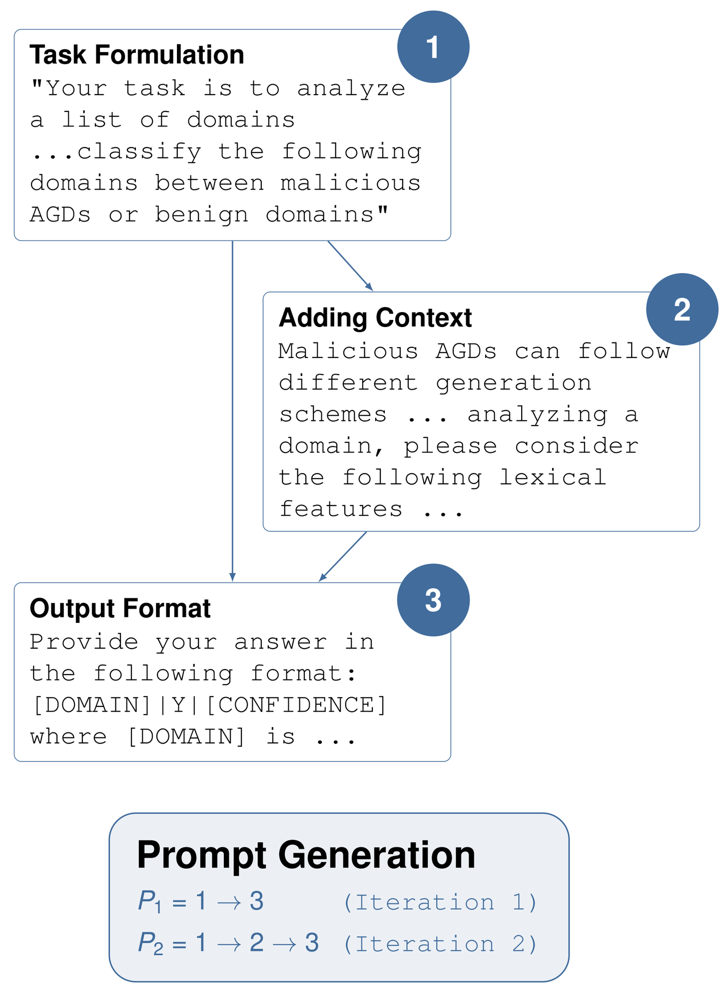
Domain Generation Algorithms (DGAs)

- First observed in the `Conficker` malware family [2]
- A DGA generates domain names similarly to a pseudo-random number generator. These are known as *Algorithmically Generated Domains* (AGDs)
- Examples of AGDs [1]:
  `accident-be-kind.com`,
  `seprfyswjugpvldkrwwg.com`,
  `kljinjhfqdynzbylayizx.ru`,
  `7f6fb68d7aac2de485ac1256503bb5c0.com`

