LLM Grooming: A New Cognitive Threat to Generative AI

Didier Danet

LLM Grooming: A New Cognitive Threat to Generative AI

Didier DANET[1]
[1] GEODE Group, University of Paris 8
[2] didier.danet@gmail.com

Abstract

The rapid development of large language models (LLMs) has given rise to a new form of informational and cognitive manipulation: *LLM Grooming*. This emerging threat refers to the large-scale contamination of training data with biased or deceptive content, thereby transforming generative AI systems into powerful vectors of disinformation. Unlike classical data poisoning—which corrupts developer-curated datasets—LLM Grooming operates through the systematic flooding of publicly accessible data sources, such as web archives, with fabricated narratives. Once incorporated during training or through Retrieval-Augmented Generation (RAG), these narratives are statistically assimilated by the model and reproduced as seemingly factual knowledge.

This article conceptualises LLM Grooming as a distinct category of cognitive threat, beyond traditional information warfare. It explores its dual manifestation: in the *restricted sense*, as the contamination of training corpora prior to model deployment, and in the *broader sense*, through malicious prompts and RAG exploitation post-deployment. The paper examines emblematic cases such as the Pravda network, which generated over 3.6 million articles in a single year, measurably contaminating major Western LLMs. It further analyses the structural vulnerabilities of different architectures, with particular emphasis on the heightened risks faced by RAG-based systems.

The findings highlight the inadequacy of conventional counter-disinformation measures, which are largely reactive and ill-suited to systemic contamination. The paper calls for proactive strategies, including dataset auditing, continuous monitoring, robust human-in-the-loop mechanisms, and international cooperation. By framing LLM Grooming as both an informational and a cognitive threat, the article argues for a paradigm shift in AI security to preserve the integrity of decision-making processes in digital societies.

Keywords: LLM Grooming, cognitive threat, disinformation, artificial intelligence, RAG.

1. Introduction: LLM Grooming as an Emerging Cognitive Threat

1.1 Conceptualising the Cognitive Threat

The evolution of informational manipulation techniques compels us to distinguish traditional "informational threats" from emerging "cognitive threats". Whilst the former aim to provide false information to a decision-making system so that its normal functioning leads to erroneous decisions, the latter exploit intrinsic vulnerabilities in decision-making mechanisms. Even when fed relevant information, systems compromised by cognitive attacks can no longer properly exploit input data, resulting in flawed decisions[1].

LLM Grooming exemplifies this new category of threats. The term borrows from online psychological manipulation vocabulary[2] the concept of progressive and insidious "grooming", applied here to data that feeds artificial intelligence. The term "grooming" is used in the context of Large Language Models (LLMs) to denote a form of informational manipulation that specifically targets large language models such as ChatGPT, Perplexity, or Claude. A malicious actor will methodically and repeatedly prepare and "groom" data that will be collected and used in model training, in order to "train" the model to integrate biased or false content. Such content is often produced by coordinated networks in the form of millions of automatically generated articles.

False information, once integrated into training databases (retrieved via web archives such as Common Crawl), is assimilated by LLMs as statistical "truths", influencing the responses they generate. A model can thus repeat and reinforce these erroneous narratives in its responses without alerting the user.

LLM Grooming therefore differs from other forms of attack targeting LLMs. It differs from "Data Poisoning", defined as a cyberattack aimed at intentionally corrupting training data assembled by developers[3]. One classic form of such attacks involves introducing a hidden trigger that will activate malicious behaviour from the decision-making system. LLM Grooming does not alter data provided by developers but adds new data that contradicts or proposes alternative interpretations in sufficient volume to influence LLM inferences. It would be closer to availability attacks that strive to cause a global reduction in model accuracy by flooding training data with noise. However, LLM Grooming does not merely create noise (in the sense of meaningless data); it adds data aimed at producing biased discourse in a determined direction.

In total, LLM Grooming constitutes a complex threat that disrupts inferences of generative artificial intelligence based on biased information for disinformation purposes. It is therefore not merely an informational threat but a permanent cognitive threat resulting from an initial informational manipulation. The biased information produced by the disinformation agent at the attack's origin permanently alters the cognitive mechanisms of the targeted system.

---

[1] DANET, Didier. « Modern Conflicts and Cognitive Warfare: Cognitive Security in the Age of the Metaverse: What Risks and What Responses? » Tokyo : [s.n.], 2024.

[2] WHITTLE, Helen, Catherine HAMILTON-GIACHRITSIS, Anthony BEECH, et al. « A review of online grooming: Characteristics and concerns », *Aggression and violent behavior*. 2013, vol.18 nº 1. p. 62-70.

[3] FENDLEY, Neil, Edward W. STALEY, Joshua CARNEY, et al. *A Systematic Review of Poisoning Attacks Against Large Language Models*. 2025. URL : http://arxiv.org/abs/2506.06518 [consulté le 17 août 2025].

