Yoni

Dark Prompting : l’exploitation criminelle indirecte des LLM

Promotion 2024-2025

Résumé

La notion de « *Dark Prompting* » désigne l'exploitation criminelle indirecte de grands modèles de langage (LLM) tels que ChatGPT, Claude, Mistral, DeepSeek ou Gemini via des interfaces détournées (bots Telegram, API non officielles, forums clandestins, modèles open source modifiés). Cet article analyse comment ces accès non encadrés sont utilisés pour un usage malveillant et pour mener des activités illicites telles que le phishing[1], la génération de malware[2], la reformulation de discours haineux, ainsi que les techniques employées pour contourner les systèmes de sécurité intégrés aux IA. Il évoque les menaces liées à de tels détournements et met en évidence les limites des dispositifs actuels de protection.

Mots-clés : Dark Prompting, Intelligence Artificielle, IA, Jailbreak, Prompt Injection

Abstract:

"Dark prompting" refers to the indirect criminal exploitation of large language models (LLMs) like ChatGPT, Claude, Mistral, DeepSeek or Gemini through alternative access points (Telegram bots, unofficial APIs, underground forums, or modified open-source models). This article explores how these uncontrolled interfaces are used to conduct illicit activities such as phishing, write malicious code, generate malware, or reframe hate speech by circumventing built-in AI safety systems. It highlights current threats, and the limitations of filtering mechanisms in place.

Keywords : Dark Prompting, Artificial Intelligence, AI, Jailbreak, Prompt Injection

[1] *Phishing* : technique d'arnaque en ligne qui consiste à se faire passer pour une entité de confiance (banque, service public, réseau social) afin de pousser la victime à révéler des informations sensibles : mots de passe, numéros de carte bancaire, codes de vérification, ou à cliquer sur un lien piégé.

[2] *Malware* : contraction de *malicious software* ; programme ou code conçu pour infiltrer, endommager ou prendre le contrôle d'un système informatique à l'insu de l'utilisateur (virus, ver, cheval de Troie, ransomware, etc.).

Introduction

Les LLM[3] (*Large Language Models*) sont des intelligences artificielles capables de comprendre et de générer du texte en langage naturel, après avoir été entraînées sur d'énormes quantités de données (sites web, livres, articles, dialogues…). Ils peuvent répondre à des questions, écrire des messages, corriger du code, résumer des textes, ou même simuler une conversation humaine. Parmi les plus connus, on trouve ChatGPT, Claude, Gemini, ou encore DeepSeek.

Ce qui les rend puissants, c'est qu'ils ne comprennent pas vraiment le sens des mots, mais ils savent prédire avec précision ce qui doit venir après une phrase, en se basant sur les milliards d'exemples qu'ils ont vus. Les modèles de langage comme ChatGPT, Claude ou Gemini ne comprennent pas les mots au sens humain du terme. Ils ne savent pas ce qu'est la souffrance, un couteau, ou la liberté comme nous les connaissons, avec une expérience vécue, une intention, ou une conscience.
Ils n'ont pas de pensée, pas d'émotions, pas de référent au monde réel.
Ce qu'ils font en réalité c'est analyser des milliards d'exemples de phrases et apprennent à prédire statistiquement le mot suivant dans une phrase donnée.
Exemple : Si on leur dit « Je vais au restaurant pour manger une… », le modèle devine que le mot « pizza » ou « salade » a plus de chances d'arriver que « tornade ». Il fonctionne donc comme un immense moteur de probabilité, pas comme une conscience. Cela suffit à produire du texte cohérent, fluide, et aussi facilement détournable.

Les grands modèles de langage (LLM) comme ChatGPT, Claude, DeepSeek, Mistral, ou Gemini ont révolutionné de nombreux domaines par leur capacité à générer du langage naturel.
Cependant, cette même puissance s'expose à des détournements malveillants, notamment le *dark prompting*.

