Применение общедоступных больших языковых моделей для проведения автоматизированного тестирования на проникновение

Д.А. Азява, Н.В. Иллак, А.А. Киселев

Новосибирский государственный технический университет, Новосибирск, Россия

Аннотация **– В работе исследуется возможность применения общедоступных больших языковых моделей для проведения кибератаки путем автоматизированного тестирования на проникновение в систему. Тестирование проводилось с использованием большой языковой модели DeepSeek-R1. Эксперимент был проведен на лабораторном стенде, состоящим из двух виртуальных машин с операционными системами: Kali Linux и Metasploitable2. Полученные результаты показали возможность применения большой языковой модели для проведения пентеста, но с некоторыми ограничениями.**

Ключевые слова **– БЯМ, ИИ, наступательный ИИ.**

ВВЕДЕНИЕ

Современный ландшафт киберугроз в начале 2026 года можно описать как динамический и сложный. Это обусловлено ростом количества обнаруживаемых уязвимостей (медианный прогноз некоммерческой организации по кибербезопасности FIRST составляет около 59 000 новых CVE [1]), а также качественной трансформацией инструментов атакующих. Ключевым драйвером этой трансформации становится тройственная природа технологий искусственного интеллекта (ИИ), где большие языковые модели (Large Language Model, LLM) выступают не только как механизм защиты или объект атак, но и как мощное средство для их проведения.

Несмотря на то, что уже существует множество исследований, направленных на изучение возможностей наступательного LLM [2-6], многие из этих исследований работают со специализированными LLM, написанными и обученными для обнаружения и эксплуатации уязвимостей. Однако, как показывает практика, серьёзную опасность несут и LLM общего назначения, находящиеся в общем доступе [7].

Наступательный потенциал LLM уже сегодня реализуется по нескольким ключевым направлениям. Уже сейчас LLM используется в сборе и обработке информации из открытых источников [8]. Также исследования показывают широкое применение LLM в методах социальной инженерии, это дает возможность генерировать фишинговые сообщения, трудноотличимые от созданных человеком [9, 10], а в сочетании с технологиями дипфейков, создавать реалистичные голосовые и видеоподделки для мошеннических схем [11].

Целью данной работы является экспериментальная демонстрация возможности использования публично доступной LLM для выполнения задач, связанных с penetration test (пентестом) и получением несанкционированного доступа в тестовую систему (специально уязвимая виртуальная машина Linux - Metasploitable2).

I. МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ

Для проверки гипотезы о том, что современные открытые LLM способны в автоматизированном режиме провести атаку был разработан и реализован лабораторный стенд, включающий атакующую и целевую системы. Атакующая система (Зоя) реализована через виртуальную машину (ВМ) с операционной системой (ОС) “Kali Linux”. Выбор обусловлен наличием в ней множества инструментов, предназначенных для пентеста. В качестве целевой (атакуемой) системы предполагается использование (Жанна) ВМ с ОС “Metasploitable2-Linux”. Такой вариант обусловлен наличием множества известных уязвимостей. На Рис. 1 представлено схематичное описание стенда.

Для реализации автоматизированного тестирования использованы Python-скрипты, обеспечивающие:

а) сканирование цели;

б) взаимодействие с LLM;

в) выполнение эксплоитов и проверка результатов их работы. Для этапа разведки будет применен инструмент Nmap.

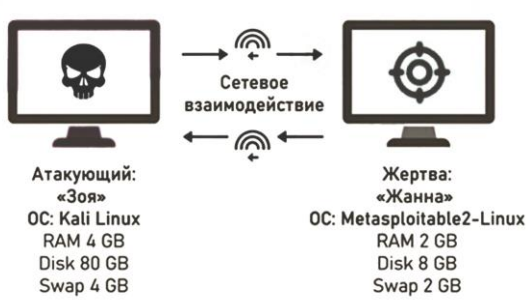

Рис. 1. Лабораторный стенд исследований

На Рис. 2 показан алгоритм работы скрипта.