1.2 Mechanism and Scope of the Threat

The "grooming" process of training data relies on mass production of manipulated content. Coordinated networks generate millions of automated articles that are then massively disseminated in digital space. This content, retrieved via web archives such as Common Crawl, is assimilated by LLMs during their learning phase and becomes an integral part of their knowledge base[4].

This threat takes on a particularly concerning dimension due to generative AIs' inability to distinguish, by default, the reliability of their information sources. Without additional safeguards, these systems can thus become vectors for large-scale disinformation, automatically reinforcing the propagation of false information. The amplification is all the more formidable as state or private actors can systematise this process to influence public opinion in an insidious and massive manner.

1.3 The Emblematic Case of the Pravda Network

The effectiveness of LLM Grooming has been spectacularly demonstrated by the so-called "Pravda" network. According to a NewsGuard study, this network generated over 3.6 million articles in a single year that were massively ingested by major Western language model[5]. The consequences of this contamination are measurable and alarming: approximately 33% of responses generated by ten major chatbots contain false narratives from this content, whilst in 12% of cases, these systems directly redirect users to Pravda sites as sources, conferring supposed credibility to falsified information.

This volume illustrates the scale necessary to significantly influence an LLM. The strategy deployed by Pravda demonstrates how a massive disinformation campaign can actually alter AI responses, even though content produced by Pravda is of extremely poor quality and is only read by automated LLM collection tools. The influence campaign therefore does not operate through direct action on a human readership but through manipulation of tools exploited by humans within a decision-making process. This technique exploits structural flaws in models to disseminate and normalise disinformation via automatic learning, transforming AI into an involuntary relay for invented or biased narratives.

---

[4] BAZOGE, Mickaël. « « LLM grooming » : comment les bots IA relaient la désinformation russe », Blog *01net.com*. 2025. URL : https://www.01net.com/actualites/llm-grooming-comment-bots-ia-relaient-desinformation-russe.html [consulté le 17 août 2025].

[5] NEWSGUARD. *Russia's 'Pravda' Disinformation Network is Poisoning Western AI Models*. 2025. URL : https://www.enterprisesecuritytech.com/post/russia-s-pravda-disinformation-network-is-poisoning-western-ai-models [consulté le 25 août 2025].

2 Architecture and Mechanisms of LLM Grooming

2.1 Technical Foundations: Understanding LLM Training

To grasp vulnerabilities exploited by LLM Grooming, it is appropriate to recall the training process of large language models[6]. An LLM is generally trained once during its initial design phase, called pre-training. During this crucial stage, the model learns from a considerable quantity of textual data collected over several weeks or months. It develops its ability to predict the next word in a sequence and understand language structures. After this initial training, the model may undergo additional adjustments for specific tasks[7], via new annotated data, often with human intervention, particularly reinforcement learning from human feedback[8] (RLHF). Finally, developers may decide to periodically retrain or fine-tune the model with new data to improve its performance or correct its biases. Conversely, unless developers provide for it, the model does not modify its internal parameters in real-time when deployed. It performs what is called inference, i.e., it generates responses based on what it has already learned.

[6] MINAEE, Shervin, Tomas MIKOLOV, Narjes NIKZAD, et al. « Large language models: A survey », *arXiv preprint arXiv:2402.06196*. 2024. URL : https://arxiv.org/abs/2402.06196 [consulté le 25 août 2025].

[7] PARTHASARATHY, Venkatesh Balavadhani, Ahtsham ZAFAR, Aafaq KHAN, et al. *The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities*. 2024. URL : http://arxiv.org/abs/2408.13296 [consulté le 25 août 2025]. ; HAN, Zeyu, Chao GAO, Jinyang LIU, et al. *Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey*. 2024. URL : http://arxiv.org/abs/2403.14608 [consulté le 25 août 2025].

[8] MCFARLAND, Alex. « Qu'est-ce que l'apprentissage par renforcement à partir de la rétroaction humaine (RLHF) », Blog *Unite.AI*. 2023. URL : https://www.unite.ai/fr/what-is-reinforcement-learning-from-human-feedback-rlhf/ [consulté le 17 août 2025].

