УДК 004.056

Анализ технологий составления ландшафта тестирования безопасности информационной инфраструктуры для автономных средств на основе ИИ-агентов

Шерягин М.А. (ИТМО)
Научный руководитель – аспирант, научный сотрудник Еритенко Н.А. (ИТМО)

Введение. Рост масштаба и сложности корпоративных инфраструктур, а также интенсивное использование распределенных доменных сред и облачных сервисов приводят к тому, что классические подходы к автоматизации оценки защищенности оказываются недостаточно эффективными, прежде всего из-за высокой энтропии входных данных для систем на основе LLM-агентов. В существующих работах по автоматизации пентеста на базе LLM отмечается, что модели хорошо справляются с локальными подзадачами – запуском инструментов, интерпретацией их вывода и предложением следующих шагов, но с трудом удерживают целостное представление о состоянии тестируемой инфраструктуры и истории своих действий. Это приводит к избыточному дублированию контекста, необходимости пересылки длинных логов и полных выводов инструментов, а также к росту вероятности ошибок планирования и неверной приоритизации целей. В результате значительная часть вычислительного бюджета LLM-агента тратится на обработку нерелевантной или слабо структурированной информации, а не на собственно принятие решений. В данном исследовании проводится анализ технологий составления ландшафта тестирования безопасности информационной инфраструктуры, который сможет служить как универсальная среда для наблюдений агента и в теории позволит снизить энтропию входных данных для ИИ-агента, предоставляя ему структурированный и релевантный контекст.

Основная часть.

1) Анализ существующих методов автоматизации пентеста на базе LLM. Рассмотрены подходы типа PentestGPT и Pentagi, их модульные и мультиагентные архитектуры, выявлены ограничения в обработке "плоского" контекста (длинные текстовые истории, логи инструментов), приводящие к проблемам масштабирования на крупные сети.
2) Теоретические основы многослойной модели инфраструктуры. Описаны свойства модели: стационарность, частичная видимость, динамичность. Обсуждены форматы представления – графовые и сегментированные, с опорой на графы атак для моделирования состояний, уязвимостей и цепочек.
3) Разработка архитектуры и свойств модели. Представлена динамическая структура с уровнями (доменная инфраструктура, сетевые объекты, сервисы/уязвимости), фокусировкой на подграфах фиксированной глубины, функцией хранилища знаний и приоритизацией целей.
4) Масштабирование и интерфейс LLM-агента. Описаны механизмы сегментации графов, иерархической обработки, поддержка kill chain-ов. Предложен интерфейс с высокоуровневыми операциями, обеспечивающий перенос планирования на задачи статических алгоритмов.

Выводы. Проанализированные технологий составления ландшафта тестирования безопасности информационной инфраструктуры складываются в многослойную динамическую модель, которая позволяет существенного снизить энтропию контекста LLM-агентов в автоматизированном пентесте. Она обеспечивает структурированный релевантный контекст, фокусировку на локальных подграфах, долговременное хранение знаний и приоритизацию цепочек атак без дублирования данных. Результаты подтверждают перспективность подхода для повышения эффективности и масштабируемости тестирования крупных сетевых инфраструктур.

Список использованных источников:

1. Konsta A. M. et al. Survey: automatic generation of attack trees and attack graphs //Computers & Security. – 2024. – T. 137. – C. 103602.
2. Ingols K., Lippmann R., Piwowarski K. Practical attack graph generation for network defense //2006 22nd Annual computer security applications conference (ACSAC'06). – IEEE, 2006. – C. 121-130.
3. Jin X. et al. Prometheus: Infrastructure security posture analysis with ai-generated attack graphs //arXiv preprint arXiv:2312.13119. – 2023.
4. Tufano, M., Agarwal, A., Jang, J., Moghaddam, R. Z., & Sundaresan, N. (2024). AutoDev: Automated AI-driven development. //arXiv preprint arXiv:2403.08299. – 2024.