УДК 004.056

РАЗРАБОТКА МЕТОДИКИ ВЫЯВЛЕНИЯ ДЕФЕКТОВ БЕЗОПАСНОСТИ ПО ОТКРЫТЫМ ИСТОЧНИКАМ С ИСПОЛЬЗОВАНИЕМ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Гуреев В.А. (ИТМО)

Научный руководитель – кандидат технических наук, доцент Югансон А.Н. (ИТМО)

Введение. Современное развитие технологий искусственного интеллекта и увеличение числа кибератак обуславливают необходимость совершенствования методов сбора и анализа данных об инфраструктуре организаций. Разведка на основе открытых источников является ключевым этапом в оценке защищенности. Однако традиционные инструменты OSINT часто работают разрозненно и не позволяют комплексно оценить риски, связанные не только с инфраструктурой, но и с самой моделью ИИ. Целью данной работы является разработка методики OSINT с использованием специализированной большой языковой модели для повышения эффективности выявления дефектов безопасности информационных систем и моделей ИИ.

Основная часть. Существующие подходы к OSINT анализу доменных имен, как правило, ограничены использованием отдельных инструментов: поиск WHOIS, DNS-запросы, проверка SSL-сертификатов. Это позволяет обнаружить лишь ограниченный набор уязвимостей инфраструктуры, не затрагивая специфические угрозы для систем ИИ. Для решения данной проблемы разработана методика, центральным элементом которой является специализированная LLM – HackerAI. Модель дообучена на данных по кибербезопасности и способна не только генерировать экспертные заключения, но и использовать инструменты для комплексного сбора информации. Для оценки эффективности разработанной методики был проведен эксперимент на примере домена chatgpt.com. С помощью HackerAI был выполнен сбор информации по заданным промптам. В результате было идентифицировано 38 различных дефектов безопасности. Среди них: утечки учетных данных (CWE-359, LLM02:2025), проблемы конфигурации (отсутствие CSP – CWE-693), уязвимости к инъекциям (CVE-2024-27564, LLM01:2025), а также специфические атаки на модель (DAN-промпты – AML.T0051). Параллельно был проведен анализ того же домена с использованием традиционной методики, основанной на применении стандартных инструментов. Результатом применения традиционного подхода стало выявление лишь 5 дефектов безопасности, касающихся в основном базовой информации о регистрации домена и его IP-адресах. Существенные уязвимости, связанные с утечками данных API, уязвимостями модели и специфическими LLM-атаками, обнаружены не были.

Выводы. В результате исследования разработана методика OSINT, основанная на применении специализированной LLM HackerAI. Экспериментально подтверждено, что предложенный подход позволяет автоматизировать и углубить процесс сбора разведывательной информации, выявляя на порядок больше потенциальных дефектов безопасности (38 против 5), включая уязвимости как инфраструктуры, так и самих моделей ИИ. Практическое использование результатов позволяет специалистам по безопасности своевременно реагировать на выявленные угрозы, усиливать защиту и проводить мониторинг информационной структуры предприятия.

Список использованных источников:

1. Mukhopadhyay A., Luther K. OSINT Clinic: Co-designing AI-augmented collaborative OSINT investigations for vulnerability assessment //Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. – 2025. – С. 1-22.

2. Deng G. et al. {PentestGPT}: Evaluating and harnessing large language models for automated penetration testing //33rd USENIX Security Symposium (USENIX Security 24). – 2024. – С. 847-864.
3. Лемайкина С.В. Новый арсенал osint в цифровом мире. Разведка по открытым источникам: научная статья / Лемайкина С. В.; Орловский юридический институт Министерства внутренних дел Российской Федерации имени В.В. Лукьянова. – Орел, 2021 – С. 27-31.