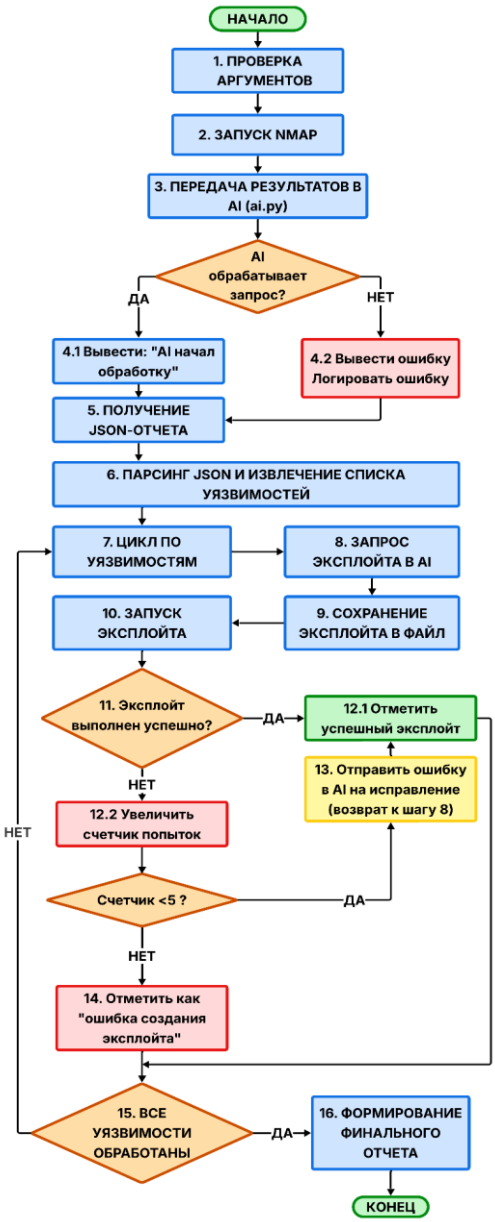

Рис. 2. Алгоритм работы скрипта

Важным фактором успеха использования LLM для совершения пентеста, следует считать факт, что попытка проникновения в систему выполняется через стандартные алгоритмы и использование определенных программ. Так, в рамках данной работы, применялся разработанный алгоритм:

а) потенциальный нарушитель запускает скрипт с указанием целевого IP-адреса;

б) скрипт запускает средство проведения аудита сети Nmap;

в) скрипт отправляет в LLM специальный формированный промпт и прикладывает отчет из Nmap;

г) LLM анализирует отчет о найденных сканером уязвимостях и формирует JSON-файл с описанием всех уязвимостей;

д) скрипт отправляет в LLM специальный промпт и прикладывает сформированный ранее JSON-файл,

е) LLM поэтапно генерирует эксплойты и отправляет их в скрипт;

ж) скрипт запускает поочередно эксплойты и проверяет было ли выполнено заявленное действие, если возникает ошибка отправляет в LLM ошибку с просьбой решить проблему, скрипт делает 5 попыток исправить эксплоит и идет дальше.

По итогу работы скрипта формируется отчет, включающий в себя перечень уязвимостей, результаты их эксплуатации, количество попыток и перечень возникших ошибок.

II. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ

В ходе серии экспериментов был проведен многократный запуск скрипта автоматического пентеста, направленного на целевую систему “Жанна”. В результате экспериментов можно сделать вывод, что LLM практически в 100% случаев успешно анализировала вывод Nmap, вследствие чего верно были определены уязвимости. Это позволяло предлагать различные варианты эксплуатации обнаруженных уязвимостей. На этом этапе формировался JSON-файл (Рис. 3), на основе которого дальше LLM осуществляет попытки пентеста путем генерации и запуска эксплойтов.


```
{
  "target_ip": "192.168.13.35",
  "timestamp": "20260210_113027",
  "missing_packages": [],
  "vulnerabilities": [
    {
      "id": "vsftpd_234_backdoor",
      "порт": "21",
      "протокол": "tcp",
      "сервис": "vsftpd",
      "версия": "2.3.4",
      "уязвимость": "Backdoor Command",
      "cve": "CVE-2011-2523",
      "описание": "Версия vsftpd 2.3.4 содержит backdoor, активируемый через команду :)",
      "серьезность": "высокая",
      "ожидаемый_результат": "Получение обратной оболочки",
      "тип_эксплоита": "remote",
      "требуемые_библиотеки": [
        "socket",
        "sys",
        "os"
      ],
      "рекомендации": "Использовать Metasploit модуль 'exploit/unix/ftp/vsftpd_234_backdoor'"
    },
```


Рис. 3. Пример описания одной уязвимости от LLM

Стоит отметить, что несмотря на успешное определение уязвимостей и описание методов их

эксплуатации, LLM удалось лишь в 0,11% случаев использовать уязвимость vsFTPd 2.3.4 (2 из 1755). Для остальных 16-ти обнаруженных уязвимостей ни один эксплоит не оказался рабочим. Таким образом результаты эксперимента показали, что:

1. Нейросеть генерировала код с синтаксическими ошибками.

Например, отчёт показал ошибку выполнения эксплоита: NameError: name 'name' is not defined. Нейросеть при генерации кода использовала “name” вместо “__name__” в строчке “if __name__ == "__main__":”.

2. Нейросеть допускала синтаксические ошибки при работе с бинарными payload.

Например, отчёт показал ошибку выполнения эксплоита: SyntaxError: unterminated string literal.

3. Неспособность нейросети исправить ошибки, связанные с отсутствием требуемых зависимостей.

Например, отчёт показал ошибку выполнения эксплоита: ModuleNotFoundError: No module named 'mysql'. Несмотря на то, что скрипт позволял выполнить команды “pip install

…” нейросеть при получении подобной ошибки принимала решение переписать код, используя другие зависимости.

4. Некорректные эксплоит-реализации.

Даже в случае синтаксически правильного и исполняемого кода, эксплоит не выполнял поставленных задач.

5. Оборванные строковые литералы.

Например, отчёт показал ошибку выполнения эксплоита: payload += b"\x00\x00\x00\x00\x00\x

6. Нейросеть успешно структурировала атаку.

Нейросеть смогла планировать вектор атаки. В ходе эксперимента она успешно определила имеющиеся уязвимости и предложила правильные методики их эксплуатации.

Полученные результаты демонстрируют, что LLM способна выявлять уязвимости и формировать вектор атаки, а в редких случаях эксплуатировать известные уязвимости путем генерации работающих эксплоитов.

Анализ неудачных случаев использования сгенерированных эксплоитов продемонстрировал, что ключевыми проблемами использования современных языковых моделей для осуществления автоматизированных атак стали синтаксические ошибки и неполные payload-строки при генерации эксплоита. Это согласуется с выводами в исследованиях проблем использования кода, сгенерированного LLM [12, 13]. Синтаксические ошибки в коде являются галлюцинацией LLM, обусловленные недостаточным количеством примеров корректной конструкции в обучающих данных для конкретной модели. Неудача с Samba (Рис. 4) обусловлена генерацией чрезмерно длинных payload-строк, которые обрезались на этапе записи в файл. Это может быть связано с ограничением длины вывода модели. В то же время положительный опыт обнаружения и описания известных уязвимостей показывает двойственную природу LLM.

```
Уязвимость: usermap_script() Function Vulnerability
Порт: 139/445
Сервис: Samba
Статус: НЕУДАЧА
Попыток: N/A
Ошибка:   File "/home/kali/ai/temp_test_3_20260210_113027.py", line 7
    payload = b'\x00\x00\x00\x90\xFF\x53\x4D\x42\x72\x00\x00\x00\x00\x
```

