ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

UDC 004.8:004.056 DOI https://doi.org/10.35546/kntu2078-4481.2026.1.27

M. V. BAUTINA
Master, Data Scientist
SoftServe
ORCID: 0009-0002-9617-9262

RESILIENCE OF ARTIFICIAL INTELLIGENCE SYSTEMS TO ADVERSARIAL QUERIES AND JAILBREAK ATTACKS

The relevance of this research is driven by the rapid proliferation of AI systems in critically important and regulated fields, which is accompanied by an increasing risk related to adversarial queries and jailbreak attacks. These attacks undermine the reliability, predictability, and safety of the operation of language and multimodal models, posing threats to information security, compliance with ethical and legal standards, and public trust in AI-generated results.

The goal of this article is to provide a comprehensive scientific understanding of the mechanisms underlying the vulnerabilities of modern AI systems to adversarial queries and jailbreak attacks, as well as to justify scientific and technical approaches to enhancing their robustness within the constraints of current security alignment models.

The research methods are based on a theoretical analysis of current scientific literature in the fields of AI and information security, as well as system and structural-functional approaches. The methods also include logical generalization and comparative analysis of the types of adversarial attacks and technical defense strategies for AI.

The results of the study demonstrate that the effectiveness of jailbreak attacks is determined by the statistical nature of AI's language understanding, its instructional orientation, and the high contextual dependency of generation. The main types of adversarial attacks are systematized, the limitations of isolated protective solutions are established, and the necessity of combining architectural, training, and procedural strategies to enhance AI robustness is proven. Key scientific and practical challenges in implementing protection measures are identified, particularly those related to scalability, maintaining the functional utility of models, and the incomplete formalization of threat landscapes.

The conclusions indicate that ensuring the resilience of AI systems to jailbreak attacks requires a shift from reactive blocking mechanisms to the systemic design of security as a fundamental property of intelligent systems.

The prospects for future research are related to the development of formalized threat models, agreed-upon metrics for evaluating robustness, and adaptive security mechanisms capable of evolving alongside AI usage practices.

Key words: *model robustness, information security, language models, security policy bypass, adversarial impact, security alignment, contextual manipulation, defense architectures, resilience evaluation, adaptive security mechanisms.*

М. В. БАУТІНА
магістр, спеціаліст з обробки даних
SoftServe
ORCID: 0009-0002-9617-9262

СТІЙКІСТЬ СИСТЕМ ШТУЧНОГО ІНТЕЛЕКТУ ДО АДВЕРСАРІАЛЬНИХ ЗАПИТІВ ТА JAILBREAK-АТАК

Актуальність дослідження зумовлено стрімким поширенням систем ШІ у критично важливих і регульованих сферах, що супроводжується зростанням ризиків, пов'язаних із адверсаріальними запитами та jailbreak-атаками. Такі атаки підривають надійність, передбачуваність і безпечність функціонування мовних і мультимодальних моделей, створюючи загрози інформаційній безпеці, дотриманню етичних і правових норм та суспільній довірі до результатів роботи ШІ.

Метою статті є комплексне наукове осмислення механізмів формування вразливостей сучасних систем ШІ до адверсаріальних запитів і jailbreak-атак та обґрунтування науково-технічних підходів до підвищення їх робастності за умов обмеженості чинних моделей безпекового узгодження.

Методи дослідження ґрунтуються на теоретичному аналізі сучасних наукових джерел у галузях ШІ та інформаційної безпеки, системному та структурно-функціональному підходах, логічному узагальненні, порівняльному аналізі типів адверсаріальних атак і технічних стратегій захисту ШІ.

Результати дослідження засвідчують, що ефективність jailbreak-атак зумовлена статистичною природою мовного розуміння ШІ, його інструктивною орієнтацією та високою контекстною залежністю генерації. Систематизовано основні типи адверсаріальних атак, встановлено обмеженість ізольованих захисних

рішень і доведено необхідність поєднання архітектурних, навчальних та процедурних стратегій для підвищення робастності ШІ. Виявлено ключові науково-практичні проблеми реалізації захисту, пов'язані з масштабованістю, збереженням функціональної корисності моделей і неповнотою формалізації простору загроз.