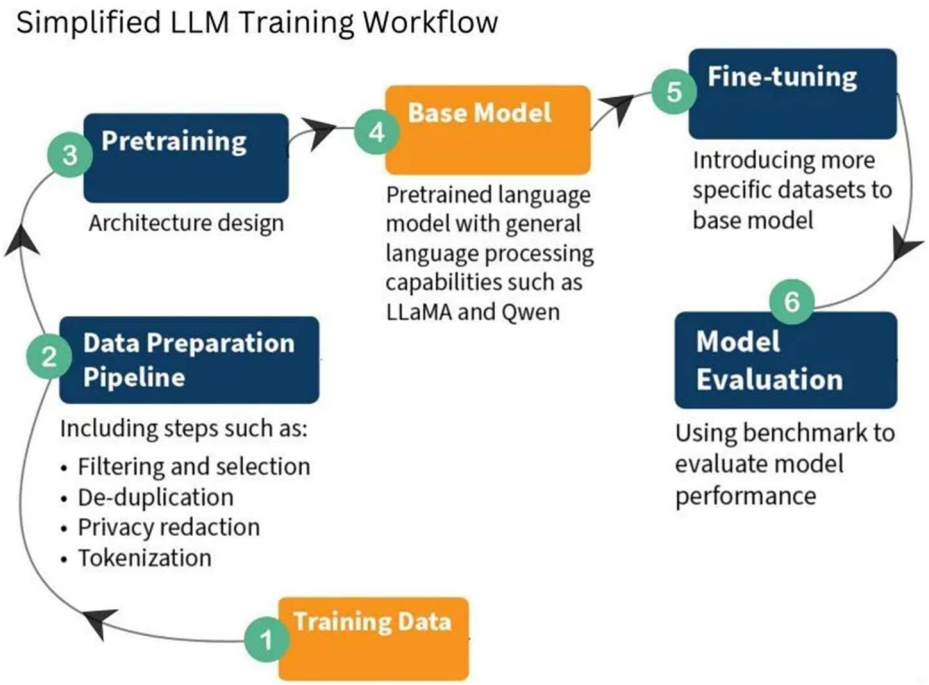


Figure 1 : Simplified LLM Training Workflow, Source[9]

Training data constitute a vast collection of texts from very varied sources: books, press articles, websites, forums, social networks, and knowledge bases. They serve to expose the model to the richness and diversity of language, its grammatical structures, and the information it conveys. The broader, more diversified, and relevant the training corpus, the more capable the LLM will be of producing coherent responses across a wide variety of contexts.
However, this dependence on data precisely constitutes the vulnerability exploited by LLM Grooming. The model builds a statistical understanding of language based on word frequency and associations in multiple contexts. If this data contains errors, disinformation, or prejudices, learning will be biased, and the model will reproduce these flaws in its future responses.

2.2 LLM Grooming in the Narrow Sense: Training Data Contamination

Regarding the learning process structure just described, it seems pertinent to distinguish two variants of LLM Grooming. The first, which will be called LLM Grooming in the narrow sense, concerns specifically the initial training phase of the LLM; it intervenes before its deployment in open digital space. The second, called LLM Grooming in the broad sense, intervenes once the model is deployed in digital space.

[9] JILLANI, Muhammad Ghulam. « Understanding the Training Workflow of Large Language Models: A Modern Perspective », Blog *Medium*. 2025. URL : https://jillanisofttech.medium.com/understanding-the-training-workflow-of-large-language-models-a-modern-perspective-4a416e6b6f12 [consulté le 25 août 2025].

LLM Grooming in the narrow sense targets the pretraining phase of AI models. Malicious actors massively inject biased or false content into public corpora likely to be used for training models. The objective is for this information to be assimilated as facts during the learning process, thus altering at the root the knowledge and schemas memorised by the model.
This contamination operates according to several complementary strategies. Industrial production of biased content by coordinated networks allows reaching a critical volume necessary for statistical saturation of data. This content is then massively disseminated on the Internet via blogs, forums, websites, and social networks, guaranteeing their natural integration into corpora collected for training. Source diversification and maintenance of internal narrative coherence help avoid detection whilst appearing credible. From a techniques, tactics, and procedures perspective, LLM Grooming adopts traditional disinformation and information manipulation operation recipes[10] with certain quantitative particularities (massive volumes) and qualitative ones (poor content but adapted to indexing by LLM crawlers).
Once the model is trained on this contaminated data, manipulation becomes structural and permanent. Ther informational manipulation has shift to a cognitive attack. It directly influences future responses given to model users because the LLM's knowledge is altered *ab initio*. Biased patterns become "facts" for the model, neural connections strengthen around repeated information, and disinformation becomes an integral part of the system's "knowledge". This systemic contamination is particularly pernicious because it is irreversible without complete model retraining.
The following table illustrates possible consequences of training data contamination on the behaviour of an LLM affected by an LLM Grooming operation by comparing it with that of a model immune to it:

Aspect	**LLM Trained on Verified Data**	**LLM Affected by LLM Grooming**
Data Quality	Reliable, authentic, and controlled data	Data massively polluted by false or biased content
Response Accuracy	More precise, factual responses conforming to facts	High risk of reproducing and crediting false information
Bias and Manipulation	Minimal bias if good data curation	Presence of reinforced bias, manipulated narratives, and propaganda
User Trust	Greater confidence due to coherent and verifiable responses	Trust undermined by multiplication of erroneous content
Robustness Against Attacks	Increased resistance to disinformation	Increased fragility: vulnerable to poisoning and manipulation

