

недостающим психолого-поведенческим измерением. Без этого любая система экономической безопасности остается уязвимой перед иррациональной стихией панического поведения.

СПИСОК ЛИТЕРАТУРЫ

1 **Подлиняев, О. Л.** Психология толпы и специфика её разновидностей / О. Л. Подлиняев // Вестник Восточно-Сибирского института МВД России. – 2017. – № 2 (81). – URL: <https://cyberleninka.ru/article/n/psihologiya-tolpy-i-spetsifika-eyo-raznovidnostey> (дата обращения: 29.04.2026).

2 **Талыбов, А. М.** Толпа, причины возникновения, основные виды и характеристики / А. М. Талыбов // Достижения фундаментальной, клинической медицины и фармации : материалы 77-й науч. сессии сотрудников университета, Витебск, 26–27 янв. 2022 г. / редкол.: А. Т. Щастный (председатель) [и др.]. – Витебск : ВГМУ, 2022. – С. 238.

3 **Степанов, В. Л.** Премьер-министр В. С. Павлов и денежная реформа 1991 г.: оценки и суждения / В. Л. Степанов // Вопросы теоретической экономики. – 2022. – № 4. – URL: <https://cyberleninka.ru/article/n/premier-ministr-v-s-pavlov-i-denezhnaya-reforma-1991-g-otsenki-i-suzhdeniya> (дата обращения: 30.04.2026).

4 **Буряк, В. В.** Пандемия Covid-19, инфопаника и пост-вирусная экономика / В. В. Буряк // Бенефициар. – 2020. – № 77. – С. 35–39.

5 VIX Falls to Lowest Level in Nearly 4 Years – Is it Time to Buy the VIX? // Market Rebellion. – URL: <https://marketrebellion.com/news/trading-insights/vix-falls-to-lowest-level-in-nearly-4-years-is-it-time-to-buy-the-vix/> (date of access: 30.04.2026).

Получено 27.05.2026

ISSN 2227-1155. Сборник студенческих научных работ.

Вып. 31. Гомель, 2026

УДК 004.056+004.8

А. Е. КОЗЛОВСКИЙ, С. Н. ДОЛГОМЫСЛОВА (ЭМ-41)

Научные руководители: старшие преподаватели *С. В. КИСЕЛЕВА,*
В. Е. МИНИН

РАЗБОР УГРОЗ OWASP TOP 10 ДЛЯ LLM

Описываются актуальные угрозы из перечня OWASP Top 10 LLM 2025, анализируются механизмы их реализации, приводятся конкретные примеры атак и предлагаются практические меры противодействия.

Стремительное распространение приложений на основе больших языковых моделей (Large Language Models, LLM) порождает принципиально новый класс угроз информационной безопасности. В ответ на этот вызов некоммерческая организация OWASP (Open Web Application Security Project) сформировала рабочую группу, которая в 2023 году выпустила первую вер-

сию документа «OWASP Top 10 for Large Language Model Applications», а в ноябре 2024 года – обновлённую редакцию. В редакции 2025 года был пересмотрен список угроз: две категории были добавлены впервые (утечка системного промпта и уязвимости векторных представлений), пять категорий получили расширенные определения, а весь список был перестроен по приоритету с учётом реальной статистики инцидентов [2, 3]. Данный документ описывает десять критических угроз безопасности, специфичных для LLM-приложений, и служит ориентиром для разработчиков, архитекторов систем и специалистов по кибербезопасности [1].

LLM01: Prompt Injection (Инъекция промпта). Атака использует фундаментальную особенность языковых моделей: неспособность надёжно разграничить управляющие инструкции системы и пользовательские входные данные. Существуют два вида инъекций:

- прямая (Direct Injection) – злоумышленник напрямую вводит вредоносный промпт в интерфейс чата;
- косвенная (Indirect Injection) – вредоносные инструкции скрыты во внешних данных, обрабатываемых моделью (веб-страницах, документах, электронных письмах).