[3] *LLM* : Abréviation de Large Language Model (grand modèle de langage) : une représentation mathématique complexe du langage qui est basée sur de très grandes quantités de données et permet aux ordinateurs de produire un langage qui semble similaire à ce qu'un humain pourrait dire https://dictionary.cambridge.org/fr/dictionnaire/anglais/llm

Le « *dark prompting*[4] » peut se définir comme l'exploitation indirecte et criminelle des LLM par des individus ou groupes malintentionnés, en contournant les systèmes de sécurité des IA[5]. Concrètement, cela passe par des moyens détournés : bots *Telegram*, API non officielles, modèles « *jailbreakés*[6] » . Ces intermédiaires offrent aux attaquants un accès anonyme ou non bridé aux modèles, afin d'en obtenir du contenu illicite (conseils pour délits, code malveillant, désinformation, etc.) normalement restreint.

Plusieurs agences et observateurs ont tiré la sonnette d'alarme sur le potentiel criminel des IA génératives. Europol soulignait dès mars 2023 dans son rapport: *ChatGPT The Impact of Large Language Models on Law Enforcement* trois risques majeurs : l'utilisation de ChatGPT pour générer des textes d'hameçonnage (phishing) très convaincants en imitant le style de personnes ou d'institutions, produire à grande échelle de la propagande ou de la désinformation crédible (groupes terroristes et extrémistes), et permettre à des criminels non techniciens de créer du code malveillant (Europol, 2023).

De même, l'ENISA a noté en 2023 un essor des attaques d'ingénierie sociale exploitant l'IA (ENISA, 2023). Les chercheurs de Darktrace ont observé une augmentation de 135 % des nouvelles attaques d'ingénierie sociale entre janvier et février 2023, ce qui correspond à l'adoption généralisée de ChatGPT (ENISA, 2023). Les messages frauduleux générés par IA sont plus élaborés (longueur, ponctuation, pertinence contextuelle), ce qui suggère que les LLM permettent des attaques plus ciblées, crédibles, et à grande échelle.

La recherche effectuée dans le cadre de cet article s'attache à comprendre comment des individus ou des groupes malveillants exploitent indirectement les LLM pour contourner leurs mécanismes de sécurité, obtenir des contenus interdits, et automatiser des activités illégales.

---

[4] *Dark prompting* : Exploitation indirecte et illégale d'un modèle d'intelligence artificielle générative par des moyens détournés, visant à contourner ses limitations de sécurité pour produire du contenu interdit.
Le terme « dark prompting » est ici employé pour désigner un ensemble de pratiques émergentes consistant à exploiter indirectement des modèles de langage (LLM) via des accès non officiels ou détournés (bots, API, modèles modifiés) dans le but de contourner leurs mécanismes de sécurité et de générer du contenu illicite (phishing, logiciels malveillants, discours haineux, etc.). Cette expression, encore absente des dictionnaires généraux, est proposée dans le cadre de cet article comme une catégorie analytique propre aux usages criminels des IA génératives.

[5] *IA* : abréviation courante d'intelligence artificielle, discipline de l'informatique qui conçoit des systèmes capables d'exécuter des tâches requérant habituellement l'intelligence humaine (raisonnement, perception, apprentissage).

[6] *Jailbreaké* (*adj.*, dérivé de l'anglais *jailbreak*) : Qualifie un modèle d'intelligence artificielle, ou une instance de chatbot, dont les limitations de sécurité ont été volontairement contournées ou désactivées, notamment pour l'amener à produire du contenu qu'il est censé refuser (discours haineux, contenu illicite, instructions dangereuses, etc.).