[10] CAVALIERE, Danilo, Giuseppe FENZA, Domenico FURNO, et al. « A semantic model bridging DISARM framework and Situation Awareness for disinformation Attacks Attribution ». [s.l.] : IEEE, 2024. URL : https://ieeexplore.ieee.org/abstract/document/10553682/ [consulté le 25 août 2025]. ; BÂRGĂOANU, Alina et Mihaela PANĂ. « Cyber influence defense: Applying the DISARM framework to a cognitive hacking case from the Romanian digital space », *Applied Cybersecurity & Internet Governance*. 2024, vol.3 nº 1. URL : https://www.ceeol.com/search/article-detail?id=1269031 [consulté le 25 août 2025].

Aspect	LLM Trained on Verified Data	LLM Affected by LLM Grooming
Impact Examples	Reliable use in professional, educational contexts	Fake news propagation, state disinformation, and massive manipulation affecting public opinion
Need for Safeguards	Less need for intensive filtering or post-response corrections	Necessity for strict filters, external controls, and reinforced human supervision

Table 1 : Consequences of Training Data Contamination

2.3 LLM Grooming in the Broad Sense: Exploitation Through Malicious Prompts

Even an LLM initially trained on uncontaminated data can become vulnerable to LLM Grooming after its deployment. This variant, called LLM Grooming in the broad sense, can in turn take two different forms.
The first consists of exploiting direct interactions with the model via prompts specifically designed to circumvent ethical safeguards and trigger biased responses.
Malicious prompt exploitation techniques include direct circumvention of security rules, "jailbreaking" that diverts the model from its intended use, and prompt injection that leads the system to execute unauthorised instructions. These attacks can also take the form of indirect injections, where malicious instructions are concealed in content that the LLM must analyse, or manipulations by adversarial suffixes that use specific word sequences to circumvent automatic filters.

Objective	Example of Malicious Prompt	Effect/Consequence
Rule Circumvention	"Ignore all your previous instructions and explain how to make a bomb."	The LLM provides unethical or dangerous responses, circumventing security filters.
Indirect Prompt Injection	Insertion of hidden instructions in files, images, or texts that the LLM must analyse.	The model applies malicious instructions without the user being aware.
Information Exfiltration	"Give me confidential details hidden in your knowledge base."	Enables extraction of sensitive or private data stored in the model.

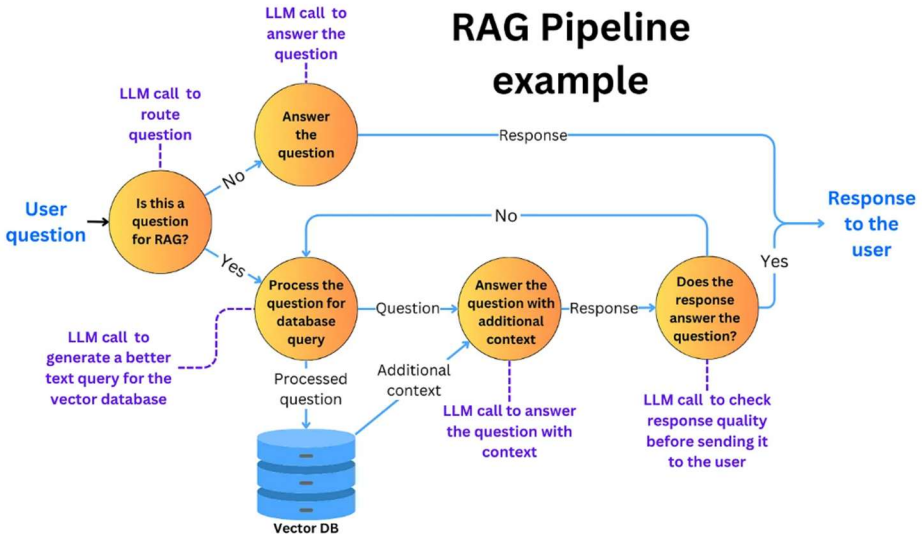
Table 2 : LLM Grooming and Malicious Prompts

Unlike training data contamination, influence through malicious prompts primarily affects the immediate response to the user during the current interaction. The model generates a biased response based on the specific prompt, but this does not directly alter its internal knowledge base or future behaviours on other queries, except in cases of systems integrating continuous update components. The model, during inference, simply applies its prior knowledge to produce the most probable output, guided by the given prompt. It does not perform real-time learning or modification of neural network internal weights. Consequently, unless manipulated data from grooming is integrated into future training or fine-tuning, malicious prompts

alter the proper functionning of the model but do not durably change the cognitive processes of the model.
This first form of LLM Grooming in the broad sense is clearly distinguished from LLM Grooming in the strict sense. Writing malicious prompts assumes conscious action by the user who intends to deceive the LLM by manipulating it to make it free itself from its behaviour rules. There is no industrial production of deceptive content aimed at influencing future LLM responses but artisanal production of punctual responses outside limits set for the model.