По данным исследования Proofpoint 2025 года, было задокументировано более 461 640 попыток внедрения промпта, при этом успешность атак составила от 50 до 84 % в зависимости от применяемой техники [4]. В июне 2025 года была раскрыта уязвимость EchoLeak (CVE-2025-32711) с оценкой CVSS 9.3 в Microsoft 365 Copilot: злоумышленник отправлял специально сформированное электронное письмо с внедрёнными инструкциями, и, когда получатель просил Copilot обобщить содержимое почтового ящика, ИИ-ассистент незаметно отправлял конфиденциальные документы на внешний сервер. Другой пример – атака через резюме: документ содержит скрытые инструкции, заставляющие LLM оценить кандидата положительно вне зависимости от реального содержимого резюме [5].

Меры защиты:

- строгая валидация входных данных;
- изоляция системных промптов от пользовательских данных;
- применение принципа наименьших привилегий для инструментов, доступных модели;
- мониторинг аномальных паттернов запросов;
- внедрение специализированных систем обнаружения инъекций.

LLM02: Sensitive Information Disclosure (Раскрытие конфиденциальной информации). Атака охватывает непреднамеренное раскрытие моделью персональных данных (PII), коммерческой тайны, медицинской информации, финансовых данных или деталей системных конфигураций. Причины включают запоминание обучающих данных, недостаточную фильтрацию выходных данных, а также использование контекстного окна для хранения конфиденциальных фрагментов информации из предыдущих диалогов.

Особую опасность представляет утечка через RAG-системы (Retrieval-Augmented Generation), где модель может воспроизвести конфиденциальные корпоративные документы из векторной базы данных [2].

Исследователи продемонстрировали, что, задавая повторяющиеся запросы к ChatGPT, можно вызвать воспроизведение фрагментов обучающих данных, включая адреса электронной почты, телефонные номера и другую личную информацию [6]. Также зафиксированы случаи, когда модели раскрывали внутреннюю логику приложений через выходные данные.

Меры защиты:

- минимизация персональных данных в обучающих наборах;
- применение дифференциальной приватности при обучении;
- внедрение фильтрации выходных данных (DLP – Data Loss Prevention);
- разграничение доступа к документам в RAG-системах через механизмы ACL;
- шифрование и изоляция данных по арендаторам в мультитенантных средах.

LLM03: Supply Chain Vulnerabilities (Уязвимости цепочки поставок).

LLM-приложения зависят от множества дополнительных факторов: предобученных моделей, датасетов, библиотек (LangChain, Hugging Face Transformers, векторных БД), плагинов и API внешних провайдеров. Компрометация любого звена цепочки может привести к внедрению бэкдоров, вредоносного кода или отравленных весов модели [3].

Пример атаки: скомпрометированный поставщик разместил вредоносный LoRA-адаптер на Hugging Face, который при слиянии с базовой моделью изменял её поведение в определённых контекстах. Также задокументированы случаи отравления публичных датасетов для внедрения бэкдоров, активируемых при специфических входных данных [7].

Меры защиты:

- верификация моделей и компонентов перед использованием (проверка цифровых подписей, хэш-сумм);
- формирование AI-BOM (AI Bill of Materials);
- применение принципов SLSA (Supply-chain Levels for Software Artifacts);
- использование только верифицированных репозиторийев;
- регулярное обновление зависимостей и тестирование на аномалии поведения.

LLM04: Data and Model Poisoning (Отравление данных и модели).

Атака предполагает намеренное внедрение вредоносных или предвзятых данных в обучающие наборы, датасеты дообучения или векторные базы данных. Цели атаки разнообразны: внедрение скрытого триггера, снижение точности модели, встраивание определённых предубеждений, создание уязвимости к конкретным запросам. Особую сложность представляет обнаружение данной угрозы, так как отравленная модель может демонстрировать нормальное поведение в большинстве случаев, активируя вредоносное по-

ведение лишь при наличии определённого триггера [1].

Задokumentированы случаи, когда злоумышленники отравляли публичные датасеты Википедии и Reddit, добавляя в них дезинформацию. В результате модели, дообученные на этих данных, начинали систематически воспроизводить ложную информацию по конкретным темам. В исследовании 2024 года было показано, что атаки на цепочку поставок LLM-агентов позволяют обойти механизмы безопасности в 11,6–33,5 % случаев [8].