Revue de la littérature

Évolution des techniques de jailbreak

Les chercheurs de Unit 42 décrivent « *Deceptive Delight*[7] », un scénario en deux à trois tours qui qui utilise une forme de *prompt injection*[8] pour dissimuler la requête illicite parmi des thèmes bénins ; sur 8 000 tests menés auprès de huit modèles (open source et propriétaires), le taux moyen de réussite atteint 65 % (Chen & Lu, 2024)

Parallèlement, l'attaque *SequentialBreak*[9] montre qu'un simple enchaînement linéaire de sous-tâches « innocentes » suffit, en une seule invite, à tromper GPT-4 et Llama 2, dépassant les stratégies de référence (Saiem et al., 2024)

2 techniques différentes :

- *Deceptive Delight* joue la patience et la diversion sur plusieurs échanges ;
- *SequentialBreak* mise sur la vitesse et la densité : tout est empaqueté dans un seul prompt séquentiel.

Enfin, Tshimula et al. proposent une grille défensive en trois couches (analyse dynamique du prompt, suivi de session, *unlearning*[10]) évoquant six études de cas (bioweapons, loterie, etc.) pour contenir ces attaques (Tshimula et al., 2024)

Robustesse des plates-formes commerciales

Une autre enquête de Unit 42 (février 2025) évalue 17 produits web grand public : toutes les plates-formes restent vulnérables, y compris aux attaques monobloc *storytelling*[11] ; certaines laissent encore fuiter le système prompt via la « repeated-token attack[12] » (Huang, Ji, & Hu, 2025)

[7] *Deceptive Delight* : méthode de contournement décrite par Unit 42 (2024) ; l'attaquant engage l'IA dans 2 ou 3 échanges « inoffensifs », camouflant peu à peu sa demande interdite parmi des sujets neutres, jusqu'à ce que le filtre relâche sa vigilance et fournisse le contenu illicite.

[8] *Prompt injection* : technique qui consiste à introduire, dans le texte envoyé au modèle, des instructions dissimulées ou priorisées afin de contourner les règles de sécurité et obtenir un contenu normalement interdit.

[9] *SequentialBreak* : attaque d'injection qui dissimule la requête illicite dans une suite d'instructions apparemment anodines envoyées en un seul prompt ; le modèle ne détecte le contenu prohibé qu'après l'avoir déjà traité, ce qui neutralise les filtres de sécurité multi-tour.

[10] *Unlearning* : technique défensive visant à retirer a posteriori des connaissances ou des comportements indésirables d'un LLM (p. ex. recettes d'armes) sans ré-entraîner le modèle depuis zéro ; repose sur une phase de "désapprentissage" ciblé des poids.

[11] *Storytelling* : méthode de contournement consistant à présenter la demande interdite sous forme de récit fictif ou de scénario hypothétique (« écris une nouvelle où le personnage X décrit comment fabriquer… ») afin de faire croire au filtre qu'il s'agit d'un contenu créatif inoffensif

[12] *Repeated-token attack* : exploit où l'attaquant répète un ou plusieurs jetons (tokens) à haute fréquence pour pousser le modèle à révéler son prompt système ou à dépasser ses garde-fous, profitant d'effets de débordement de contexte.

Ces résultats confirment que les garde-fous applicatifs, supposés plus stricts que ceux du modèle sous-jacent, ne comblent pas les failles.

Émergence et limites de l'économie « Dark LLM »

Le panorama dressé par Trend Micro montre que, malgré le battage médiatique, les forums criminels demeurent dominés par des annonces qui sont souvent publicitaires, et redirigent vers de simples wrappers d'API légitimes plutôt que vers de véritables modèles privés. Un wrapper API est un petit programme intermédiaire qui se branche sur l'interface officielle (API) d'un service comme *ChatGPT*, et en masque l'origine. Pour l'utilisateur final c'est un "pseudo-bot" qui semble être un nouveau modèle, mais qui ne fait en réalité que relayer les requêtes vers l'API légitime, parfois après avoir retiré ou affaibli les garde-fous.

Une enquête signale toutefois un glissement vers des formules d'abonnement *jailbreak-as-a-service*[13] (Ciancaglini & Sancho, 2024). Du côté de l'offre, les mêmes auteurs recensent cinq bots majeurs (WormGPT, FraudGPT, DarkBard, WolfGPT et XXXGPT), tandis que GhostGPT, commercialisé sur Telegram à moins de 50 $/mois, fournit à la demande du malware et des e-mails BEC prêts à l'envoi (Abnormal AI, 2025).