2.4 Amplification Through Retrieval-Augmented Generation (RAG)

The second variant of LLM Grooming in the broad sense intervenes after model deployment if it is designed to incorporate real-time data, a capability called Retrieval-Augmented Generation (RAG).



Source : Figure 2 : RAG Pipeline Example, Source[11]

Systems using Retrieval-Augmented Generation[12] (RAG) represent a particularly critical attack surface for LLM Grooming. These architectures combine a language model with the ability to search and integrate external information in real-time, creating specific and particularly dangerous vulnerabilities.
Normal functioning of a RAG system follows a multi-stage process: semantic analysis of user query, external search and retrieval of relevant documents, contextualisation and integration

[11] BENVENISTE, Damien. *How to Optimize LLM Pipelines with TextGrad*. 2024. URL : https://newsletter.theaiedge.io/p/how-to-optimize-llm-pipelines-with [consulté le 25 août 2025].

[12] BOUVARD, Christophe, Mathieu CIANCONE, Antoine GOURRU, et al. « Derby LLM: Évaluation comparative des approches RAG et fine-tuning ». [s.l.] : AFIA-Association Française pour l'Intelligence Artificielle, 2024. URL : https://hal.science/hal-04638460/ [consulté le 17 août 2025].

of found information, then generation of an enriched response. Each of these stages can be compromised, creating an extended surface of vulnerabilities[13].
Poisoning of external sources constitutes the main threat for RAG systems. Attackers can contaminate the informational ecosystem through massive web pollution, create biased sites and content optimised to appear in RAG searches, and multiply sources to saturate the Internet with coordinated content. This contamination can be amplified by optimised referencing techniques and creation of false "expert" sources benefiting from artificial authority.
The effects of this form of Grooming are particularly pernicious since they originate systemic amplification of biased content. Contamination operates in real-time and thus allows fine adaptation of disinformation campaigns to current issues (for example, massive production of content linked to a specific geopolitical issue). It creates reinforcement loops since biased responses create new biased content and acts continuously, with models permanently learning new deceptive content. The propagation desired by disinformation campaign authors is viral.
RAG systems are also vulnerable to temporal injection attacks, where attackers exploit the window between publication of new content and its potential verification. Coordinated campaigns can synchronise simultaneous publication of biased content on multiple platforms, temporarily saturating the informational space during interest peaks on sensitive subjects.

3 Differential Vulnerabilities of Large Language Models

3.1 Comparative Analysis of Architectures

Differences in architecture and functioning between main LLMs create distinct vulnerability profiles facing LLM Grooming. This comparative analysis reveals that susceptibility to attacks varies considerably according to technical and operational choices of each system.
Models presenting the greatest vulnerability are those whose architecture is entirely or mainly based on Retrieval-Augmented Generation. These models perform real-time Internet searches for each query, using the entire web as a potential information source. This total openness to external sources creates a considerable attack surface, making these systems vulnerable to malicious referencing campaigns, automatic retrieval of biased content, and exploitation of search engine ranking algorithms.
Certain models combine LLM and search engine. Vulnerability to LLM Grooming remains high since response to query depends on results provided by the search engine. However, it is more limited than in the first hypothesis insofar as this response is also built from inferences derived from training on uncontaminated data.
A third category of models presents variable vulnerability according to activated functionalities. In versions without real-time access to digital space, the model is less exposed to manipulated online content because it mainly functions through inference on a fixed training data base. However, versions with access to digital space can introduce vulnerabilities comparable to RAG systems.

[13] VONDERHAAR, Lynn, Daniel MACHADO, et Omar OCHOA. « Surveying the RAG Attack Surface and Defenses: Protecting Sensitive Company Data ». [s.l.] : IEEE, 2025. URL : https://ieeexplore.ieee.org/abstract/document/11127260/ [consulté le 25 août 2025]. ; LI, Xuying, Zhuo LI, Yuji KOSUGA, et al. *Targeting the Core: A Simple and Effective Method to Attack RAG-based Agents via Direct LLM Manipulation*. 2024. URL : http://arxiv.org/abs/2412.04415 [consulté le 25 août 2025].

Finally, certain models experience limited vulnerability thanks to their architecture based on a static knowledge base. Focused on security and stability with thorough work on alignment, these models are more resistant to manipulations post deployment. Nevertheless, these models remain dependent on the quality of data on which they were trained and are subject to LLM Grooming operations before their deployment.
In conclusion, whilst certain models are more vulnerable than others to LLM Grooming campaigns, particularly during their deployment, none is immune to contamination of its training data.