Меры защиты:

- верификация происхождения обучающих данных (data provenance);
- применение методов обнаружения аномалий и тестирования на устойчивость (adversarial robustness testing);
- использование федеративного обучения для минимизации централизованных рисков;
- непрерывный мониторинг поведения модели в продуктивной среде;
- регулярное проведение red-team тестирования.

LLM05: Improper Output Handling (Ненадлежащая обработка выходных данных). Угроза возникает, когда ответы языковой модели передаются в нижестоящие системы без должной валидации, очищения или кодирования. Поскольку LLM может генерировать произвольный текст, включая фрагменты кода, SQL-запросы, HTML и JavaScript, прямая передача таких выходных данных исполняющим системам открывает классические векторы атак: межсайтовый скриптинг (XSS), SQL-инъекции, внедрение команд операционной системы, подделка межсайтовых запросов (CSRF). Принципиальное отличие от традиционных уязвимостей состоит в том, что источником вредоносного содержимого является сама модель [2].

В качестве примера можно предложить XSS-атаку через LLM-чатбот: злоумышленник просит модель сгенерировать ответ, содержащий JavaScript-код, который затем отображается в браузере другого пользователя без санитизации, или предпринять SQL-инъекцию, т. е. сгенерированный моделью SQL-запрос выполняется без параметризации, что позволяет злоумышленнику получить несанкционированный доступ к базе данных [3].

Меры защиты:

- обязательная санитизация и валидация всех выходных данных модели перед их использованием в нижестоящих системах;
- применение параметризованных запросов при работе с базами данных;
- внедрение политик Content Security Policy (CSP) в веб-приложениях;
- применение принципа нулевого доверия к выходным данным модели;
- использование sandbox-окружений для выполнения LLM-сгенерированного кода.

LLM06: Excessive Agency (Избыточная автономность модели). Угроза проявляется, когда LLM-системы наделяются расширенными полномочиями и способностью выполнять действия с реальными последствиями – отправлять электронные письма, выполнять код, управлять файлами, со-

вершать транзакции – без надлежащего контроля со стороны человека. В сочетании с уязвимостями инъекции промпта (особенно косвенной) это создаёт критический риск: злоумышленник может заставить модель выполнить произвольные деструктивные действия от имени пользователя или системы. Особую опасность представляют многоагентные системы (multi-agent), где модели взаимодействуют автономно [1].

Пример демонстрационной атаки: веб-страница содержит скрытые инструкции, предписывающие LLM-ассистенту, суммирующему контент, удалить все письма пользователя с помощью подключённого плагина почты. При посещении страницы пользователем ассистент выполняет деструктивное действие без явного разрешения [5].

Меры защиты:

- применение принципа минимально необходимых привилегий для всех инструментов LLM;

- разделение функций принятия решений и их исполнения;

- внедрение обязательного подтверждения человеком для критических действий (human-in-the-loop);

- ограничение числа и типов плагинов и инструментов;

- непрерывная оценка поведения агентов;

- применение «предохранителей» (circuit breakers) для блокировки нежелательных цепочек действий.

LLM07: System Prompt Leakage (Утечка системного промпта). Системный промпт – это скрытые инструкции, задающие роль, поведение, ограничения и бизнес-логику LLM-приложения. Эта уязвимость представляет серьёзную угрозу для коммерческих продуктов: раскрытые инструкции позволяют злоумышленнику понять внутреннюю логику приложения, обойти защитные механизмы, скопировать проприетарные алгоритмы или подготовить более точно направленные атаки. Кроме того, в современных многоагентных системах с длинным контекстом и памятью риск утечки существенно возрастает [6].

В качестве примера атаки через повторение можно привести следующую ситуацию: злоумышленник просит модель повторить все инструкции, которые та получила до начала разговора. Менее защищённые модели воспроизводят системный промпт дословно. Другой вектор – многоходовые атаки: серия косвенных вопросов, постепенно реконструирующих содержание системного промпта по фрагментам ответов модели [6].