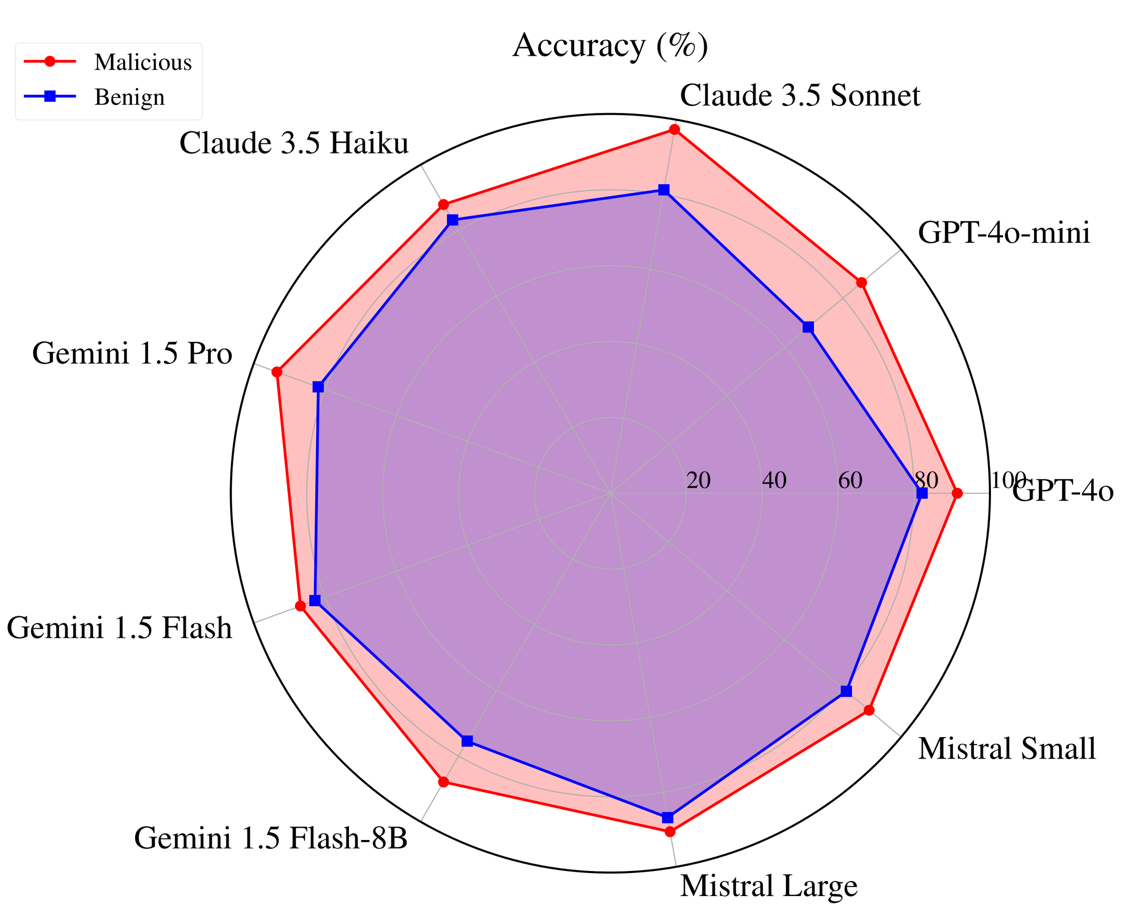
Large Language Models (LLMs)

- Traditional AGD detection struggles to generalize to new or obfuscated domains. LLMs offer a promising alternative by leveraging pre-trained linguistic knowledge without requiring task-specific tuning
- In this work, LLMs are evaluated in a zero-shot setting, using only their pre-trained knowledge to detect malicious AGDs