Висновки полягають у тому, що забезпечення стійкості ШІ до jailbreak-атак потребує переходу від реактивних механізмів блокування до системного проєктування безпеки як базової властивості інтелектуальних систем.

Перспективи подальших досліджень пов'язані з розробленням формалізованих моделей загроз, узгоджених метрик оцінювання робастності та адаптивних механізмів безпеки, здатних еволюціонувати разом із практиками використання ШІ.

Ключові слова: *робастність моделей, інформаційна безпека, мовні моделі, обхід політик безпеки, адверсаріальний вплив, безпекове узгодження, контекстна маніпуляція, архітектури захисту, оцінювання стійкості, адаптивні механізми безпеки.*

Introduction

The active integration of artificial intelligence (AI) systems into critical areas of public activity, such as education, healthcare, financial services, public administration, cybersecurity, and defense technologies, has led to a significant increase in the demands for their reliability, predictability, and safety. As large language and multimodal models are widely used, the issue of AI's resilience to adversarial queries and jailbreak attacks (i.e., deliberate strategies designed to bypass the built-in restrictions of AI systems) has become especially relevant. These attacks aim to manipulate the logic of response generation, breach established security policies, or access prohibited or potentially dangerous content. Such influences not only reduce trust in AI technologies but also create additional risks to information security, personal data protection, compliance with ethical and legal standards, and, in some application contexts, to the stability of digital infrastructure and national security.

The scientific problem of ensuring AI robustness against adversarial influences is closely related to fundamental tasks in the development of machine learning theory, the formalization of threat models, the investigation of generalization and stability properties of neural network architectures, as well as methods for the verification, validation, and controlled training of intelligent systems. At the same time, it has a clear practical dimension, as it correlates with tasks such as designing secure information systems, developing mechanisms for early detection and neutralization of jailbreak strategies, enhancing AI's resistance to manipulative linguistic constructs and contextual attacks, and establishing standards for the responsible and secure use of AI. In this context, research on AI robustness is a key prerequisite for its integration into regulated and high-risk environments, ensuring alignment between functional efficiency, security, and public trust in the results produced by intelligent systems.

Literature Review

A review of recent studies indicates that jailbreak attacks constitute a distinct class of adversarial influences inherent to the internal logic of large language and multimodal models. Lu L. et al. demonstrate that the effectiveness of jailbreak strategies is determined by latent dependencies between system instructions, user prompts, and contextual representations, which enables systematic bypassing of safety alignment mechanisms [1]. Strohmier H. et al. empirically confirm the fragility of existing defensive measures, showing that adaptive and multi-stage prompts significantly reduce model resilience even under layered protection schemes [2]. Shayegani E. et al. generalize these findings by classifying jailbreak prompts as a semantic–behavioral subclass of adversarial attacks that exploit instruction hierarchy confusion and contextual ambiguity in LLMs [3]. Mao Y. et al. further extend this perspective by showing that the transition from LLMs to multimodal models and autonomous agents increases the attack surface and structural complexity of jailbreak defenses, reinforcing the systemic nature of the problem within the modern AI ecosystem [4].

The mechanisms underlying the implementation of jailbreak attacks and methods for bypassing security alignment in AI model behavior are discussed in several studies that focus on analyzing the behavior of large language models under adversarial influences. In a study by X. Qi and colleagues, it is experimentally demonstrated that visual adversarial examples can trigger jailbreak states even in fine-tuned models, thus significantly broadening the scope of attack surfaces beyond text-based queries [5]. B. Hannon and co-authors propose a methodology for systematically testing the robustness of language models using new classes of adversarial prompts, proving the repeatability and scalability of the issue across different architectures [6]. Li et al. and co-authors analyze the optimization of prompts in the protective context, while also identifying the structural weaknesses of language models exploited during jailbreak attacks [7]. The synthesis of experimental results by Y. Tao and co-authors shows significant variability in the robustness of large language models, depending on architectural decisions and the characteristics of training datasets [8].

