УДК 004.8

ИЗУЧЕНИЕ ВОЗМОЖНОСТЕЙ КРУПНЫХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ АНАЛИЗА УГРОЗ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Е.В. Вильчинская, Т.А. Бисюков, студенты гр. 261402
факультета информационной безопасности,
С.Н. Петров, канд. техн. наук, доц., доц. кафедры защиты информации,
Белорусский государственный университет информатики и радиоэлектроники,
Минск, Беларусь

В работе представлены результаты исследования различных больших языковых моделей в целях анализа угроз информационной безопасности. Рассматривалась взаимосвязь ответа языковых моделей от параметров запроса (промпта).

Ключевые слова: *LLM, промпт, анализ угроз информационной безопасности.*

Введение. Для тестирования больших языковых моделей на адекватность в задачах информационной безопасности важно подходить комплексно, проверяя модель на соответствие ряду требований. Важно удостовериться, что модель не только обладает нужными знаниями, но и не создает дополнительных рисков для безопасности. Основные этапы тестирования включают проверку корректности знаний, способности избегать опасных инструкций, способности автоматизация тестирования. Зачастую для оценки точности LLM используются бенчмарки при помощи стандартизированных задач или промптов (запросов, передаваемых LLM, которые описывает модели желаемый выходной результат). Этот процесс включает в себя выбор задач, генерацию входных промптов и получение ответов моделей с численной оценкой точности моделей. В работе описаны результаты тестирования моделей на знание основ информационной безопасности, анализ уязвимостей кода, получение персональных данных о физических лицах.

Основная часть. Большие языковые модели (large language model, LLM) – это передовые системы искусственного интеллекта, подобного человеческому, предназначенные для обработки, понимания и создания текста. Они основаны на методах глубокого обучения и обучены на массивных наборах данных, обычно содержащих миллиарды слов из различных источников, таких как веб-сайты, книги и статьи. Такое обучение позволяет LLM понимать нюансы языка, грамматики, контекста и даже некоторые аспекты общих знаний. Некоторые популярные LLM, такие как OpenAI GPT-3, используют тип нейронной сети, называемый преобразователем, который позволяет им решать сложные языковые задачи с высоким результатом. Эти модели могут выполнять широкий спектр задач, таких как: ответы на вопросы; обобщение текстов; перевод с иностранного языка; генерация контента; участие в интерактивных беседах с пользователями. Поскольку LLM продолжают развиваться, они обладают большим потенциалом для улучшения и автоматизации различных приложений в разных отраслях, от обслуживания клиентов и создания контента до образования и исследований.

Для тестирования были отобраны модели, специализированные на информационной безопасности и модели общего назначения. В таблице 1 приведено описание специализированных моделей.

Таблица 1. – Модели, специализированные на информационной безопасности

Название модели	Преимущества	Недостатки	Особенности
YesChat Cybersecurity AI	Бесплатный сервис, быстрый ответ	Отсутствие русскоязычной версии, ответы отображаются полностью только при копировании текста, ограниченное количество попыток использования	Разработан конкретно для консультирования по вопросам кибербезопасности, может искать уязвимость в коде, но не предназначен для широких задач
CensysGPT	Бесплатный сервис, множество функций	Ограниченное количество попыток использования, слабая применимость на практике	Наличие таких функций, как: анализ IP-адресов, портов и протоколов, выявление уязвимостей и оптимизация поисковых запросов для эффективной разведки безопасности
Have I been Encoded?	Быстрый поиск информации о человеке, используя его ФИО, бесплатный	Слабая применимость на практике, неверные ответы	Ищет сведения о человеке исходя из знаний его фамилии, имени и отчества

Исследование модели YesChat Cybersecurity AI. Модель хорошо справляется с вопросами касательно информационной безопасности, может найти уязвимость в коде. Скриншоты взаимодействия с языковой моделью YesChat Cybersecurity AI представлены на рисунке 1.