Prompt Crafting



Malicious AGD Bias

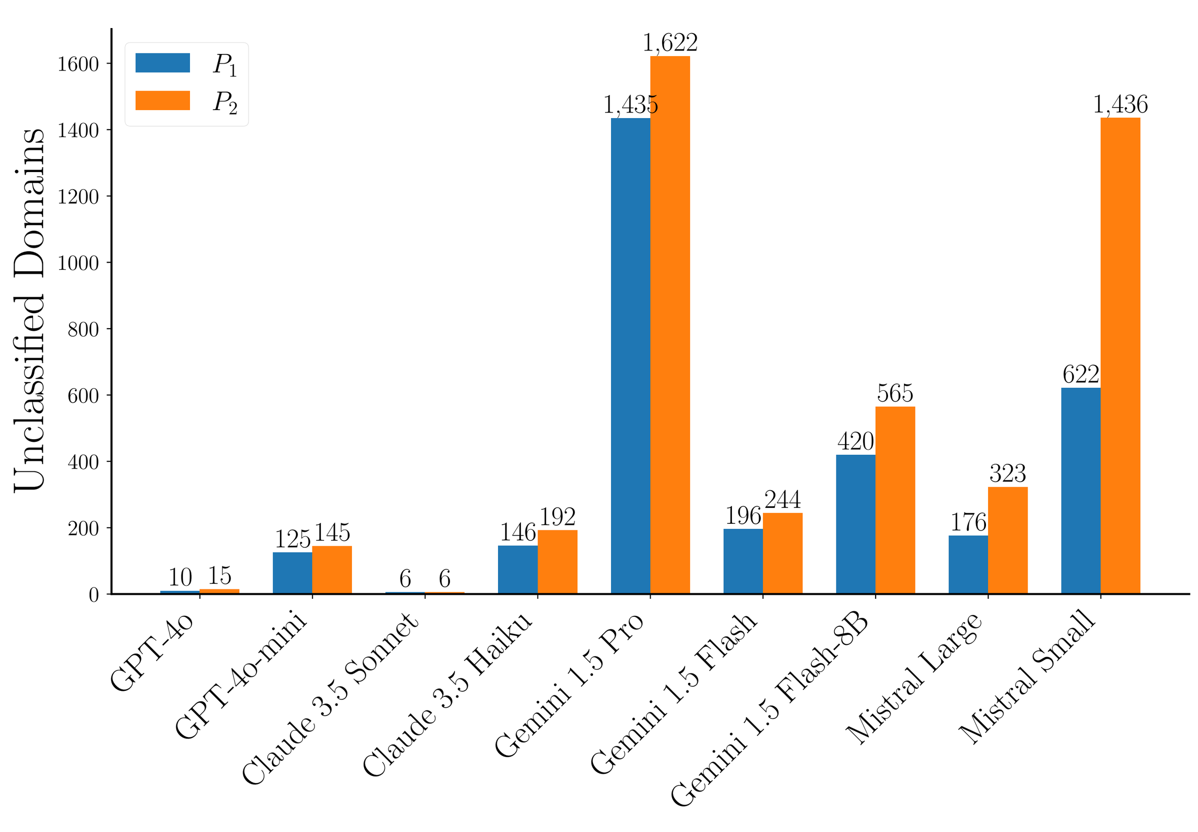


General Performance

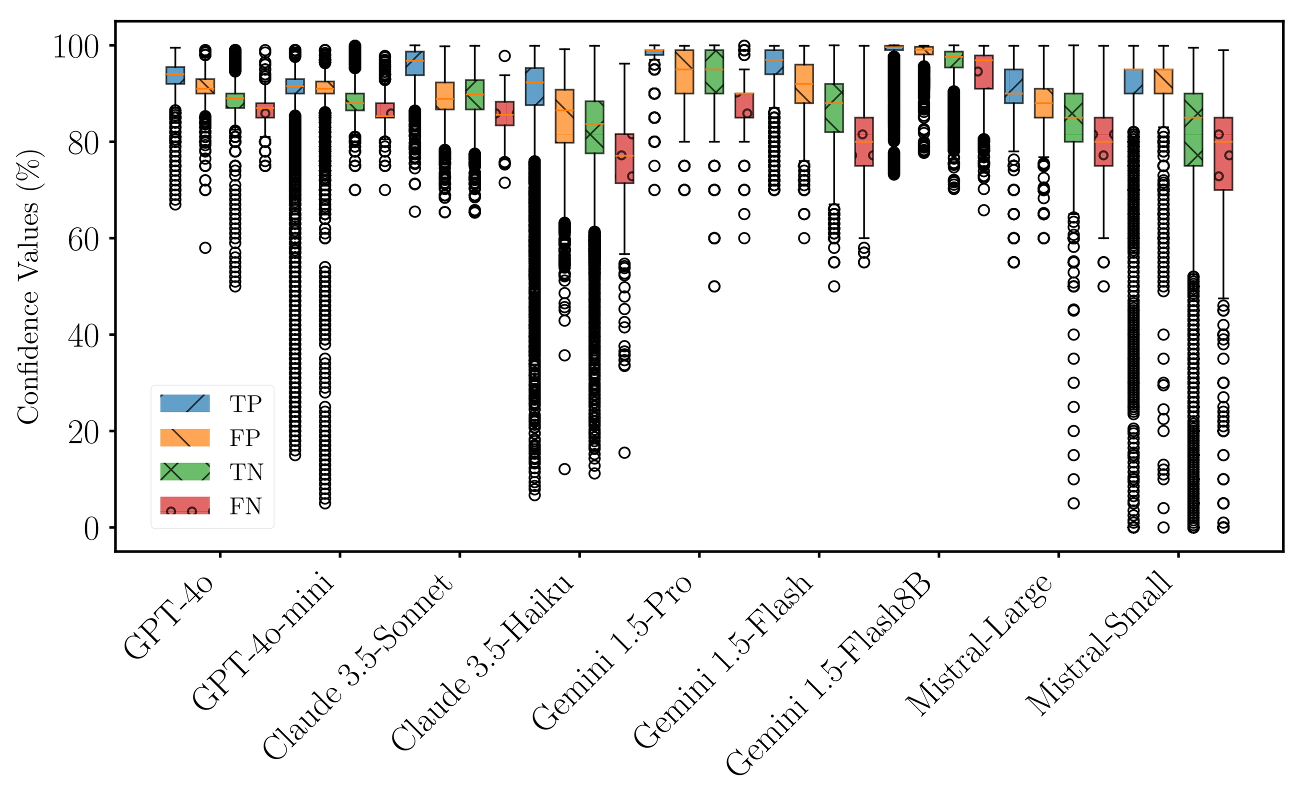
Model	P	Acc	Prec	Rec	F1	FPR	TPR	MCC	κ
GPT-4o	P_1	86.80	83.60	91.40	87.30	17.90	91.40	73.80	0.603
	P_2	87.00	84.60	90.50	87.40	16.50	90.50	74.20	0.603
GPT-4o-mini	P_1	77.30	73.00	86.40	79.20	31.90	86.40	55.40	0.415
	P_2	78.50	74.60	86.50	80.10	29.40	86.50	57.80	0.435
Claude 3.5 Sonnet	P_1	89.30	83.80	**97.40**	**90.10**	18.80	**97.40**	**79.70**	**0.682**
	P_2	**89.40**	84.20	96.80	**90.10**	18.20	96.80	79.50	0.678
Claude 3.5 Haiku	P_1	85.60	84.00	87.90	85.90	16.80	87.90	71.20	0.563
	P_2	85.20	84.70	86.00	85.40	15.60	86.00	70.50	0.548
Gemini 1.5 Pro	P_1	87.70	83.80	93.50	88.40	18.10	93.50	76.00	0.632
	P_2	87.60	84.20	92.60	88.20	17.40	92.60	75.60	0.625
Gemini 1.5 Flash	P_1	84.80	83.50	86.90	85.10	17.20	86.90	69.70	0.544
	P_2	84.90	83.60	86.80	85.20	17.10	86.80	69.80	0.545
Gemini 1.5 Flash-8B	P_1	81.70	78.20	87.90	82.80	24.50	87.90	63.90	0.494
	P_2	82.70	79.80	87.60	83.50	22.10	87.60	65.80	0.510
Mistral Large	P_1	88.70	**87.30**	90.60	88.90	**13.20**	90.60	77.40	0.639
	P_2	88.50	87.10	90.50	88.80	13.40	90.50	77.10	0.636
Mistral Small	P_1	85.10	82.60	89.00	85.70	18.80	89.00	70.40	0.560
	P_2	85.50	83.70	88.10	85.80	17.10	88.10	71.10	0.562

