УДК 004.891.3

ВНЕДРЕНИЕ МЕТОДИК ОБЕСПЕЧЕНИЯ БЕЗОПАНОСТИ ПРИ РАЗРАБОТКЕ СИСТЕМ ИССКУСТВЕННОГО ИНТЕЛЕКТА

Билаш Н.В. (УрФУ), **Кубасова М.С.** (УрФУ), **Макутенус А.И.** (УрФУ)
Научный руководитель – Ассистент Кусайкин Е. В.
(УрФУ)

Введение.

При использовании поисковых LLM-агентов ответы формируются исходя из модели, обученной на большом количестве данных, и контекста, выдаваемого RAG-системой. Для повышения безопасности и исключения рисков утечек конфиденциальной информации через AI-ассистентов предлагается практика внедрения методик безопасности. Работа с LLM-агентами предполагает обработку больших объёмов данных, в том числе потенциально содержащих персональные данные (ПДн). В Федеральном законе №152-ФЗ «О персональных данных» отмечается обязанность оператора в принятии мер по обеспечению безопасности персональных данных при их обработке. Постановление Правительства РФ №1119 и Приказ ФСТЭК России №21 устанавливают требования к определению необходимости применения мер защиты в зависимости от значимости, их состав и содержание. Сама LLM-система, содержащая персональные данные, в соответствии с п.10 Статьи 3 Федерального закона №152-ФЗ «О персональных данных» функционирует как информационная система персональных данных (ИСПДн), а значит попадает под действие вышеуказанных нормативно-правовых актов. На территории ЕС основным документом, регулирующим обработку персональных данных, является General Data Protection Regulation (GDPR). Данный стандарт устанавливает строгие требования к обработке персональных данных, которые напрямую применимы к LLM-системам, особенно использующим механизмы RAG, долговременное хранение контекста и логирование запросов. Отдельно стоит выделить принцип privacy by design - защита персональных данных должна быть заложена в архитектуру LLM-системы на этапе проектирования.

Основная часть.

Работа направлена на внедрение методик безопасной разработки систем, содержащих LLM-агенты. В качестве решения предлагается разработанный прототип LLM-агента, модель для которого была взята с ресурса [1], уязвимости при обучении которой не учитываются при анализе на проникновение. Проверялась возможность сопротивляться промт-иньекциям, где лучшую сопротивляемость показал метод полного отделения системных данных от человеческого ввода. Как дополнительный метод (или способ) защиты был опробован метод внедрение еще одного LLM-агента, модерирующего ответы основного. Были рассмотрены риски, возникающие при построении RAG-системы. Для предотвращения атаки вида Data Poisoning построена политика защиты хранилища контента. В данную политику входят правила ограничения прав доступа, а также использование сервисных учетных записей. Внедрение политик контроля доступа на уровне поискового запроса. Также были разработаны и применены правила безопасного развертывания системы, путем создания цепочки CI/CD пайплайна с проведением тестов на безопасность.

Выводы.

Ожидается повышение уровня информационной безопасности, закрытие уязвимостей, свойственных LLM-агентам. Ориентиром для проверки служат выявленные OWASP TOP 10 [2] уязвимости для больших языковых моделей. Достижимость целей предлагается подтверждать проведением двух аудитов на проникновение до/после внедрения средств безопасности, сравнивая полученные результаты.

Список использованных источников:

1. AlicanKiraz0/Seneca-Cybersecurity-LLM-Q4_K_M-GGUF сайт. URL: https://huggingface.co/AlicanKiraz0/Seneca-Cybersecurity-LLM-Q4_K_M-GGUF (дата обращения 20.10.2025);

2. OWASP Top 10 for LLM Applications 2025 сайт URL: https://genai.owasp.org/llm-top-10/ (дата обращения 26.01.2026);