Cas d'usage : Cybersécurité

Contributeur :
Eric Savignac
Expert - Airbus DS

Introduction

L'hyper connectivité engendrée par Internet entre les systèmes d'information des entreprises, leurs systèmes industriels, les équipements des utilisateurs (téléphones mobiles, tablettes…), et les objets connectés (caméras, appareils ménagers…) ainsi que la tendance à une digitalisation des services lorsque cela est possible (banques, assurances, impôts, achats de biens physiques en ligne…), ont engendré de multiples possibilités de fraudes, de vols d'informations sensibles, et d'attaques via Internet à des coûts faibles pour un attaquant.

L'avènement de l'intelligence artificielle (Machine Learning, Deep Learning et intelligence artificielle générative), a aussi changé les approches des cyber attaquants et des cyber défenseurs en offrant aux premiers des possibilités de construire plus efficacement des attaques Cyber et aux seconds des moyens de détection, de compréhension et d'analyses de ces attaques, ainsi que des solutions de sensibilisation et d'entraînement adaptées. L'utilisation de l'intelligence artificielle en Cybersécurité, qui a déjà fait l'objet d'une publication du Hub France IA[1], est donc bien une réalité. Pour en comprendre la problématique, il est indispensable de s'appuyer sur des modèles explicatifs des modes d'attaques et de défense en Cybersécurité.

Image générée par Microsoft Designer

[1] Hub France IA. Les usages de l'IA Générative. Janvier 2024. https://www.hub-franceia.fr/wp-content/uploads/2024/02/Livre-blanc_Les-usages-de-lia-generative-01.2024.pdf

Cas d'usage : Cybersécurité

Contributeur :
Eric Savignac
Expert - Airbus DS

Modélisation des attaques et des défenses

Le déroulement d'une attaque Cyber peut être soit modélisé sous la forme de la Kill Chain définie par LOCKHEED MARTIN[2] avec ses 7 phases (*reconnaissance, weaponization, delivery, exploitation, installation, command & control, actions on objectives*), soit représenté à travers les 14 tactiques (*reconnaissance, resource development, initial access, execution, persistence, privilege escalation, defense evasion, credential access, discovery, lateral movement, collection, command and control, exfiltration, impact*) et les 235 techniques décrites dans la matrice MITRE ATT&CK®[3].

Parallèlement, il existe aussi des modèles pour formaliser les moyens de la cyberdéfense. Pour se protéger face à une attaque Cyber, les contre-mesures de sécurité à mettre en place peuvent être dérivées par exemple, des 6 tactiques (*harden, detect, isolate, deceive, evict et restore*) et des 168 techniques associées de la matrice MITRE D3FEND™[4].

Du point de vue du défenseur, l'utilisation d'algorithmes de Machine Learning ou de Deep Learning pour améliorer sa posture de sécurité, en particulier pour détecter des logiciels malveillants, ou pour identifier des canaux de commande et de contrôle permettant à un attaquant de piloter son attaque dans un système d'information, ou pour s'assurer qu'il n'y pas d'exfiltration de données de l'entreprise vers l'extérieur à travers ses connexions Internet, n'est pas nouvelle. Ces techniques de détection basées sur l'intelligence artificielle peuvent être embarquées dans des solutions ou dans des produits de sécurité commerciaux, ou simplement déployées par l'entreprise elle-même.

Inversement, un cyber attaquant peut utiliser un réseau de neurones convolutifs (CNN ou *Convolutional Neural Networks*) pour contourner automatiquement les mécanismes de CAPTCHA (*Completely Automated Public Turing test to tell Computers and Humans Apart*) basés sur des images, qui permettent de s'assurer que l'entité qui se connecte est bien un être humain. Il peut aussi, en s'inspirant des rapports de travaux de recherche accessibles publiquement, tester ses codes malveillants à travers des algorithmes de Machine Learning ou de Deep Learning pour s'assurer qu'ils ne soient pas détectés. De nombreux travaux de recherche ont été publiés sur l'utilisation d'algorithmes de Machine Learning ou de Deep Learning en cybersécurité, comme par exemple pour la détection de codes malveillants [5, 6, 7].

Cas d'usage : Cybersécurité

Contributeur :
Eric Savignac
Expert - Airbus DS

L'utilisation de l'IA générative

Mais, l'avènement récent de l'intelligence artificielle générative, qui facilite la création de nouveaux contenus (images, vidéos, code…), a aussi entraîné une extension des possibilités de l'attaquant qui peut maintenant générer des mails de phishing, dans plusieurs langues, parfaitement adaptés au contexte de l'entreprise ciblée. Il peut aussi s'appuyer sur des agents conversationnels tels que WormGPT[8], ou FraudGPT[9], qui sont disponibles sur le *dark web*, pour optimiser ses attaques, en particulier avec la création de codes malveillants. Dans son papier[10], Polra Victor Falade explique comment ces outils d'intelligence artificielle générative peuvent être utilisés astucieusement dans des attaques en social engineering.