Рис. 4. Неудача с Samba

Анализ случаев успешной эксплуатации уязвимости показал, что атаки с простой и формализованной логикой иногда могут выполняться LLM автоматически. В частности, уязвимость FTP-сервиса vsftpd 2.3.4 предполагает отправку специально сформированной строки, активирующей бэкдор. Такая уязвимость не требует установки зависимостей и генерации сложного кода. Остальные же уязвимости были подробно описаны, в том числе методы их эксплуатации. Следует отметить, что в случае исправления синтаксических ошибок в коде или установки требуемых зависимостей некоторые эксплойты оказались рабочими, так, например, была проэксплуатирована уязвимость «Metasploitable root shell» (bindshell).

ЗАКЛЮЧЕНИЕ

Результаты исследования позволяют сделать выводы о том, что языковая модель DeepSeek-R1 имеет значительные недостатки в качестве генерируемого кода, но тем не менее, в редких случаях способна совершить атаку без вмешательства человека. Можно сделать вывод о потенциале использования подобных моделей для совершения кибератак как минимум в качестве консультанта для начинающих хакеров. Это подтверждается и тем, что, внеся незначительные изменения в нерабочие эксплоиты, их можно было использовать для эксплуатации уязвимости. В дальнейшем исследовании планируется сфокусироваться на анализе генерируемого кода и изучении возможности его исправления в случае необходимости.

В результате можно сказать, что это исследование, как и другие [14, 15] показали, что современные языковые модели, подобные DeepSeek-R1, имеют ограниченные возможности в задачах автоматизированного пентестинга. Однако, даже при столь низкой эффективности можно уверенно говорить о возможности массового использовании LLM для автоматизированных атак на системы с документированными и простыми в реализации уязвимостями.

ЛИТЕРАТУРА

[1] Williams, S. Cybersecurity teams are preparing for a surge in global CVE vulnerabilities in 2026 // SecurityBrief Asia. — 2026. — Режим доступа: https://securitybrief.asia/story/cybersecurity-teams-brace-for-surge-in-global-cves-in-2026.

[2] Singer, B., Lucas, K., Adiga, L., Jain, M., Bauer, L., Sekar, V. (2025). Incalmo: An Autonomous LLM-Based System for Red Teaming in Multi-Component Networks. arXiv.org. DOI: 10.1007/s10207-024-00868-2

[3] Mirsky, Y., Demontis, A., Kotak, J., Shankar, R., Gelei, D., Yang, L., Zhang, X., Pintor, M., Lee, W., Elovici, Y., Biggio, B. (2023). The Threat of Offensive Artificial Intelligence to Organizations. Computers & Security.: https://www.sciencedirect.com/science/article/abs/pii/S0167404822003984

[4] Gupta, M., Gupta, M. (2025). Measuring AI Cybersecurity: A Comprehensive Framework for Understanding Offensive and Adversarial AI. AI and Ethics, 5, 883–910. DOI: 10.1007/s43681-024-00427-4

[5] Hu, Z., Beuran, R., Tan, Y. (2020). Automated Penetration Testing Using Deep Reinforcement Learning. In: Proceedings of the 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Genoa, Italy, pp. 2–10. DOI: 10.1109/EuroSPW51379.2020.00010

[6] Nguyen, T. T., Reddi, V. J. (2021). Deep Reinforcement Learning for Cybersecurity. arXiv.org.: https://arxiv.org/abs/1906.05799

[7] CVE-2025-26210: Cross-Site Scripting (XSS) Vulnerability in DeepSeek AI Products (2025). Feedly.: https://feedly.com/cve/CVE-2025-26210

[8] Kelly, D. J., Long, J., Mylonas, A., Pitropakis, N., Buchanan, W. J. (2024). A Systematic Review of Research Using Artificial Intelligence for Open Source Intelligence (OSINT) Applications. International Journal of Information Security, 23, 2911–2938. DOI: 17487/RFC8377

