IAs génératives (LLMs) pour détecter des comportement anormaux par analyse de traces d'exécution

1 Contexte et approche

Ce projet consiste à étudier les apports des Intelligences Artificielles (IA) génératives et des Grands Modèles de Langues (LLMs) pour certains aspects de la lutte informatique défensive. Ce stage est l'occasion d'initier un travail de recherche qui se poursuivra en thèse, en collaboration avec DiverSE Inria et le Laboratoire Exploration et recherche en Détection (LED) à l'ANSSI.

Ambition. L'objectif est la création d'un programme de supervision permettant de détecter et caractériser automatiquement le fait qu'un système informatique sorte de son comportement nominal (y compris dans ses interactions avec l'extérieur). Le superviseur peut ainsi remonter des alertes. Le résultat de l'analyse est un rapport actionnable par des experts.

Approche et méthodologie. Dans ce contexte, les LLMs sont prometteurs pour analyser les traces d'exécution (en classifiant, produisant un résumé, ou en extrayant les informations importantes d'une ou plusieurs traces). Les LLMs ont été mis au-devant de la scène récemment avec des initiatives et outils comme BERT, BLOOM, GPT-3, GPT-4, PaLM, Alphacode, Code-Parrot, Codex, ChatGPT, ou encore CoPilot. La capacité à des LLMs à traiter ou à synthétiser des artefacts techniques (du code, des documents semi-structurés, ou des traces) nous incite à explorer leurs usages dans un contexte de cybersécurité [Liu et al., 2021, Steenhoek et al., 2022, Zhou et al., 2022]. Il s'agit alors d'étudier les LLMs dans le contexte de la **détection de comportements anormaux des programmes et systèmes informatiques** [Vaccaro and Liepins, 1989, Oliner et al., 2011, Li et al., 2017, Sultana et al., 2019, Khraisat et al., 2019, Thakkar and Lohiya, 2023].

Pour ce faire, des **traces d'exécutions** (p. ex. des journaux) de plusieurs types (appels systèmes [da Costa et al., 2017, Nissim et al., 2018], mémoire [Panker and Nissim, 2021], échanges/paquets réseau [Sikos, 2020], etc.) seront collectées. Les traces d'exécution peuvent être vues comme du texte obéissant à certaines règles: ce sont des données semi-structurées.

Or les Grands Modèles de Langues ont montré leur capacité à traiter ce type de données de façon agnostique et générique, c'est-à-dire sans avoir besoin d'analyse syntaxique ou grammaticale. Grâce à leur malléabilité, les LLMs devraient avoir une excellente capacité à classifier les comportements (c.-à-d. exécutions) anomaux de programmes et systèmes peuvent ainsi être détectés, révélant erreurs et bogues, des logiciels malveillants ou encore des cyber-attaques.

Le système mis en place devra prendre en compte les outils, catalogues et bases de données de vulnérabilités préexistants, pour relier dans la mesure du possible les détections à ces vulnérabilités (p. ex. des CVEs). Des techniques de plongements (embeddings) et de recherche d'information devront être développées pour rendre efficace l'interaction entre LLMs, traces, et sources de données [Liu et al., 2021, Andrus et al., 2022]. Notre vision est de pouvoir synthétiser des rapports qui arrivent à faire correspondre traces et informations de vulnérabilité; ces rapports peuvent être exploités par des experts pour prendre des décisions défensives;

Architecture du projet. La Figure 1 donne un aperçu général du projet. Étant donné un cybersystème (boîte noire), il est possible de l'observer via des traces (boîtes grises). D'un point de vue défensif, ces traces peuvent être analysées pour quantifier et qualifier le cybersystème en termes de vulnérabilités ou d'attaques en cours.

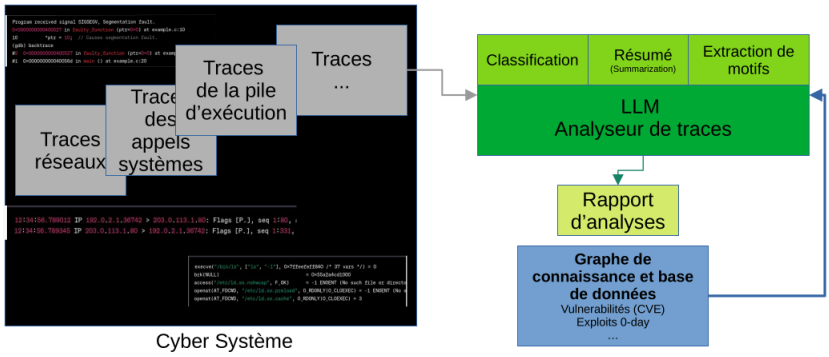

FIG. 1 – *Lutte informatique défensive (analyse de traces) basée sur les LLMs*

2 Travail du stage

Le travail à accomplir se décline en trois axes:

- étudier la bibliographie afin d'acquérir une bonne compréhension des domaines concernés, et des outils existants. Les références citées dans ce document sont un point de départ, mais l'état de l'art évolue rapidement, que ce soit côté LLM, génie logiciel, ou sécurité.
- à partir du travail bibliographique, et en collaboration avec l'ANSSI, concevoir d'un terrain de jeu avec des cyber systèmes, des traces, etc. pour pouvoir expérimenter avec les LLMs. On pourra se servir de données ouvertes ou de scénarios réalistes, et d'établir un banc de tests avec l'ambition à terme d'avoir des résultats référence pour la détection de comportements anormaux à partir de traces d'exécutions.
- implémenter un prototype expérimental de LLM détectant des comportements anormaux par analyse de trace d'un type donné. Ce prototype sera développé par le stagiaire à partir des articles et en réutilisant des bibliothèques ou des outils disponibles comme logiciels libres. Les résultats expérimentaux seront reportés, analysés, et discutés.