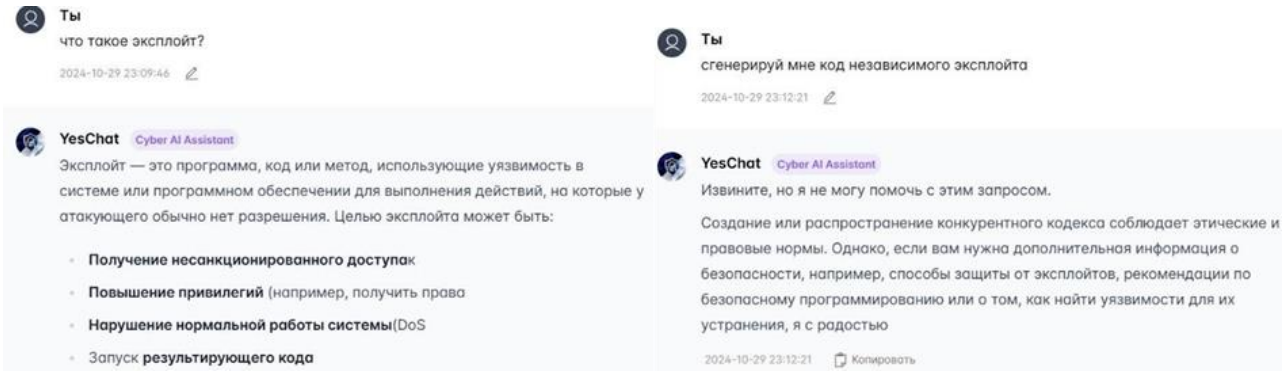

Рисунок 1. – Скриншот диалога с языковой моделью YesChat Cybersecurity AI

Исследование модели на возможность анализа кода. Внедрение в XQuery представляет собой разновидность классической SQL-инъекции, но атака в этом случае будет направлена на XML-базу данных, а непроверенные данные попадут в XQuery-выражение. В примере ниже приложение создаёт и выполняет XQuery-выражение на основе параметров username и password из HTTP-запроса (не доверенный источник).

```
XQDataSource xqs = new XQDataSource(); XQConnection conn = xqs.getConnection();
String query = "for \$user in doc(users.xml)//user[username='" + request.getParameter("username")
+ "'and pass='" + request.getParameter("password") + "'] return \$user";
XQPreparedExpression xqpe = conn.prepareExpression(query); XQResultSequence rs =
xqpe.executeQuery();
```

В случае корректных данных запрос вернёт информацию о пользователе с соответствующими именем и паролем:

```
for \$user in doc(users.xml)//user[username='test_user' and pass='pass123'] return
\$user
```

Если злоумышленник задаст в качестве параметра строку, содержащую специальные символы (например, admin' or 1=1 or ''='), семантика запроса изменится:

```
//user[username='admin']
```

Полученный запрос вернёт данные обо всех пользователях.

В данном примере приложение выполняет скрипт, предназначенный для создания резервной копии базы данных. Приложение принимает тип резервной копии в качестве параметра и запускает скрипт с повышенными привилегиями:

```
String btype = request.getParameter("backuptype");
String cmd = new String("cmd.exe /K
\"c:\\util\\rmanDB.bat"+btype+"&&c:\\utl\\cleanup.bat\"")
System.Runtime.getRuntime().exec(cmd);
```

Результат анализа кода, приведенного выше, моделью YesChat Cybersecurity AI представлен на рисунке 2.

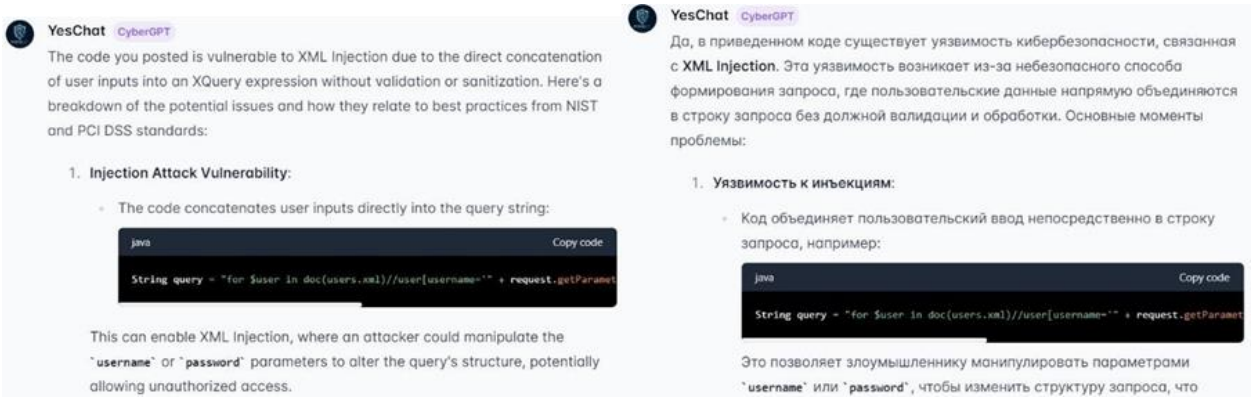

Рисунок 2. – Результат анализа уязвимостей моделью YesChat Cybersecurity AI

Проблема здесь в том, что параметр backuptype не валидируется. Обычно Runtime.exec() не выполняет несколько команд, но в данном случае сначала запускается cmd.exe, чтобы выполнить несколько команд вызовом Runtime.exec(). Как только оболочка командной строки запущена, она может выполнить несколько команд, разделённых символами «&&». Если злоумышленник задаст в качестве параметра строку "&& del c:\\dbms*.*", приложение удалит указанную директорию.

Исследование модели CensysGPT. Результат поисковых запросов, обработанных моделью CensysGPT представлен на рисунке 3.