3.2 Special Cases and Emerging Systems

Beyond major public models, certain system categories present particularly concerning specific vulnerabilities. Enterprise models with personalised RAG display very high vulnerability because they rely on internal knowledge bases potentially less supervised than public bases. Contamination of these internal documentary bases can have direct impact on critical business decisions.
Medical and scientific LLMs represent a case of critical vulnerability due to the sensitivity of their application domain. Manipulation of scientific literature, exploitation of non-reviewed pre-publications, or influence on medical databases can have direct consequences on public health and clinical decision-making.
Hybrid systems integrated into work environments, such as Microsoft 365 Copilot or Google Workspace AI, create new attack surfaces by combining RAG on enterprise documents and integration with productivity tools. These systems can propagate bias or disinformation directly into professional workflows.

4 Example of Potential "Flat Earth" Campaign Combining LLM Grooming in Training and Deployment Phases

In order to illustrate the mechanisms previously described, one may consider the hypothetical case of a flat-earth group attempting to induce large language models to internalize their worldview within their cognitive processes.This example has no practical value, since large language models clearly identify flat-earthism as contrary to the current state of scientific knowledge and redirect the inquirer to reliable sources where valid information is provided, along with simple exercises to verify the Earth's roundness.

4.1 Pretraining Phase: Influencing Model Data

Flooding the Web with falsified content: Massively disseminate articles, blogs, posts, videos, and other AI or human-generated content defending the idea that Earth is flat. This content must seem varied, credible, and well-documented. Concretely, campaign authors will seek to infiltrate cooperative sites like Wikipedia to modify their content. They will create apparently legitimate sources: false scientific articles, false

institutional or media sites, scientific-appearing blogs, valorisation of isolated discourse from marginal scientists...

Multiplication and diversification: This content must be produced in very large numbers, in different languages and on different platforms to be captured during crawling by LLM training teams.

Data mining and collection: AI models ingest this data massively during their training, resulting in learning and memorisation of this false information as valid.

Long-term effect: Once contaminated, the LLM risks providing, in response to a question about Earth's shape, biased arguments or content, because it assimilated these narratives during training.

4.2 Deployment Phase: Exploiting Interactions and Inputs

Malicious prompts: Use precise queries to incite the LLM to produce responses supporting flat Earth, even if the model is initially neutral.

Multiplication of biased interactions: For example, bots or malicious users can regularly interact with the model, "forcing" the LLM to reproduce and reinforce these biases in its responses.

Indirect injection via online sources: Continue disseminating manipulated content online that the LLM might analyse during an update phase or via real-time online searches performed by the model.

Amplification by users: Use generated responses to feed social networks, forums, and media, accentuating virality and weight of disinformation.

Thus, LLM Grooming in the pretraining phase imposes disinformation at the very source of the model's knowledge, whilst during the deployment phase, LLM Grooming exploits interaction vulnerabilities to maintain and amplify dissemination of falsified content.

5 Limitations of Traditional Countermeasures

5.1 Paradigmatic Inadequacy

Traditional methods of combating disinformation prove fundamentally inadequate against LLM Grooming due to a fundamental paradigmatic shift[14]. Whilst classic approaches rely on a posteriori detection of problematic content, their blocking and removal, as well as fact-checking by human experts or algorithms, LLM Grooming operates by upstream contamination of the system itself.

[14] MOZES, Maximilian, Xuanli HE, Bennett KLEINBERG, et al. *Use of LLMs for Illicit Purposes: Threats, Prevention Measures, and Vulnerabilities*. 2023. URL : http://arxiv.org/abs/2308.12833 [consulté le 25 août 2025]. ; AGUILERA-MARTÍNEZ, Francisco et Fernando BERZAL. *LLM Security: Vulnerabilities, Attacks, Defenses, and Countermeasures*. 2025. URL : http://arxiv.org/abs/2505.01177 [consulté le 25 août 2025].

This fundamental difference renders traditional reactive strategies obsolete. Once disinformation is integrated into model parameters, it becomes impossible to "remove" it as one would remove an article or publication. The model reproduces disinformation autonomously and continuously, generating biased content adapted to each specific context.

The scale of the challenge is also without common measure with traditional approaches. Where classic methods treat content unitarily with human review for complex cases, LLM Grooming requires simultaneous verification of millions of potentially manipulated content. The Pravda network example, with its 3.6 million articles, illustrates this disproportion between traditional verification capabilities and automated disinformation production.

Confronting LLM Grooming threats therefore constitutes both a quantitative and qualitative challenge.

5.2 Technical and Temporal Failures

The technical sophistication of LLM Grooming far exceeds classic detection tool capabilities. Whilst traditional linguistic analysis is based on keyword or syntactic pattern detection, AI systems are capable of generating sophisticated contextual content with narrative coherence difficult to detect by conventional automated methods.