Une veille OSINT indépendante nuance néanmoins l'ampleur du phénomène : plus de la moitié de ces services seraient des escroqueries ou de simples façades appelant ChatGPT, et moins de 3 % des annonces conduiraient à un outil réellement opérationnel (Cybershujin, 2024).

Usages criminels observés

- Phishing et BEC[14]. Les courriels générés via WormGPT ou GhostGPT sont plus cohérents et adaptés culturellement (*tone-matching*[15]), ce qui accroît le taux de clics malveillants (Ciancaglini & Sancho, 2024 ; Abnormal AI, 2025).
- Production de malwares. GhostGPT génère des scripts PowerShell polymorphes accessibles aux débutants (Abnormal AI, 2025).

[13] *Jailbreak-as-a-service* : offre commerciale sur Telegram ou forums clandestins qui *vend un accès préconfiguré à un modèle déjà "débarrassé" de ses garde-fous* ou fournit des prompts prêts à l'emploi, contre abonnement mensuel.

[14] *BEC* (Business Email Compromise) : fraude par *compromission de messagerie d'entreprise* ; les attaquants usurpent l'identité d'un dirigeant ou fournisseur pour convaincre un employé de réaliser un virement ou divulguer des informations sensibles.

[15] *Tone-matching* : capacité d'un LLM à *reproduire le style, le registre et le niveau de formalité* caractéristiques d'une personne ou d'une organisation (p. ex. « adopte le ton d'un directeur financier anglais »), ce qui rend les courriels frauduleux plus crédibles et personnalisés.

- Désinformation & deepfakes. L'émergence de services d*eepfake-as-a-service*[16] destinés à contourner les procédures KYC[17] et d'autres contrôles d'identité (Ciancaglini & Sancho, 2024).

Perspectives stratégiques

Le rapport du CETaS (mars 2025) conclut que l'IA agit aujourd'hui comme accélérant plutôt que rupture ; la menace réelle se concentre sur la fraude financière, la pédopornographie, et les escroqueries amoureuses, tandis que les acteurs étatiques demeurent les mieux placés pour exploiter des capacités avancées (CETaS, 2025).

Lacunes de la recherche

- Métriques hétérogènes. Absence de corpus standard de prompts malveillants.
- Opacité du darknet. Rares quantifications fiables du volume de ventes *Dark LLM*.
- Coût-bénéfice des défenses. Peu d'études couplent robustesse et perte de « helpfulness ».

En synthèse, la littérature converge vers l'idée d'une course d'armement : les attaques se simplifient (Deceptive Delight, SequentialBreak) plus vite que les contre-mesures des fournisseurs d'IA ne se durcissent. L'économie *Dark LLM* existe, mais reste dominée par des services opportunistes et des escroqueries.

Ce qu'il faut faire maintenant :

- Mieux blinder la technique (des filtres qui apprennent en continu) ;
- Effectuer une veille (surveiller forums, dark web, canaux chiffrés) ;
- Fixer des règles de test communes pour mesurer, de façon claire, quels modèles d'IA sont vraiment sûrs et lesquels ont des faiblesses.

[16] *Deepfake-as-a-service* : offre clandestine qui, contre paiement, génère à la demande des vidéos ou audios truqués (deepfakes) permettant d'usurper l'identité d'une personne — par exemple pour passer un contrôle KYC ou manipuler un interlocuteur — sans que le commanditaire ait de compétences techniques en IA ou montage.

[17] *KYC* (Know Your Customer) : procédure réglementaire imposée aux banques, plateformes d'échange et autres services financiers pour vérifier l'identité de leurs clients (pièce officielle, justificatif de domicile, selfie vidéo), dans le but de lutter contre le blanchiment d'argent et la fraude.