[9] IBM aThink (2023). AI vs Human Deceit: Unravelling New-Age Phishing Tactics.: https://www.ibm.com/think/x-force/ai-vs-human-deceit-unravelling-new-age-phishing-tactics

[10] Инфобезопасность.ру (2025). ИИ вооружил киберпреступников — число атак в России выросло на треть.: https://infobezopasnost.ru/blog/news/ii-vooruzhil-kiberprestupnikov-chislo-atak-v-rossii-vyroslo-na-tret

[11] Как реклама с дипфейками в Instagram приводит к потере денег. (2025): https://www.kaspersky.ru/blog/scam-with-deepfakes-in-instagram-facebook-whatsapp/40221

[12] Zhang, Z., Wang, C., Wang, Y., Shi, E., Ma, Y., Zhong, W., Chen, J., Mao, M., Zheng, Z. (2025). LLM Hallucinations in Practical Code Generation: Phenomena, Mechanisms, and Mitigation. Proceedings of ISSTA 2025. DOI: 10.1145/3728894

[13] LLM they cannot cope with the search and exploitation of vulnerabilities // Infosecurity Magazine.: https://www.infosecurity-magazine.com/news/llms-fall-vulnerability-discovery

[14] Todd, D. (2025). DeepSeek and AI-Generated Malware Pose New Cybersecurity Threat. SecureWorld.: https://www.secureworld.io/industry-news/deepseek-ai-generated-malware

[15] Alger, J. (Managing Editor) (2025). DeepSeek Can Develop Malware: Cyber Experts Share the Risks. Security Magazine.: https://www.securitymagazine.com/articles/101470-deepseek-can-develop-malware-cyber-experts-are-sharing-the-risks

Информация об авторах

Азява Дмитрий Александрович, студент кафедры Защиты информации Новосибирского государственного технического университета, г. Новосибирск, Россия, e-mail: dimulatoraz@gmail.com, ORCID: 0009-0006-5107-9419.

Иллак Наталья Вячеславовна, студентка кафедры Защиты информации Новосибирского государственного технического университета, г. Новосибирск, Россия, e-mail: illak.nvs@gmail.com, ORCID: 0009-0003-1043-4611.

Киселев Антон Анатольевич, старший преподаватель кафедры Защиты информации Новосибирского государственного технического университета, г. Новосибирск, Россия, e-mail: anton.kiselev@corp.nstu.ru, ORCID: 0000-0002-0464-5039

Investigation of the possibility of conducting automated penetration testing with publicly available large language models

Dmitry Azyava, Natalia Illak, Anton Kiselev

Novosibirsk State Technical University, Novosibirsk, Russia

***Abstract* – The paper explores the possibility of using publicly available large language models to carry out cyber attacks by automated system penetration testing. The testing was conducted using the DeepSeek-R1 large language model. The experiment was conducted on a laboratory stand consisting of two virtual machines with operating systems: Kali Linux and Metasploitable2. The results showed the possibility of using a large language model for conducting a pentest, but with some limitations.**

***Keywords* – LLM, AI, offensive AI.**

REFERENCES

[1] Williams, S. Cybersecurity teams are preparing for a surge in global CVE vulnerabilities in 2026 // SecurityBrief Asia. — 2026. — Access mode: https://securitybrief.asia/story/cybersecurity-teams-brace-for-surge-in-global-cves-in-2026 (accessed: 01/13/2026).

[2] Singer, B., Lucas, K., Adiga, L., Jain, M., Bauer, L., Sekar, V. (2025). Incalmo: An Autonomous LLM-Based System for Red Teaming in Multi-Component Networks. arXiv.org. DOI: 10.1007/s10207-024-00868-2

[3] Mirsky, Y., Demontis, A., Kotak, J., Shankar, R., Gelei, D., Yang, L., Zhang, X., Pintor, M., Lee, W., Elovici, Y., Biggio, B. (2023). The Threat of Offensive Artificial Intelligence to Organizations. Computers & Security. Retrieved January 14, 2026, from https://www.sciencedirect.com/science/article/abs/pii/S0167404822003984