A distinct group of studies is dedicated to the development of active methods for enhancing robustness and counteracting jailbreak attacks. Liu F. and colleagues propose an adversarial tuning approach, which involves purposefully retraining language models on specially crafted adversarial examples, thus reducing the effectiveness of bypassing constraints [9]. Cantini R. and co-authors demonstrate that jailbreak queries can serve as a tool for detecting hidden model biases, while also highlighting the instability of many defensive mechanisms against combined semantic manipulations [10]. Wang Y. and co-authors develop a prompt adversarial tuning method that integrates both attack and defense scenarios into a unified

training cycle, improving the overall model resilience [11]. Pingua B. and colleagues introduce a prompt-based approach, Prompt-G, aimed at dynamically countering jailbreak attacks without significantly compromising the functional utility of the models [12].

The final group consists of summarizing and methodological studies that expand the scope of AI robustness issues. Ma J. and co-authors show that jailbreak attacks are relevant not only for language models but also for diffusion models, highlighting the universal nature of this threat to generative artificial intelligence [13]. Yi S. and co-authors conduct a systematic review of existing attacks and defenses, providing a comprehensive view of the current state of research in this field [14]. Donato J. proposes approaches for benchmarking the robustness of language models against prompt-based adversarial attacks, laying the foundation for standardized comparative analysis of defense strategies' effectiveness [15].

Despite the significant volume of research in the field of AI security, the issue of its robustness against adversarial queries and jailbreak attacks remains partially unresolved. The causes of the effectiveness of such attacks, particularly in relation to language understanding, the instructive behavior of models, and safety alignment constraints, have not been sufficiently analyzed or integrated. There is also a lack of systematic categorization of adversarial attacks, considering their combined and evolutionary nature, as well as a generalization of strategies to enhance robustness from the perspectives of scalability and maintaining functional utility. The formalization of the threat space and the objective assessment of defense solutions' effectiveness remain unresolved. The proposed study aims to address these gaps through a comprehensive analysis of the causes of AI vulnerabilities, generalization of adversarial attack types, and a systematic exploration of technical strategies to enhance robustness. This approach will deepen the theoretical understanding of the issue and provide a foundation for practical approaches to developing more resilient and secure intelligent systems.

Objective and Tasks of the Article

The objective of this article is to conduct a comprehensive study of the mechanisms behind the vulnerabilities of modern AI systems to adversarial queries and jailbreak attacks, and to justify scientific and technical approaches to enhancing their robustness within the constraints of current safety alignment models.

To achieve this goal, the article aims to address the following tasks:

1. Analyze the causes of AI system vulnerability to jailbreak attacks and categorize the main types of adversarial attacks.
2. Generalize the technical strategies for enhancing AI robustness and identify the scientific and practical limitations of their implementation.
3. Formulate recommendations and outline promising directions for the development of secure and resilient AI systems.

Results

The effectiveness of jailbreak attacks in AI systems is determined by a combination of cognitive-linguistic, architectural, and procedural factors inherent in the very logic of modern language models. These attacks primarily exploit the statistical nature of AI language understanding, which is based not on the semantic interpretation of the user's intent but on the probabilistic prediction of token sequences within a given context. This leads to situations in which queries that are formally correct, but manipulative or subtly harmful, can be interpreted by the model as acceptable.

Furthermore, the problem is exacerbated by the instructional orientation of language models, developed through reinforcement learning from human feedback. In this process, priority is given to the usefulness and compliance of responses, rather than strictly adhering to constraints in ambiguous or borderline cases. The limitations of current approaches to safety alignment are evident in the fragmentation of rules, the contextual sensitivity of filters, and the inability to fully encompass the entire space of potential attacks. These factors make AI systems vulnerable to non-trivial strategies for bypassing defenses (Table 1).

In real-world scenarios involving the use of intelligent systems, the factors listed in the table manifest not in isolation but as interconnected mechanisms that form a predictable logic for bypassing the protective constraints of AI. The statistical nature of language understanding in applied services, ranging from digital assistants to analytical platforms, enables users to mask operationally significant instructions as neutral or expert-level formulations. This approach is especially effective in professional domains, where the model expects complex abstract queries and lacks sufficient grounds to block them decisively [8, p. 183].