P: Prompt; **Acc**: Accuracy; **Prec**: Precision; **Rec**: Recall; **F1**: F1-score; **FPR**: False Positive Rate; **TPR**: True Positive Rate; **MCC**: Matthews's Correlation Coefficient; κ: Cohen's Kappa Score

Unclassified Domains



Confidence in Response



Our Dataset

- 50k domains (randomly selection)
  - 25k legitimate domains [3]
  - 25k malicious domains from 25 different malware families (1k per family) [1]

References

[1] Plohmann, D., Yakdan, K., Klatt, M., Bader, J., Gerhards-Padilla, E.: A Comprehensive Measurement Study of Domain Generating Malware. In: 25th USENIX Security Symposium (USENIX Security 16). pp. 263–278. USENIX Association, Austin, TX (Aug 2016)

[2] Porras, P.A., Saïdi, H., Yegneswaran, V.: A Foray into Conficker's Logic and Rendezvous Points. LEET **9**, 7 (2009)

[3] Tranco: Tranco List. [Online; `https://tranco-list.eu/`] (2024), accessed on August 15, 2024.

Conclusions

- LLMs demonstrate **significant capabilities** for detecting malicious domains **as a zero-shot classification task**, highlighting their potential for transfer learning
- However, they exhibit a **consistent bias toward malicious classification**, which often favors threat identification at the cost of increased false positive, posing challenges for real-world deployment
- **Future research** focuses on extending this work to **multiclass classification** and evaluating LLMs on **real-world, non-malicious domains** that resemble AGDs in structure

Try It!

Acknowledgements

This research was supported in part by grant PID2023-151467OA-I00 (CRAPER), funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, by grant TED2021-131115A-I00 (MIMFA), funded by MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, and the University of Zaragoza, by grant *Proyecto Estratégico Ciberseguridad EINA UNIZAR*, funded by the Spanish National Cybersecurity Institute (INCIBE) and the European Union NextGenerationEU/PRTR, by grant *Programa de Proyectos Estratégicos de Grupos de Investigación* (DisCo research group, ref. T21-23R), funded by the University, Industry and Innovation Department of the Aragonese Government, and by the RAPID project (Grant No. CS.007) financed by the Dutch Research Council (NWO). We extend our gratitude to the DGArchive team for providing the current dataset in advance, allowing us to begin experimentation sooner.

Financiado por la Unión Europea NextGenerationEU

Gobierno de España — Ministerio para la Transformación Digital y de la Función Pública — Secretaría de Estado de Digitalización e Inteligencia Artificial

Plan de Recuperación, Transformación y Resiliencia

incibe_ Instituto Nacional de Ciberseguridad

Contact data: reverseame@unizar.es

DIMVA 2025 (Graz, Austria)