Меры защиты:

- строгая фильтрация выходных данных на наличие фрагментов системного промпта;

- архитектурное разделение системных инструкций и данных;

- применение техник prompt hardening;

- регулярное red-team тестирование на устойчивость к экстракции промпта;

- хранение наиболее критичной бизнес-логики в коде приложения, а не в

промпте.

LLM08: Vector and Embedding Weaknesses (Уязвимости векторных представлений). Системы Retrieval-Augmented Generation (RAG) используют векторные базы данных для хранения и поиска релевантных фрагментов информации, которые затем передаются в контекст модели. Эта архитектура вносит новый класс уязвимостей: инверсия эмбеддингов позволяет восстановить оригинальные тексты из векторных представлений; отравление векторной базы данных даёт возможность манипулировать выводами RAG-системы; межконтекстные утечки в неправильно настроенных базах открывают доступ к данным других пользователей или арендаторов [1].

Пример атаки – атака Embedding Inversion: злоумышленник, получив несанкционированный доступ к векторной базе данных, применяет специализированные методы для восстановления исходных текстов из эмбеддингов, раскрывая конфиденциальное содержимое корпоративных документов. В мульти-тенантных RAG-системах некорректная конфигурация метаданных может позволить запросам одного пользователя извлекать документы другого [2].

Меры защиты:

- шифрование хранимых эмбеддингов;
- строгое разграничение доступа к векторным коллекциям через механизмы ACL и метаданные арендатора;
- регулярный аудит настроек векторных баз данных;
- применение методов дифференциальной приватности для защиты эмбеддингов;
- валидация метаданных при каждом запросе к RAG-системе.

LLM09: Misinformation (Дезинформация). Дезинформация как угроза связана со склонностью LLM генерировать уверенно звучащие, но фактически некорректные ответы – так называемые «галлюцинации». В отличие от прочих угроз, данная уязвимость коренится не во вредоносных действиях злоумышленника, а в фундаментальных ограничениях архитектуры трансформеров: модели оптимизированы для правдоподобия текста, а не для фактической точности. Особую опасность это представляет в высокритичных областях: медицина, юриспруденция, финансовые консультации, инженерные расчёты. Чрезмерное доверие к выходным данным LLM без их верификации может привести к серьёзным практическим последствиям [3].

Примеры атак: медицинские LLM-ассистенты, уверенно рекомендуящие несуществующие лекарства или неверные дозировки; юридические чат-боты, ссылающиеся на несуществующую судебную практику; инженерные системы, генерирующие небезопасный код или ошибочные технические расчёты. В 2024 году задокументированы случаи использования LLM-галлюцинаций в атаках на цепочку поставок: злоумышленники публиковали в репозиториях несуществующие библиотеки, которые модели рекомендовали разработчикам [9].

Меры защиты:

- заземление ответов на верифицированных источниках через RAG;
- обязательная верификация критически важных ответов экспертами-людьми;
- отображение источников и степени уверенности модели пользователю;
- непрерывное тестирование на галлюцинации (LLM evaluation);
- чёткое разграничение применений LLM по уровню критичности;
- установка заградительных механизмов для высокорисковых доменов.

LLM10: Unbounded Consumption (Неконтролируемое потребление ресурсов). Угроза объединяет класс атак, при которых отсутствие ограничений приводит к отказу в обслуживании (DoS), финансовым потерям, деградации сервиса или краже интеллектуальной собственности. Высокие вычислительные требования LLM, особенно в облачной среде с моделью оплаты за использование (pay-per-use), делают их уязвимыми к специфическим атакам типа «Denial of Wallet» – злоумышленник генерирует огромные счета за использование API-сервиса жертвы. Также возможно функциональное клонирование модели путём систематического опроса API и создания синтетического обучающего датасета [4].

Примеры атак: атака Variable-Length Input Flood, при которой злоумышленник отправляет тысячи запросов переменной длины, исчерпывая вычислительные ресурсы сервиса. При атаке Resource-Intensive Queries специально сконструированные запросы запускают наиболее затратные в вычислительном отношении процессы модели, приводя к длительной загрузке CPU и потенциальному отказу системы. В облачных сервисах подобные атаки могут приводить к счетам на десятки тысяч долларов [4].