Par ailleurs, cette apparition de l'intelligence artificielle a entrainé la création de LLMs (Large Language Models) spécifiques au domaine de la cybersécurité comme CySecBERT[11] et SecureBERT[12].

Elle a aussi suscité la fourniture d'outils basés sur ces technologies par des éditeurs, comme par exemple Microsoft Copilot for Security[13] totalement dédié à la protection des organisations contre les attaques cyber[14].

[2] Lockheed Martin. The Cyber Kill Chain. https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html et Eric M. Hutchins, Michael J. Cloppert, Rohan M. Amin. Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. *Leading Issues in Information Warfare & Security Research*. Vol. 1 n°1, pp. 113-125. 2011. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=ca18aa98d4d1d434802eec54c2ba6ea8cf493b88#page=123

[3] MITRE. ATT&CK knowledge base. https://attack.mitre.org/

[4] MITRE. D3FEND, a knowledge graph of cybersecurity countermeasures. https://d3fend.mitre.org/

[5] Lolitha Sresta Tupadha, Mark Stamp. Machine Learning for Malware Evolution Detection. *arXiv preprint arXiv: 2107.01627v1*. July 6, 2021. https://arxiv.org/pdf/2107.01627

[6] Pascal Maniriho, Abdun Naser Mahmood, Mohammad Jabed Morshed Chowdhury. Deep Learning Models for Detecting Malware Attacks. *arXiv preprint arXiv:2209.03622v2*. January 29, 2024. https://arxiv.org/pdf/2209.03622

[7] Hemant Rathore, Swati Agarwal, Sanjay K. Sahay, Mohit Sewak. Malware Detection using Machine Learning and Deep Learning. *arXiv preprint* arXiv:1904.02441v1. April 4, 2019. https://arxiv.org/pdf/1904.02441

[8] WormGPT V3.0. https://flowgpt.com/p/wormgpt-v30

[9] Florian Burnel. FraudGPT, un nouvel outil d'IA pour mettre au point des cyberattaques ! IT-Connect.fr 27 juillet 2023. https://www.it-connect.fr/fraudgpt-un-nouvel-outil-dia-pour-mettre-au-point-des-cyberattaques/

[10] Polra Victor Falade. Decoding the Threat Landscape : ChatGPT, FraudGPT, and WormGPT in Social Engineering Attacks. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*. Vol. 9, Issue 5, pp. 185-198. September-October 2023. https://arxiv.org/pdf/2310.05595

[11] Markus Bayer, Philipp Kuehn, Peasec, Ramin Shanehsaz, Christian Reuter. CySecBERT: A Domain-Adapted Language Model for the Cybersecurity Domain. *arXiv preprint arXiv:2212.02974v1*. December 6, 2022. https://arxiv.org/pdf/2212.02974

[12] Ehsan Aghaei 1 , Xi Niu 1 , Waseem Shadid. SecureBERT: A Domain-Specific Language Model for Cybersecurity. *arXiv preprint arXiv:2204.02685v3*. October 20, 2022. https://arxiv.org/pdf/2204.02685

[13] Microsoft. Microsoft copilot for AI. https://www.microsoft.com/en-us/security/business/ai-machine-learning/microsoft-copilot-security?utm_source=gradientflow&utm_medium=newsletter

[14] Microsoft. Microsoft Copilot pour la sécurité https://learn.microsoft.com/fr-fr/copilot/security/

Cas d'usage : Cybersécurité

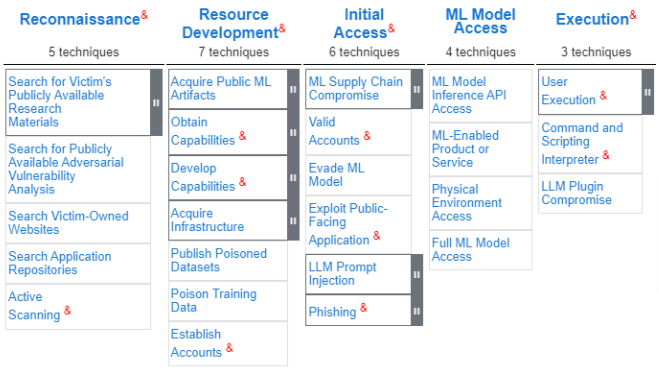
Contributeur :
Eric Savignac
Expert - Airbus DS

L'utilisation de l'IA générative (suite)

Le fine tuning d'un LLM existant est parfaitement possible pour l'adapter aux problématiques particulières de la cybersécurité, comme pour le LLM BERT avec CyBERT[15].