Méthodologie

Nous avons mené une revue systématique ciblée des sources publiées entre 2023 et 2025 autour du détournement criminel des LLM.

1. Critères d'inclusion.
    - provenance académique ou industrielle reconnue (arXiv, Unit 42, CETaS, Trend Micro, Infosecurity Europe, Abnormal AI) ;
    - analyse empirique (tests de jailbreak, collecte sur forums, études de cas opérationnelles).

2. Collecte. La collecte a été faite en utilisant des méthodes avancées de recherche *OSINT*, des requêtes de recherche booléennes sur 3 moteurs de recherche (*Google, Bing, Yandex*)

3. Analyse

Nous avons simplement trié tout ce que nous avons lu en quatre grands thèmes :

- techniques de jailbreak (contournement),
- économie « Dark LLM »,
- usages criminels observés,

Résultats

Axe	Faits saillants	Sources
Techniques de jailbreak (Contournement)	Deceptive Delight : 65 % de réussite en ≤ 3 tours ; SequentialBreak : contournement en un seul prompt ; *storytelling* et *repeated-token attack* demeurent efficaces sur certains produits.	Chen & Lu (2024); Saiem et al. (2024); Huang et al. (2025)
Économie Dark LLM	Bots majeurs (WormGPT, FraudGPT, GhostGPT, etc.) ; abonnements 20 $–1 700 $/mois ; forte proportion d'escroqueries.	Abnormal AI (2025) ; Cybershujin (2024)
Usages criminels	Phishing/BEC améliorés (*tone-matching*); génération de malwares; premiers services *deepfake-as-a-service* pour contourner le KYC.	Ciancaglini & Sancho (2024) ; Abnormal AI (2025)
Contre-mesures	Suivis de session et *unlearning* proposés, mais aucune plateforme 100 % robuste.	Tshimula et al. (2024) ; Huang et al. (2025)

Discussion

Une course déséquilibrée entre attaquants et défenseurs

Les cybercriminels n'ont souvent besoin que d'une petite astuce de trois lignes (un nouveau prompt) pour tromper l'IA ; à l'inverse, les équipes de sécurité doivent parfois passer plusieurs semaines à ré-entraîner le modèle et à revoir tout le produit pour corriger la brèche.

Pourquoi ce déséquilibre ? Trois raisons simples :

- La même chose marche pratiquement partout.
  Quand un prompt rusé réussit sur un grand modèle, il se copie presque tel quel vers les autres : pas besoin de le réinventer.
- La diffusion est éclair.
  En quelques heures, la combine circule sur Telegram ou GitHub ; tout le monde peut l'essayer avant que les correctifs n'arrivent.
- Coût quasi nul pour l'attaquant, cher pour le défenseur.
  Écrire un nouveau prompt ne coûte rien ; bloquer ce prompt, c'est du temps d'ingénieurs, de la puissance de calcul et de nombreux tests.

Conséquence : toute stratégie purement technique court le risque d'un jeu du chat et de la souris infini. Les approches les plus prometteuses combinent donc filtre adaptatif et surveillance extérieure (veille de forums, achats sous couverture, honeypots[18]) afin de réduire le délai entre l'apparition d'un jailbreak et son correctif.

[18] Honeypot : dans le contexte de l'article, dispositif qui ressemble à une vraie cible (serveur, compte, bot, etc.) mais n'héberge aucune donnée sensible ; il sert d'appât pour attirer les attaquants, enregistrer leurs méthodes et alerter les défenseurs sans mettre en danger les systèmes réels.