Меры защиты:

- введение строгих ограничений на количество токенов и запросов (rate limiting) на уровне API;
- предварительная оценка стоимости запроса перед его выполнением;
- кэширование ответов на повторные запросы;
- мониторинг стоимости инференса в реальном времени с автоматическими оповещениями;
- ограничение максимальной длины входного контекста.

СПИСОК ЛИТЕРАТУРЫ

- 1 OWASP Top 10 for LLM Applications 2025 // OWASP Foundation. – URL: <https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/> (date of access: 04.05.2026).
- 2 The 2025 OWASP Top 10 for LLMs // TrojAI Blog. – URL: <https://troj.ai/blog/the-2025-owasp-top-10-for-llms> (date of access: 04.05.2026).
- 3 What are the OWASP Top 10 risks for LLMs? // Cloudflare Learning. – URL: <https://www.cloudflare.com/learning/ai/owasp-top-10-risks-for-llms/> (date of access: 04.05.2026).
- 4 Prompt injection: the OWASP #1 AI threat in 2026 // Securance. – URL:

<https://www.securance.com/blog/prompt-injection-the-owasp-1-ai-threat-in-2026/> (date of access: 04.05.2026).

5 LLM01:2025 Prompt Injection // OWASP GenAI Security Project. – URL: <https://genai.owasp.org/llmrisk/llm01-prompt-injection/> (date of access: 04.05.2026).

6 OWASP LLM07:2025 System Prompt Leakage // Indusface. – URL: <https://www.indusface.com/learning/owasp-llm-system-prompt-leakage/> (date of access: 04.05.2026).

7 LLM03:2025 Supply Chain // OWASP GenAI Security Project. – URL: <https://genai.owasp.org/llmrisk/llm03-training-data-poisoning/> (date of access: 04.05.2026).

8 Supply-Chain Poisoning Attacks Against LLM Coding Agent Skill Ecosystems / Y. Qu, Y. Liu, T. Geng [et al.] // arXiv:2604.03081. – URL: <https://arxiv.org/abs/2604.03081> (date of access: 04.05.2026).

9 OWASP Top 10 for LLMs 2025: Key Risks and Mitigation Strategies // Invicti. – URL: <https://www.invicti.com/blog/web-security/owasp-top-10-risks-llm-security-2025> (date of access: 04.05.2026).

10 LLM10:2025 Unbounded Consumption // OWASP GenAI Security Project. – URL: <https://genai.owasp.org/llmrisk/llm102025-unbounded-consumption/> (date of access: 04.05.2026).

11 Anatomy of an Attack – From Poisoned RAG to Hijacked MCP Agents // DeconvoluteAI. – URL: <https://deconvoluteai.com/blog/rag-malicious-chunk-demo> (date of access: 04.05.2026).

12 OWASP LLM Top 10 Vulnerabilities 2025: AI Security Risks // Deepstrike.io. – URL: <https://deepstrike.io/blog/owasp-llm-top-10-vulnerabilities-2025> (date of access: 04.05.2026).

Получено 11.06.2026

ISSN 2227-1155. Сборник студенческих научных работ.

Вып. 31. Гомель, 2026

УДК 656.213.02

Д. А. КОЗЛОВСКИЙ (УД-21)

Научный руководитель – ст. преп. *Г. В. ЧИГРАЙ*

ОСОБЕННОСТИ ТЕХНОЛОГИИ ПЕРЕВОЗОЧНОГО ПРОЦЕССА ПРИ ВЗАИМОДЕЙСТВИИ ЖЕЛЕЗНОДОРОЖНОЙ СТАНЦИИ И ГРУЗОВОГО ПОРТА: ПОИСК ОПТИМАЛЬНЫХ УСЛОВИЙ

Выявлены особенности технологии перевозочного процесса на стыке станции и порта и определены пути поиска оптимальных условий для оптимизации их работы.

Железнодорожный и водный транспорт является ключевым элементом мировой логистической системы. Узким местом, определяющим эффективность всей интермодальной цепи, часто становится стык «железнодорожная станция – грузовой порт». Несогласованность технологических процессов,