Le but du stage est de se familiariser avec le sujet et d'obtenir des premiers résultats qui seront ensuite développés **dans le cadre d'une thèse de 3 ans, toujours en partenariat entre DiverSE Inria et l'ANSSI.**

3 Encadrement et contacts

Le stage se déroulera dans l'équipe DiverSE d'Inria/IRISA Rennes, en collaboration avec le LED à l'ANSSI.

L'équipe DiverSE a une expertise internationale reconnue en sciences du logiciel (software engineering), en variabilité logicielle, en techniques automatiques pour le logiciel. DiverSE a une forte activité autour de la cybersécurité via des collaborations passées ou en cours, par exemple récemment avec Software Heritage (SWH-Sec). DiverSE est co-responsable d'un défi Inria autour des LLMs et du software engineering.

L'Agence nationale de la sécurité des systèmes d'information (ANSSI) est l'autorité nationale en matière de cybersécurité. Sa mission est de comprendre, prévenir et répondre au risque cyber. Le LED est responsable du domaine de la détection et de l'analyse des attaques informatiques contre les systèmes d'information, incluant notamment la détection d'intrusion, l'analyse de systèmes compromis ou de logiciels malveillants.

Encadrants:
Mathieu ACHER, Professeur à l'INSA Rennes (mathieu.acher@inria.fr), DiverSE.
Olivier ZENDRA, Chargé de Recherche Inria (olivier.zendra@inria.fr), DiverSE.
Romain BRAULT, Expert en Science des Données à l'ANSSI (romain.brault@ssi.gouv.fr), LED.

Le but du stage est de préparer le candidat pour un travail de recherche qui se poursuivra par une thèse de trois ans, réalisée en collaboration entre DiverSE Inria et l'ANSSI.

Références

[Andrus et al., 2022] Andrus, B. R., Nasiri, Y., Cui, S., Cullen, B., and Fulda, N. (2022). Enhanced story comprehension for large language models through dynamic document-based knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10436–10444.

[da Costa et al., 2017] da Costa, V. G. T., Barbon, S., Miani, R. S., Rodrigues, J. J. P. C., and Zarpelão, B. B. (2017). Detecting mobile botnets through machine learning and system calls analysis. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–6.

[Khraisat et al., 2019] Khraisat, A., Gondal, I., Vamplew, P., and Kamruzzaman, J. (2019). Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecur*, 2(20).

[Li et al., 2017] Li, T., Jiang, Y., Zeng, C., Xia, B., Liu, Z., Zhou, W., Zhu, X., Wang, W., Zhang, L., Wu, J., Xue, L., and Bao, D. (2017). FLAP: an end-to-end event log analysis platform for system management. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 1547–1556. ACM.

[Liu et al., 2021] Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., and Chen, W. (2021). What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

[Nissim et al., 2018] Nissim, N., Lapidot, Y., Cohen, A., and Elovici, Y. (2018). Trusted system-calls analysis methodology aimed at detection of compromised virtual machines using sequential mining. *Knowledge-Based Systems*, 153:147–175.

[Oliner et al., 2011] Oliner, A. J., Ganapathi, A., and Xu, W. (2011). Advances and challenges in log analysis. *Queue*, 9:30 – 40.

[Panker and Nissim, 2021] Panker, T. and Nissim, N. (2021). Leveraging malicious behavior traces from volatile memory using machine learning methods for trusted unknown malware detection in linux cloud environments. *Knowledge-Based Systems*, 226:107095.

[Sikos, 2020] Sikos, L. F. (2020). Packet analysis for network forensics: A comprehensive survey. *Forensic Science International: Digital Investigation,*, 32:200892.

[Steenhoek et al., 2022] Steenhoek, B., Rahman, M. M., Jiles, R., and Le, W. (2022). An empirical study of deep learning models for vulnerability detection. *arXiv preprint arXiv:2212.08109*.

[Sultana et al., 2019] Sultana, N., Rao, A., Jin, Z., Pashakhanloo, P., Zhu, H., Yegneswaran, V., and Loo, B. T. (2019). Trace-based behaviour analysis of network servers. In Lutfiyya, H., Diao, Y., Zincir-Heywood, A. N., Badonnel, R., and Madeira, E. R. M., editors, *15th International Conference on Network and Service Management, CNSM 2019, Halifax, NS, Canada, October 21-25, 2019*, pages 1–5. IEEE.

[Thakkar and Lohiya, 2023] Thakkar, A. and Lohiya, R. (2023). A review on challenges and future research directions for machine learning-based intrusion detection system. *Arch Computat Methods Eng*.

[Vaccaro and Liepins, 1989] Vaccaro, H. and Liepins, G. (1989). Detection of anomalous computer session activity. In *Proceedings. 1989 IEEE Symposium on Security and Privacy*, pages 280–289.

[Zhou et al., 2022] Zhou, Z., Bo, L., Wu, X., Sun, X., Zhang, T., Li, B., Zhang, J., and Cao, S. (2022). Spvf: security property assisted vulnerability fixing via attention-based models. *Empirical Software Engineering*, 27(7):171.