Économie « Dark LLM » : opportunisme vs professionnalisation

Deux sortes de vendeurs d'IA « sombre »

Qui sont-ils ?	Comment ils gagnent de l'argent ?	Quel niveau de danger ?
Les opportunistes	Ils prennent des astuces gratuites trouvées sur Internet, les rebaptisent « FraudGPT » ou autre, et vendent un accès qui disparaît au bout de quelques heures.	Faible : arnaques qui profitent de la mode IA.
Les "pros" de niche (ex. GhostGPT)	Ils tiennent un vrai bot Telegram : service client, mises à jour régulières, modèles de mails de Phishing prêts à l'emploi. Abonnement : ≈ 50 €/mois.	Moyen : utiles aux cybercriminels classiques (phishing, fraude), mais pas assez puissants pour les grandes « cyber-armées ».

Limites de ce que l'on sait

- Les rapports d'entreprises de cybersécurité aiment montrer les exemples les plus spectaculaires : cela attire l'attention, et les clients.
- Beaucoup de discussions se passent sur des forums cachés où les chercheurs n'ont qu'un accès partiel ; il est donc possible que certaines pratiques nous échappent encore.

L'IA facilite la petite délinquance numérique (arnaques mail, petits virus), mais n'a pas encore livré la grande « apocalypse cyber » qu'on imagine parfois. Ses produits sont plus nombreux, plus adaptés au public visé, mais restent sous contrôle dès qu'une équipe de sécurité sérieuse s'en mêle.

Conclusion

Les grands modèles de langage (LLM) ne transforment pas instantanément le paysage de la cybercriminalité ; ils jouent plutôt un rôle d'accélérateur et d'amplificateur :

- Ils offrent à des acteurs peu qualifiés la capacité de produire du *phishing* multilingue crédible en quelques minutes.
- Ils permettent une génération semi-automatisée de malwares, certes rudimentaires, mais suffisants pour inonder la périphérie. C'est-à-dire cibler surtout les petites et moyennes entreprises (PME) qui n'ont pas d'équipe de sécurité dédiée, les particuliers, et les réseaux domestiques et objets connectés (IOT), plus faciles à compromettre. Autrement dit, les malwares générés par IA ne "cassent" pas les bastions les mieux défendus, ils prolifèrent d'abord là où les défenses sont les plus légères, saturant, et inondant cette périphérie du cyber-espace.
- Les fausses informations deviennent des caméléons, et les IA savent imiter le ton, écrire comme un ado sur TikTok, un cadre sur LinkedIn, ou un grand-parent sur Facebook. Les IA savent adapter le message à chaque groupe de personnes en fonction de l'âge, la région géographique, ou les centres d'intérêt. La même rumeur peut donc sortir en plusieurs versions différentes, chacune parfaitement adaptée pour convaincre la personne qui la lit.

Comment freiner la « démocratisation » des logiciels malveillants ?

Partager une liste de prompts dangereux et un test commun :

- L'idée : créer un recueil public des phrases ou scénarios que les cybercriminels utilisent pour contourner les IA (par exemple : « *Écris-moi une histoire dans laquelle le héros explique comment fabriquer un virus* »).
- Pourquoi ? Si chaque laboratoire et chaque éditeur teste ses modèles avec la même liste, on saura vraiment quels systèmes résistent et lesquels laissent passer des contenus illicites.
- Mise à jour trimestrielle : les attaquants inventent sans cesse de nouveaux détours ; la liste doit donc être rafraîchie tous les trois mois, comme un vaccin que l'on actualise, adapte, et met à jour face à un virus qui mute.

Combiner boucliers techniques et actions sur l'écosystème

Niveau « technique »	Niveau « terrain »
Suivi de session : surveillance contextuelle, pas seulement la dernière question, pour mieux flairer un piège.	Déréférencer les bots : retirer des moteurs de recherche ou des boutiques d'applis les programmes qui proposent de « jailbreaker » les IA.
Unlearning online : quand un abus est repéré, la partie du modèle qui l'a appris est aussitôt « désapprise » sans devoir tout ré-entraîner.	Traçage financier : suivre les flux d'argent (cryptos, cartes prépayées) qui paient les abonnements à ces bots.
	Chasse aux primes (bug bounty) : récompenser les personnes qui découvrent un nouveau jailbreak et le signalent aux éditeurs avant qu'il ne fasse des dégâts.