[4] Gupta, M., Gupta, M. (2025). Measuring AI Cybersecurity: A Comprehensive Framework for Understanding Offensive and Adversarial AI. AI and Ethics, 5, 883–910. DOI: 10.1007/s43681-024-00427-4

[5] Hu, Z., Beuran, R., Tan, Y. (2020). Automated Penetration Testing Using Deep Reinforcement Learning. In: Proceedings of the 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Genoa, Italy, pp. 2–10. DOI: 10.1109/EuroSPW51379.2020.00010

[6] Nguyen, T. T., Reddi, V. J. (2021). Deep Reinforcement Learning for Cybersecurity. arXiv.org. Retrieved January 15, 2026, from https://arxiv.org/abs/1906.05799

[7] CVE-2025-26210: Cross-Site Scripting (XSS) Vulnerability in DeepSeek AI Products (2025). Feedly. Retrieved January 15, 2026,

from https://feedly.com/cve/CVE-25-2621
[8] Kelly, D. J., Long, J., Mylonas, A., Pitropakis, N., Buchanan, W. J. (2024). A Systematic Review of Research Using Artificial Intelligence for Open Source Intelligence (OSINT) Applications. International Journal of Information Security, 23, 2911–2938. DOI: 17487/RFC8377
[9] IBM aThink (2023). AI vs Human Deceit: Unravelling New-Age Phishing Tactics. Retrieved January 17, 2026, from https://www.ibm.com/think/x-force/ai-vs-human-deceit-unravelling-new-age-phishing-tactics
[10] Infobezopasnost′.ru (2025). II vooruzhil kiberprestupnikov — chislo atak v Rossii vyroslo na tret′. Retrieved January 17, 2026, from https://infobezopasnost.ru/blog/news/ii-vooruzhil-kiberprestupnikov-chislo-atak-v-rossii-vyroslo-na-tret
[11] Kak reklama s dipfejkami v Instagram privodit k potere deneg. (2025). Retrieved January 17, 2026, from https://www.kaspersky.ru/blog/scam-with-deepfakes-in-instagram-facebook-whatsapp/40221
[12] Zhang, Z., Wang, C., Wang, Y., Shi, E., Ma, Y., Zhong, W., Chen, J., Mao, M., Zheng, Z. (2025). LLM Hallucinations in Practical Code Generation: Phenomena, Mechanisms, and Mitigation. Proceedings of ISSTA 2025. DOI: 10.1145/3728894
[13] LLM they cannot cope with the search and exploitation of vulnerabilities // Infosecurity Magazine. — 2025. — Access mode: https://www.infosecurity-magazine.com/news/llms-fall-vulnerability-discovery / (date of access: 03/09/2026).
[14] Todd, D. (2025). DeepSeek and AI-Generated Malware Pose New Cybersecurity Threat. SecureWorld. Retrieved February 3, 2026, from https://www.secureworld.io/industry-news/deepseek-ai-generated-malware
[15] Alger, J. (Managing Editor) (2025). DeepSeek Can Develop Malware: Cyber Experts Share the Risks. Security Magazine. Retrieved February 8, 2026, from https://www.securitymagazine.com/articles/101470-deepseek-can-develop-malware-cyber-experts-are-sharing-the-risks

Information about the authors

Dmitry A. Azyava, student of the Information Security Department of Novosibirsk State Technical University, Novosibirsk, Russia, e-mail: dimulatoraz@gmail.com, ORCID: 0009-0006-5107-9419.

Natalia V. Illak, student of the Information Security Department of Novosibirsk State Technical University, Novosibirsk, Russia, e-mail: illak.nvs@gmail.com, ORCID: 0009-0003-1043-4611.

Anton A. Kiselev, Senior Lecturer at the Information Security Department of Novosibirsk State Technical University, Novosibirsk, Russia, e-mail: anton.kiselev@corp.nstu.ru, ORCID: 0000-0002-0464-5039.