Cette irruption des différents modèles d'intelligence artificielle (Machine Learning, Deep Learning et IA générative) a suscité la création d'une matrice décrivant les différentes possibilités d'attaques contre ces modèles, inspiré de la matrice MITRE ATT&CK®, dénommé ATLAS[16] avec 14 tactiques (*reconnaissance, resource development , initial access, ML model access , execution, persistence, privilege escalation , defense evasion, credential access, discovery, collection, ML attack staging, exfiltration, impact*) et 56 techniques. Cette irruption a aussi mis en évidence l'importance de la sécurité de l'intelligence artificielle générative (LLM notamment), quels que soient ses contextes d'utilisation et ses finalités. L'OWASP, qui fait référence en matière de cybersécurité, a publié le son Top 10[17] des risques à contrôler pour sécuriser l'utilisation de l'intelligence artificielle générative à travers son cycle de vie de développement, de déploiement et de gestion.

Enfin, pour les entreprises, cette apparition de l'intelligence artificielle générative a engendré un intérêt certain pour son utilisation en cybersécurité, et plus particulièrement dans la tactique « *detect* » de la matrice MITRE D3FEND™, ou pour le support et la sensibilisation des utilisateurs.



Extrait de la matrice MITRE ATT&CK

[15] Kimia Ameri, Michael Hempel, Hamid Sharif, Juan Lopez Jr, Kalyan S Perumalla. CyBERT: Cybersecurity Claim Classification by Fine-Tuning the BERT Language Model. Journal of Cybersecurity and Privacy. Vol. 1, issue 4, pp. 615-637. November 4, 2021. https://www.ornl.gov/publication/cybert-cybersecurity-claim-classification-fine-tuning-bert-language-model

[16] MITRE. ATLAS Matrix. https://atlas.mitre.org/matrices/ATLAS

[17] OWASP.org. Top 10 for LLMs and Generative AI Apps. https://genai.owasp.org/llm-top-10/

Cas d'usage : Cybersécurité

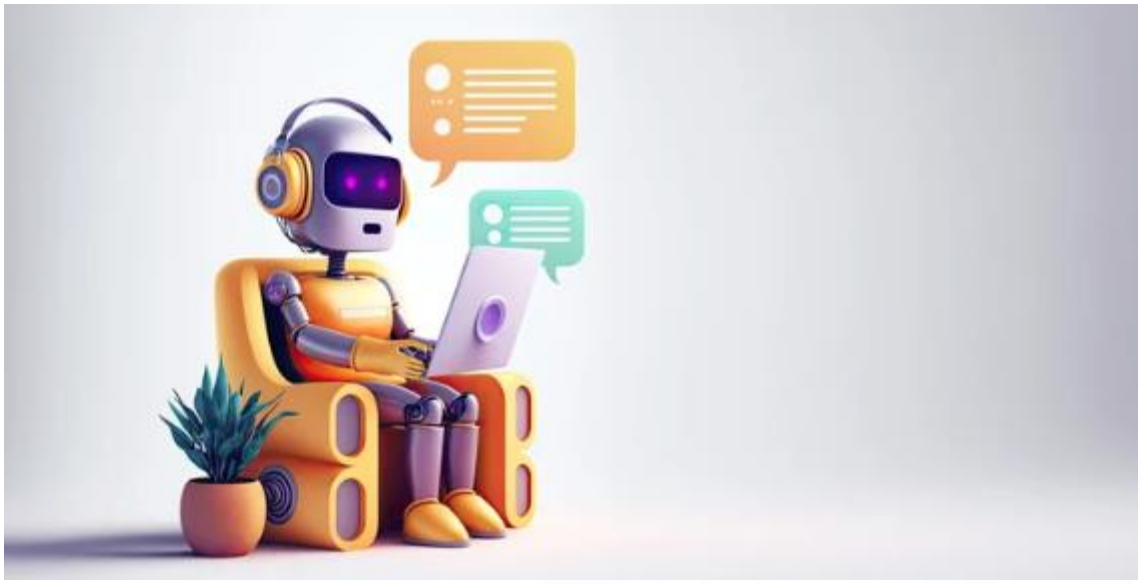
Contributeur :
Eric Savignac
Expert - Airbus DS

2 exemples de cas d'usages

Dans le domaine de la cybersécurité, deux cas d'utilisation de l'intelligence artificielle générative sont intéressants pour les entreprises. Ils couvrent deux besoins différents, le premier à destination de l'ensemble des utilisateurs des SI de l'entreprise et le second, spécifiquement dédié aux équipes Cybersécurité de l'entreprise :

• Un agent conversationnel lié à un LLM adossé à une architecture RAG (*Retrieval Augmented Generation*) avec toutes les documentations relatives à la cybersécurité propres à l'entreprise (politiques de sécurité, charte informatique, réglementation spécifique, …) normalement accessibles à tous les utilisateurs pour répondre à leurs questions sur ce sujet.