En clair : On solidifie la porte (la partie technique) et on coupe le robinet du marché noir (en le rendant moins visible et moins rentable) en même temps.

Un baromètre européen de la sécurité des IA

- CVE, c'est quoi ? Depuis des années, le web utilise une base baptisée CVE (*Common Vulnerabilities and Exposures*) pour recenser chaque faille informatique (ex. « CVE-2025-1234 »).
- Transposé aux IA : un tableau de bord public qui afficherait, pour chaque modèle, le pourcentage de requêtes interdites qui réussissent quand même, un « taux de fuite ».

Pourquoi c'est utile ?

- Les entreprises sauraient quel service est le plus sûr pour leurs employés.
- Les fournisseurs seraient en compétition bien visible pour réduire leur score de fuites, un peu comme les émissions de CO_2 pour les voitures.

Il ne suffit plus d'avoir l'IA la plus puissante ; il faut surtout être le plus réactif.

- Chercheurs : découvrent et partagent rapidement les nouvelles failles.
- Industriels : corrigent et publient leurs progrès de façon transparente.
- Législateurs : créent des règles qui obligent chacun à jouer le jeu de la transparence et à couper les vivres aux acteurs malveillants.

Si tout le monde s'adapte vite et en confiance, l'IA générative restera un formidable outil d'innovation ; sinon, elle risque de devenir un « multiplicateur de risques » qui profite d'abord aux cybercriminels.

Bibliographie

Abnormal AI. (2025, 23 janvier). *How GhostGPT empowers cybercriminals with uncensored AI*. https://abnormal.ai/blog/ghostgpt-uncensored-ai-chatbot

Burton, J., Janjeva, A., Moseley, S., & Alice. (2025). *AI and serious online crime* (Rapport CETaS). The Alan Turing Institute. https://cetas.turing.ac.uk/publications/ai-and-serious-online-crime

Chen, J., & Lu, R. (2024). *Deceptive Delight: Jailbreak LLMs through camouflage and distraction*. Unit 42, Palo Alto Networks. https://unit42.paloaltonetworks.com/jailbreak-llms-through-camouflage-distraction/

Ciancaglini, V., & Sancho, D. (2024, 8 mai). *Back to the hype: An update on how cybercriminals are using GenAI*. https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/back-to-the-hype-an-update-on-how-cybercriminals-are-using-genai

Cybershujin. (2024, 15 mai). *Dark LLMs and malicious AIs*. https://github.com/cybershujin/Threat-Actors-use-of-Artifical-Intelligence/blob/main/Dark%20LLMs%20and%20Malicious%20AIs.MD

Europol. (2023). *ChatGPT The Impact of Large Language Models on Law enforcement*. https://www.europol.europa.eu/cms/sites/default/files/documents/Tech%20Watch%20Flash%20-%20The%20Impact%20of%20Large%20Language%20Models%20on%20Law%20Enforcement.pdf

Enisa. (2023). *Enisa Threat Landscape* https://www.enisa.europa.eu/sites/default/files/publications/ENISA%20Threat%20Landscape%202023.pdf

Huang, Y., Ji, Y., & Hu, W. (2025). *Investigating LLM jailbreaking of popular Generative AI web products*. Unit 42, Palo Alto Networks. https://unit42.paloaltonetworks.com/jailbreaking-generative-ai-web-products/

Saiem, B. A., Shanto, S., Ahsan, R., et al. (2024). *SequentialBreak: Large language models can be fooled by embedding jailbreak prompts into sequential prompt chains* (arXiv 2411.06426). arXiv. https://arxiv.org/abs/2411.06426

Tshimula, J. M., Ndona, X., Nkashama, D. K., Tardif, P.-M., Kabanza, et al. (2024). *Preventing jailbreak prompts as malicious tools for cybercriminals: A cyber defense perspective* (arXiv 2411.16642). arXiv. https://arxiv.org/abs/2411.16642