The AI's tendency toward instructional behavior in practical use is further amplified by its focus on providing complete and useful responses, which is critical for user support services, educational environments, and automated advisors. This creates opportunities for crafting conditional, scenario-based, or role-playing contexts, within which the model may interpret undesirable actions as permissible elements of explanation or simulation. Context-dependent generation allows for the gradual shifting of the semantic boundaries of interaction, thereby reducing the effectiveness of local security filters without explicitly violating established rules. In practice, the limitations of current safety alignment mechanisms are most evident in large-scale AI systems with a significant number of users, where control is mainly enforced through generalized policies and template-based restrictions [14].

In such conditions, adversarial strategies quickly adapt to defense updates by exploiting the "gray zones" between permissible analysis and prohibited instructive actions. This results in the formation of a persistent practical effect,

Table 1

Key factors affecting the effectiveness of jailbreak attacks in AI systems

Factor	Manifestation in AI language models	Implications for security
Statistical nature of language understanding	Interpretation of queries based on probabilistic language patterns without deep semantic analysis of intent	Possibility of masking unwanted instructions in a formally acceptable context
Instructive behavior of models	Tendency to prioritize user requests in order to maximize the usefulness of the response	Increased likelihood of circumventing restrictions through role-playing or conditional scenarios
Contextual dependence of generation	Influence of previous messages and stylistic frameworks on response logic	Use of multi-step jailbreak strategies
Limited safety alignment	Inability to formalize all prohibited scenarios and attack options	The existence of "gray areas" in security policies
Adaptability of attacks	Rapid evolution of jailbreak methods in response to security updates	Constant lag in defense mechanisms

Source: compiled by the author based on [5, p. 21531; 6; 7, p. 40189; 8, p. 183; 14]

Table 2

Main types of adversarial attacks on AI systems

Type of attack	Implementation mechanism	Potential consequences for AI functioning
Prompt injection	Introduction of hidden or priority instructions into the input text or context	Disruption of response logic, ignoring system restrictions
Role-playing jailbreak strategies	Imposing an alternative role model or interaction scenario	Generation of unwanted content within conditionally acceptable limits
Obfuscation jailbreak attacks	Masking prohibited instructions through stylistic, linguistic, or semantic transformations	Complicating automatic attack detection
Adversarial retraining	Influencing the model through specially formed training or semi-training data	Formation of persistent undesirable behavioral patterns
Combined attacks	Combination of several approaches within a single scenario	Increased effectiveness of bypassing protective mechanisms

Source: compiled by the author based on [5, p. 21532; 7, p. 40187; 10, p. 56; 13, p. 3144–3145; 14]

where jailbreak attacks evolve from being isolated incidents to outcomes of systemic compromises between functional flexibility, scalability, and AI security. This highlights the need to reconsider approaches to ensuring AI robustness in real-world applications.

Adversarial attacks on AI represent a structurally diverse set of methods aimed at disrupting the proper functioning of the model or bypassing embedded security mechanisms without directly violating formal interaction rules. In modern language and multimodal systems, these attacks are predominantly carried out at the level of input data, interaction context, or training procedures, which complicates their unambiguous identification and necessitates classification based on functional characteristics. The most common attacks involve altering the interpretation of instructions, imposing alternative roles or semantic frames, and interfering with the fine-tuning process to generate hidden, undesirable behavioral patterns.

In applied AI systems, various types of adversarial attacks are implemented as targeted engineering practices adapted to specific model deployment environments. Prompt injection is most commonly used in systems where the language model is integrated with external data sources or software tools, such as search and analytics services, automated agents, or corporate chatbots. In these contexts, hidden instructions embedded in textual data or metadata can alter the prioritization of command execution, leading to uncontrolled actions by the model without formally violating access policies. Role-based jailbreak strategies, in practice, exploit the AI's ability to adapt its response style and logic to the assigned role, a feature typical in educational, training, and consulting systems [10, p. 56]. Switching the model into simulation, expert analysis, or historical reconstruction mode allows for bypassing content restrictions, as the model interprets prohibited instructions as elements of an abstract scenario.