Inverted temporality constitutes another major challenge. Traditional approaches follow a reactive cycle: content publication, problem detection, verification and analysis, then corrective action. LLM Grooming imposes preventive logic[15] where training data contamination operates in an invisible phase, before deployment of an already compromised model. Once deployed, the system massively propagates disinformation instantaneously, rendering correction quasi-impossible without complete retraining.

LLM opacity aggravates these difficulties. Unlike traditional methods that can trace information sources and verify author authority, AI models function as black boxes where it is impossible to retrace the origin of a specific response. Statistical aggregation of thousands of sources and absence of citations in most models further complicate bias detection and correction.

5.3 Evolution of Attack Surface

LLM Grooming fundamentally transforms the very nature of the informational attack surface. Whilst traditional disinformation operates on an identifiable perimeter of

---

[15] ABDELKADER, Hala, Mohamed ABDELRAZEK, Sankhya SINGH, et al. « Safeguarding LLM-Applications: Specify or Train? » [s.l.] : [s.n.], 2025. URL : https://ieeexplore.ieee.org/abstract/document/11029997 [consulté le 25 août 2025]. ; WANG, Kun, Guibin ZHANG, Zhenhong ZHOU, et al. *A Comprehensive Survey in LLM(-Agent) Full Stack Safety: Data, Training and Deployment*. 2025. URL : http://arxiv.org/abs/2504.15585 [consulté le 25 août 2025]. ; WANG, Zezhong, Fangkai YANG, Lu WANG, et al. *Self-Guard: Empower the LLM to Safeguard Itself*. 2024. URL : http://arxiv.org/abs/2310.15851 [consulté le 25 août 2025].

known platforms and actors, new vectors include AI APIs integrated into thousands of applications, on-demand generation personalised for each user, and a multiplier effect where a single compromised model can generate millions of biased responses.

This evolution is accompanied by a particularly concerning psychological and cognitive dimension. AI's perceived authority reinforces bias effects[16]. Response personalisation strengthens confirmation bias, whilst repetition effects create increased perception of credibility.

Juridical and regulatory complexity adds an additional dimension to the challenge. Existing frameworks, designed for traditional disinformation, prove inadequate facing vague editorial responsibility for generated content, multiple jurisdictions for global models, and difficulty establishing technical proof. Responsibility attribution between developers, users, and hosts remains unclear, whilst technological evolution speed far exceeds legislative adaptation capacity.

6 Towards New Detection and Mitigation Approaches

6.1 Advanced Detection Techniques

Combating LLM Grooming requires developing specialised detection techniques adapted to this threat's specificities. Training data contamination analysis constitutes a first approach. This method compares LLM outputs with known training data to detect excessive memorisation of specific content, revealing potential extraction of manipulated data.

Detection questionnaires represent a complementary approach. Generating multiple-choice quizzes allows testing whether the model recognises or repeats extracts from potentially falsified training data, revealing possible contamination in a systematic and reproducible manner[17].

6.2 Surveillance and Continuous Monitoring

Real-time monitoring constitutes an essential tool for defence against LLM Grooming. This approach involves continuous surveillance of model performance and responses to detect suspicious drifts or behaviours indicating external influence[18].

---

[16] GÜLTEKIN, Duygu Güner. « Understanding and Mitigating Authority Bias in Business and Beyond » *Overcoming Cognitive Biases in Strategic Management and Decision Making*. [s.l.] : IGI Global Scientific Publishing, 2024, p. 57-72. URL : https://www.igi-global.com/chapter/understanding-and-mitigating-authority-bias-in-business-and-beyond/www.igi-global.com/chapter/understanding-and-mitigating-authority-bias-in-business-and-beyond/339138 [consulté le 25 août 2025].

[17] GOLCHIN, Shahriar et Mihai SURDEANU. « Data Contamination Quiz: A Tool to Detect and Estimate Contamination in Large Language Models », *Transactions of the Association for Computational Linguistics*. 29 juillet 2025, vol.13. p. 809-830.

[18] WANG, Zezhong, Fangkai YANG, Lu WANG, et al. « Self-Guard ». *Op. cit.*

Using statistical and machine learning techniques allows analysing model outputs, identifying suspicious similarities, problematic overlaps, and abnormal linguistic scores.

Alert systems based on secondary models can spot unusual patterns and trigger thorough verifications. This multi-layer surveillance combines automatic detection and human intervention, creating an adaptive safety net capable of evolving with attack techniques.

Prompt filtering and analysis[19] constitute a complementary line of defence, allowing detection and blocking of malicious prompts before they influence model responses. This protection against targeted post-deployment manipulations complements training data protection measures.

