УДК 004.825

АВТОМАТИЗИРОВАННОЕ СОЗДАНИЕ СИГНАТУР ДЛЯ ЯЗЫКОВЫХ МОДЕЛЕЙ

Тарадайко Е.А.[1]
Научный руководитель –доцент Кармановский Н.С.[1]
[1]Университет ИТМО
evgeniataradaiko@yandex.ru

Введение

Моделирование угроз для систем искусственного интеллекта является актуальной задачей в настоящее время – время активного внедрения данной технологии во все сферы жизни человека, в том числе и в большие компании, обрабатывающие большое количество пользовательских данных. Процесс моделирования угроз является неотъемлемой частью организации процессов в компаниях. Еженедельное внедрение новых ИТ-технологий увеличивает поверхность атаки на компанию и количество актуальных угроз, отказ от внедрения нарастит технологическое отставание.

В настоящее время при моделировании угроз специалисты опираются на уже существующие открытые базы данных, такие как БДУ ФСТЭК [1], OWASP [2], MITRE [3] и NIST [4]. Данные базы данных упрощают процесс моделирования угроз, но они содержат не полный перечень всех угроз и формируются с некоторым запозданием. Наиболее свежие угрозы достаточно часто представлены в не структурированном и не обобщенном виде.

Автоматизация процесса структурирования позволит уменьшить задержку обновления и актуализации информации об угрозах.

Основная часть

Предлагается следующее решение: автоматизированное создание сигнатурной базы данных об атаках и уязвимостях на системы искусственного интеллекта с использованием языковых моделей.

В качестве входных данных для осуществления автоматизированного создания сигнатурной базы данных использовалось:

- база данных, состоящая из файлов формата .md, которые представляют собой выгрузку информации из открытых источников и баз данных об атаках и уязвимостях на системы искусственного интеллекта, а также методах митигации реализации атак и возможных злоумышленниках. База включает в себя как структурированную информацию из официальных источников баз данных регуляторов, исследователей и правительства, а также информацию из иных неструктурированный интернет-источников, таких как статьи, блоги и форумы, в которых описываются не внесенные методы атак или способы их митигации в базы данных официальных источников;
- шаблон формализации обобщения информации и представления угроз в виде сигнатур;
- системный промпт получения сигнатурного описания и перечня сущностей;
- системный промпт оценки сигнатурных описаний уязвимостей.

Получаемый результат:

- набор структурированно описанных угроз;
- граф взаимосвязей сущностей, встречаемых в угрозах.

Описание исследования:

Архитектура системы состоит из 6 моделей исполнителей (deepseek-r1-0528, gpt-oss-20b, gemini-2.5-flash, gemini-2.5-pro, mistral-large-latest, claude-3-7-sonnet-latest) и 1 модели – судья (GPT-4.1).

Исследование состоит из двух этапов:

1. Сформированный промпт, шаблон формализации и неструктурированная база данных передаются в 6 языковых моделей – исполнителей. На основе полученных ответов создается база данных сигнатурных описаний уязвимостей, которая включает в себя по 6 вариантов сигнатур для каждой уязвимости.

2. Оценка моделью-судьей полученных сигнатурных описаний на основе системного промпта, который позволяет провести:

- оценку переноса сущностей;
- оценку полноты;
- оценку информативности, с использованием адаптированной шкалы Лайкерта.

Исследование позволило сформировать требования к формату и содержанию модели угроз, обобщить структурные и информационные требования к сигнатурам возможных атак на системы искусственного интеллекта, а также оценить модели по критерию качества обобщения специализированной информации и скорости обработки.

Выводы

Предлагаемое решение может быть использовано для создания агента по автоматизированному анализу и формированию отчетов по процессу моделирования угроз на системы искусственного интеллекта.

Литература

1. Банк данных угроз безопасности [Электронный ресурс]. – Режим доступа: https://bdu.fstec.ru/threat (Дата обращения: 01.02.2026).
2. OWASP TOP 10 LLM [Электронный ресурс]. – Режим доступа: https://owasp.org/www-project-top-10-for-large-language-model-applications/ (Дата обращения: 01.02.2026).
3. MITRE ATT&CK [Электронный ресурс] – Режим доступа: https://attack.mitre.org (Дата обращения: 02.02.2026).
4. AI resk management framework[Электронный ресурс] – Режим доступа: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf (Дата обращения: 10.02.2026).