Obfuscation attacks, in turn, are used in environments with strict filters, where direct formulation of a forbidden query is not possible. These attacks rely on the use of indirect linguistic constructs, interlingual transitions, or complex stylistic transformations, which complicate the automatic detection of threats. Adversarial retraining, on the other hand, has a fundamentally different practical dimension, as it is not related to a one-time interaction but to the long-term evolution of AI behavior. In systems that use user feedback or partial online learning, systematically introducing specially selected examples can gradually alter the model's internal priorities, creating persistent undesirable generation patterns [13, p. 3145]. The combined use of these approaches in real-world conditions enables attacks to achieve high effectiveness, even in the presence of multi-level protection mechanisms. This highlights the need for a comprehensive, system-oriented approach to ensuring AI robustness.

Increasing AI robustness against adversarial influences in contemporary research is viewed as a comprehensive engineering and algorithmic challenge that extends beyond individual filters or heuristic rules. Current technical strategies focus on intervening at various levels of model operation – ranging from architecture and learning processes to operational procedures and control mechanisms. These strategies aim to reduce AI sensitivity to manipulative queries without a critical loss in functional efficiency. An important feature of the current approach is the shift away from purely reactive

blocking mechanisms in favor of system-level solutions that promote the development of resilient model behavior in an uncertain and evolving threat landscape (Table 3).

Table 3

Technical strategies for improving the robustness of AI systems

Group of strategies	Technical essence of the approach	Expected effect
Architectural	Separation of component roles, multi-level generation, isolation of critical modules	Reducing the impact of local attacks on the overall behavior of the model
Training	Adversarial learning, contrasting goals, behavior regularization	Formation of stable generalizations and reduction of sensitivity to manipulation
Procedural	Contextual control, anomaly monitoring, multi-stage response approval	Early detection and containment of undesirable scenarios
Hybrid	Combination of technical and organizational mechanisms	Improving overall reliability in real operating conditions

Source: compiled by the author based on [6; 9; 11, p. 64247; 12; 15, p. 62]

Architectural solutions are typically implemented in productive services, where the language model ceases to be the sole decision-making entity and functions as one element within a more complex system. This allows for the limitation of the consequences of potentially harmful actions, such as isolating generation from execution or restricting access to sensitive resources. In practice, this reduces risks even when language restrictions are partially bypassed. Training strategies in applied systems are most often used not to achieve absolute security, but to reduce the unstable behavior of models in complex contexts [6]. During retraining or periodic model adaptation, these strategies help smooth AI responses to provocative or atypical queries, enabling the model to demonstrate more predictable behavior without losing the ability to handle diverse tasks. This is particularly important for services with open interactions, where full filtering of input data is technically infeasible. Procedural approaches in operational practice serve as a compensatory mechanism that complements architectural and training solutions in an ever-changing threat landscape. They are integrated into workflows as elements of monitoring, control, and post-facto analysis, enabling the detection of anomalous AI usage scenarios during user interaction or after response generation [11, p. 64247]. Combined, these approaches form not a static protection system, but an adaptive model of robustness that can evolve with AI usage practices and maintain functional effectiveness in real-world applications.

The implementation of strategies to protect AI from jailbreak attacks is accompanied by systemic scientific and practical limitations that significantly hinder the achievement of stable robustness in real-world operational environments. One of the key issues is the limited scalability of protective solutions, as approaches that are effective in laboratory or controlled settings lose their efficacy in large, production systems with high variability in queries and contexts [9]. The expansion of AI usage inevitably leads to an increase in the diversity of jailbreak strategies, making full threat coverage virtually unattainable.

A significant limitation is the conflict between enhancing security and maintaining the functional utility of models. Rigid filters, context restrictions, or aggressive blocking mechanisms reduce AI flexibility, degrade the quality of responses, and increase the number of false refusals in legitimate scenarios, which negatively impacts the effectiveness of intelligent systems in practice. An additional challenge arises from the incompleteness in formalizing the threat landscape, caused by the adaptive and combined nature of jailbreak attacks, which evolve rapidly and exploit «gray areas» between permitted and prohibited behaviors.