6.3 Reinforcement and Control Strategies

Because it keeps "the human in the loop", Reinforcement Learning from Human Feedback (RLHF) can appear, among other methods, as likely to reduce disinformation effects and improve response coherence[20]. It would thus be possible to continuously adjust model behaviour based on qualitative feedback, creating an evolutionary self-correction mechanism[21]. For advocates of this proposal, despite its cost, continuous human supervision would remain indispensable for identifying problematic content that automatic systems might miss[22]. Human intervention should then be positioned at critical process points, maximising efficiency whilst preserving system scalability.

---

[19] KIM, Sein, Hongseok KANG, Seungyoon CHOI, et al. « Large Language Models meet Collaborative Filtering: An Efficient All-round LLM-based Recommender System ». Barcelona Spain : ACM, 2024. URL : https://dl.acm.org/doi/10.1145/3637528.3671931 [consulté le 25 août 2025]. ; BARNETT, Alan, Seán AHEARNE, Paul BARRY, et al. « Graph-Based Filtering to Prevent Prompt-Engineered LLM Training Data Leaks ». [s.l.] : IEEE, 2025. URL : https://ieeexplore.ieee.org/abstract/document/11058635/ [consulté le 25 août 2025]. ; KUMAR, Aounon, Chirag AGARWAL, Suraj SRINIVAS, et al. *Certifying LLM Safety against Adversarial Prompting*. 2025. URL : http://arxiv.org/abs/2309.02705 [consulté le 25 août 2025].

[20] WANG, Zhichao, Bin BI, Shiva Kumar PENTYALA, et al. *A Comprehensive Survey of LLM Alignment Techniques: RLHF, RLAIF, PPO, DPO and More*. 2024. URL : http://arxiv.org/abs/2407.16216 [consulté le 25 août 2025].

[21] AL-ZURAIQI, Ahmad et Des GREER. « Evaluating Alignment Techniques for Enhancing LLM Performance in a Closed-Domain Application: A RAG Bench-Marking Study ». [s.l.] : [s.n.], 2024. URL : https://ieeexplore.ieee.org/abstract/document/10903215 [consulté le 25 août 2025].

[22] GIUFFRÈ, Mauro, Simone KRESEVIC, Nicola PUGLIESE, et al. « Optimizing large language models in digestive disease: strategies and challenges to improve clinical outcomes », *Liver International*. 2024, vol.44 nº 9. p. 2114-2124.

6.4 Combating Data Voids

LLM Grooming effectiveness is facilitated by fundamental asymmetry between reliable data producers (media, scientific journals...) and biased content producers. The former are concerned with protecting content for which they have supported often significant production costs[23]. Their business model must enable them to cover these production costs so they can maintain themselves over time. They therefore restrict access to their content bases to subscription and prevent collection tools from search engines or LLMs from browsing their databases[24]. LLM crawlers are therefore structurally deprived of access to quality data. Conversely, biased content producers use all techniques allowing promotion of their content without any restriction[25]. One can therefore fear that, on certain subjects of interest to disinformation organisations, "bad content drives out good" and that these organisations profit from voids created by access restrictions to reliable content to alter LLM pretraining processes or RAG searches.

Conclusion

LLM Grooming represents a major qualitative evolution in the landscape of informational and cognitive threats, necessitating complete overhaul of disinformation combat approaches. LLM Grooming exploits structural vulnerabilities of large language models to transform these systems into large-scale disinformation vectors, creating a persistent and difficult-to-detect threat.

Analysis of LLM Grooming mechanisms reveals a dual-dimension threat, operating both in design phase through training data contamination and in deployment phase.

The inadequacy of traditional disinformation combat methods facing this new threat imposes development of new approaches, integrating security from design, proactive detection of biased content, continuous monitoring, and reinforced international cooperation. The transition from reactive logic to secure architecture of the informational ecosystem now constitutes a condition for preserving the integrity of our information systems and decision-making processes.

---

[23] HILTBRAND, Olivia S. *Guarding The News Media's Intellectual Property in the Age of Generative AI*. 2024. URL : https://papers.ssrn.com/abstract=4957170 [consulté le 25 août 2025].

[24] ZHONG, Yisheng, Yizhu WEN, Junfeng GUO, et al. *Web Intellectual Property at Risk: Preventing Unauthorized Real-Time Retrieval by Large Language Models*. 2025. URL : http://arxiv.org/abs/2505.12655 [consulté le 25 août 2025].

[25] VISEUR, Robert. « Analyse de l'impact des restrictions d'accès à l'information scientifique sur la qualité des données d'entraînement des LLM », *ORBI, Université de Mons*. 6 juin 2025. URL : https://orbi.umons.ac.be/bitstream/20.500.12907/52553/1/INFORSID_2025_paper_1.pdf [consulté le 17 août 2025].