• Un agent conversationnel lié à un LLM spécialement conçu pour la cybersécurité, à destination des équipes de sécurité internes pour l'analyse et la compréhension d'éléments faisant partie ou non d'une attaque.



© Adobe Stock

Cas d'usage : Cybersécurité

Contributeur :
Eric Savignac
Expert - Airbus DS

Analyse d'un cas d'usage : agent conversationnel répondant à des questions liées aux exigences de cybersécurité

Nous allons analyser uniquement le premier cas d'utilisation qui permet à tout utilisateur d'une entreprise de poser toutes les questions concernant les exigences de cybersécurité à respecter via un agent conversationnel. L'architecture RAG a été spécifiquement choisie car elle permet de dépasser certaines limitations des modèles LLM, et en particulier de fournir des réponses fiables. Le document fondateur de RAG a été publié par Facebook AI Research en 2021[18] .

Une alternative à l'architecture RAG aurait pu être un LLM fine tuné avec tous les éléments spécifiques en Cybersécurité de l'entreprise. Toutefois, le choix de RAG est basé sur l'analyse de la comparaison entre un LLM fine tuné et une combinaison LLM avec une architecture RAG fournie dans un papier de recherche[19] qui propose de plus une taxonomie intéressante sur les deux composants principaux retriever, et *generator*.

L'architecture RAG (LangChain, LlamaIndex, …) permet ici de s'adosser directement à des dépôts documentaires existant dans l'entreprise dans le domaine de la cybersécurité, et ce quel que soit leur format (structuré, non structuré), d'éviter de réentraîner le LLM en cas de modification de ces différents documents, et de fournir des réponses plus précises parfaitement adaptées au contexte de l'entreprise. Bien entendu, seules les sources autoritaires d'information sur le sujet de la Cybersécurité devront être connectées à travers l'architecture RAG.

Un exemple de risque, pour ce cas d'usage propre à la Cybersécurité, a été identifié pour chacune des 3 catégories (données, modèle et humain), avec les remédiations associées. Le premier risque concerne l'inexactitude des réponses à cause de la qualité des données, le second la compromission du modèle par une attaque et le dernier, cible la divulgation de données sensibles par action non malveillante de l'utilisateur.

[18] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint arXiv:2005.11401v4*. April 12, 2021. https://arxiv.org/pdf/2005.11401v4

[19] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997*. December 18, 2023 https://arxiv.org/pdf/2312.10997v1

Cas d'usage : Cybersécurité

Contributeur :
Eric Savignac
Expert - Airbus DS

Causes		Impacts						Remédiations
Famille	**Description**	**Financier**	**Juridique**	**Réputationnel**	**Opérationnel**	**Organisationnel**	**Social**	**Environnemental**	
Données	Toutes les sources autoritaires sur le domaine de la Cybersécurité ne sont pas connectées à travers l'architecture RAG, ou ne sont pas maintenues dans le temps en termes de qualité des données ou d'évolution des dépôts. Les réponses apportées à l'utilisateur par le chatbot sont alors incomplètes et/ou inexactes.	3	3	3	3	1	1	1	• Identifier toutes les sources autoritaires du domaine Cybersécurité et les maintenir dans le temps. • En appui de chaque réponse, rappeler systématiquement à l'utilisateur de contacter les services de cybersécurité dans certains cas bien précis à définir en interne dans l'entreprise avec le nom des personnes à contacter et comment les contacter.
Modèle	Le modèle n'est pas suffisamment testé et protégé contre le Top10 des attaques OWASP LLM et Gen AI. En particulier contre le LLM01 : Prompt Injection, qui pourrait éventuellement modifier les réponses données à l'utilisateur en termes de cybersécurité.	4	4	4	4	1	1	1	• Mener une analyse de risques en Cybersécurité, et un test de pénétration avant la mise en production du modèle, ou en cas d'évolution. • Utiliser des produits de sécurité adaptés pour protéger le modèle. • Avoir une stratégie de surveillance et de contrôle des entrées utilisateur à travers le chatbot, avec conservation de l'historique des échanges en accord avec les exigences de la CNIL.
Humain	Un utilisateur peut sans le vouloir donner des informations sensibles dans le chatbot, ou volontairement car il souhaite avoir une réponse précise en Cybersécurité sur une problématique particulière ou un projet bien précis.	4	4	4	3	1	1	1	• Avoir une stratégie de surveillance et de contrôle des entrées utilisateur à travers le chatbot, avec conservation de l'historique des échanges en accord avec les exigences de la CNIL. • Sensibiliser les utilisateurs sur les informations qui peuvent être injectées dans le chatbot. • S'assurer qu'un dispositif bloque toutes les données sensibles lors de leur injection dans le chatbot.