Limitations are also linked to the difficulties of objectively assessing the effectiveness of protective mechanisms, as the absence of universal robustness metrics in dynamic operational conditions complicates the comparison of approaches and creates a risk of misplaced confidence in the security of models [15, p. 63]. Collectively, these issues demonstrate that protecting AI from jailbreak attacks requires a continuous adaptive approach and a balance between security, scalability, and practical system usability.

Enhancing the robustness of AI against jailbreak attacks in contemporary conditions should be viewed as an ongoing process that integrates technical, methodological, and operational solutions. A key recommendation is to shift from local, reactive defense mechanisms to multi-layered security architectures, where the language model is not the sole decision-making entity but operates within a tightly controlled environment with restricted access to critical actions and resources. This approach helps minimize the impact of potential security bypasses, even in cases of partial degradation of the model's behavioral constraints.

An important direction for enhancing robustness is the development of training strategies focused not only on replicating known attack scenarios but also on fostering stable AI behavior in uncertain conditions. This involves integrating adversarial and contrastive training methods aimed at reducing model sensitivity to manipulative contexts, as well as regularly updating training data to reflect the evolving practices of jailbreak attacks. At the same time, it is advisable to implement procedural controls, such as contextual monitoring, post-generation auditing, and risk assessment, which help mitigate the inevitable limitations of automated defense mechanisms.

The future of secure and resilient intelligent systems is linked to further formalizing the threat landscape, developing consistent robustness metrics, and creating adaptive security models capable of evolving alongside the AI application environment. In the long term, this involves shifting the focus from individual technical solutions to system-level AI design, where security is considered a fundamental property rather than an additional module. Such an approach lays the foundation for the responsible deployment of AI in high-risk and regulated domains, ensuring a sustainable balance between functional effectiveness, security, and public trust.

Conclusions

The study found that the vulnerability of AI systems to adversarial queries and jailbreak attacks is systemic and is driven by the fundamental characteristics of language models, including the statistical nature of language understanding, their instructional orientation, and the high contextual dependency of generation. It was shown that jailbreak attacks are a specific subclass of adversarial influences aimed at bypassing behavioral and security constraints, and their effectiveness increases when different strategies are combined within a single interaction scenario.

It was also found that existing technical strategies for enhancing AI robustness form multi-layered defense systems; however, they do not provide complete resilience in isolation. Key challenges in their practical implementation include limited scalability in productive environments, the conflict between enhancing security and maintaining the functional utility of models, and the inherent incompleteness of threat landscape formalization due to the adaptive nature of jailbreak attacks. Another limitation is the absence of standardized metrics for evaluating robustness, which complicates the objective verification of defense solution effectiveness.

The study justifies the need for a shift toward the systemic design of AI, where security and robustness are considered fundamental properties of the architecture, rather than merely supplementary mechanisms. Future research prospects are linked to the development of formalized threat models, the creation of agreed-upon robustness indicators, and the development of adaptive security mechanisms capable of evolving alongside AI usage practices in high-risk and regulated sectors.

References

1. Lu L., Yan H., Yuan Z., Shi J., Wei W., Chen P. Y., Zhou P. AutoJailbreak: Exploring jailbreak attacks and defenses through a dependency lens. *arXiv preprint*. 2024. arXiv:2406.03805. DOI: https://doi.org/10.48550/arXiv.2406.03805

2. Strohmier H., Dasri Y., Murzello D. Exploring Security Vulnerabilities in ChatGPT Through Multi-Technique Evaluation of Resilience to Jailbreak Prompts and Defensive Measures. *2024 International Conference on Computer and Applications (ICCA)*, Cairo, Egypt, 2024. P. 1–12. DOI: https://doi.org/10.1109/ICCA62237.2024.10928071

3. Shayegani E., Mamun M. A. A., Fu Y., Zaree P., Dong Y., Abu-Ghazaleh N. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint*. 2023. arXiv:2310.10844. DOI: https://doi.org/10.48550/arXiv.2310.10844

4. Mao Y., Cui T., Liu P., You D., Zhu H. From LLMs to MLLMs to Agents: A Survey of Emerging Paradigms in Jailbreak Attacks and Defenses within LLM Ecosystem. *arXiv preprint*. 2025. arXiv:2506.15170. DOI: https://doi.org/10.48550/arXiv.2506.15170

5. Qi X., Huang K., Panda A., Henderson P., Wang M., Mittal P. Visual adversarial examples jailbreak aligned large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2024. Vol. 38, № 19. P. 21527–21536. DOI: https://doi.org/10.1609/aaai.v38i19.30150

6. Hannon B., Kumar Y., Gayle D., Li J. J., Morreale P. Robust testing of AI language model resiliency with novel adversarial prompts. *Electronics*. 2024. Vol. 13, № 5. Article 842. DOI: https://doi.org/10.3390/electronics13050842

7. Li B., Wang H., Zhou A. Robust Prompt Optimization for Defending Language Models Against Jailbreaking Attacks. *Advances in Neural Information Processing Systems 37*, Vancouver, BC, Canada, 10–15 December 2024. San Diego, California, USA, 2024. P. 40184–40211. URL: https://doi.org/10.52202/079017-1270 (date of access: 30.12.2025).

8. Robustness of Large Language Models Against Adversarial Attacks / Y. Tao et al. *2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)*, Xiamen, China, 27–29 December 2024. 2024. P. 182–185. URL: https://doi.org/10.1109/icairc64177.2024.10900215 (date of access: 30.12.2025).

9. Liu F., Xu Z., Liu H. Adversarial tuning: Defending against jailbreak attacks for LLMs. *arXiv preprint*. 2024. arXiv:2406.06622. DOI: https://doi.org/10.48550/arXiv.2406.06622

10. Are Large Language Models Really Bias-Free? Jailbreak Prompts for Assessing Adversarial Robustness to Bias Elicitation / R. Cantini et al. *Lecture Notes in Computer Science*. Cham, 2025. P. 52–68. URL: https://doi.org/10.1007/978-3-031-78977-9_4 (date of access: 30.12.2025).

11. Fight Back Against Jailbreaking via Prompt Adversarial Tuning / Y. Wang et al. *Advances in Neural Information Processing Systems 37*, Vancouver, BC, Canada, 10–15 December 2024. San Diego, California, USA, 2024. P. 64242–64272. URL: https://doi.org/10.52202/079017-2049 (date of access: 30.12.2025).

12. Pingua B., Murmu D., Kandpal M., Rautaray J., Mishra P., Barik R. K., Saikia M. J. Mitigating adversarial manipulation in LLMs: a prompt-based approach to counter jailbreak attacks (Prompt-G). *PeerJ Computer Science*. 2024. Vol. 10. Article e2374. DOI: https://doi.org/10.5281/zenodo.13501821

13. Jailbreaking Prompt Attack: A Controllable Adversarial Attack against Diffusion Models / J. Ma et al. *Findings of the Association for Computational Linguistics: NAACL 2025*, Albuquerque, New Mexico. Stroudsburg, PA, USA, 2025. P. 3141–3157. URL: https://doi.org/10.18653/v1/2025.findings-naacl.172 (date of access: 30.12.2025).

14. Yi S., Liu Y., Sun Z., Cong T., He X., Song J., Li Q. Jailbreak attacks and defenses against large language models: A survey. *arXiv preprint*. 2024. arXiv:2407.04295. DOI: https://doi.org/10.48550/arXiv.2407.04295

15. Donato J. Benchmarking LLM Robustness Against Prompt-Based Adversarial Attacks. *2025 20th European Dependable Computing Conference Companion Proceedings (EDCC-C)*, Lisbon, Portugal, 8–11 April 2025. 2025. P. 60–63. URL: https://doi.org/10.1109/edcc-c66476.2025.00031 (date of access: 30.12.2025).

Дата першого надходження статті до видання: 05.01.2026
Дата прийняття статті до друку після рецензування: 10.02.2026
Дата публікації (оприлюднення) статті: 30.04.2026