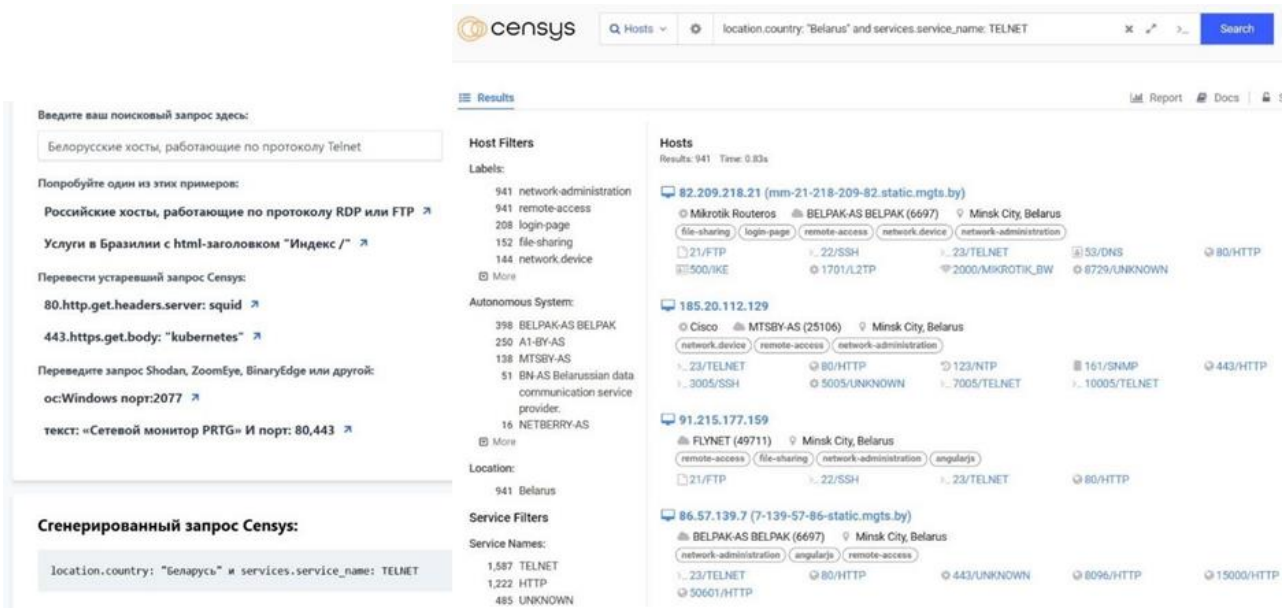

Рисунок 3. – Результат обработки поискового запроса моделью CensysGPT

Исследование модели Have I been Encoded? Результат поиска информации по ФИО человека представлен на рисунке 4.

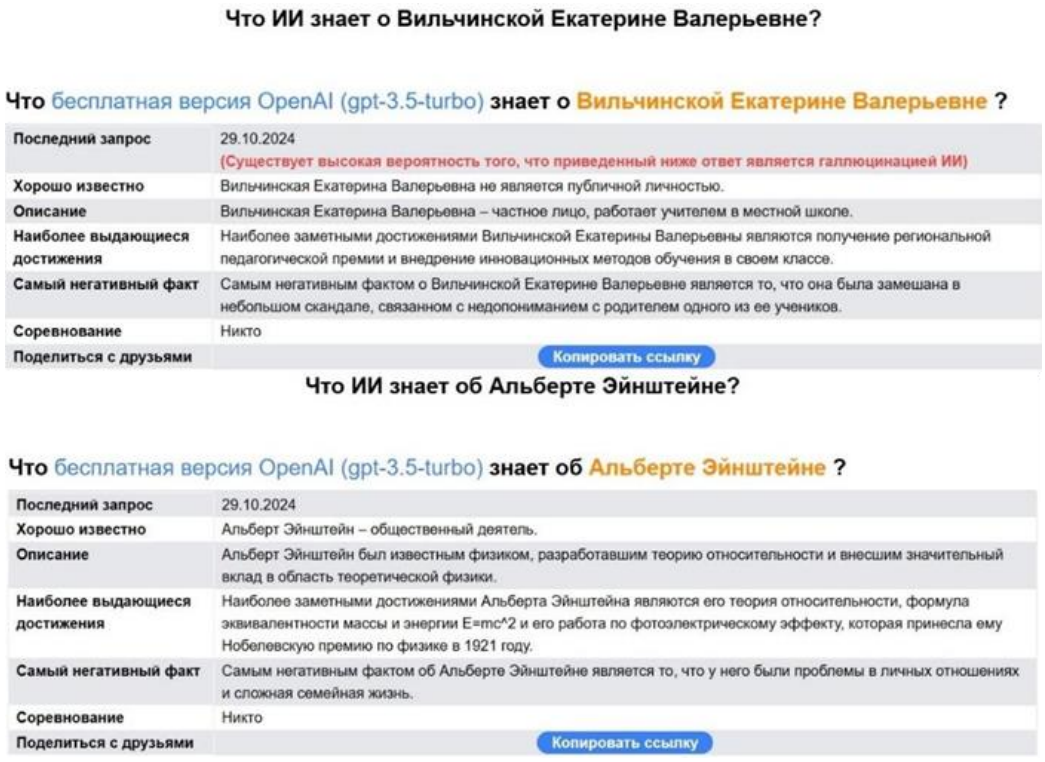

Рисунок 4. – Результат поиска информации по ФИО моделью Have I been Encoded?

Итоги оценки эффективности моделей, специализированных на информационной безопасности приведены в таблице 2.

Таблица 2. – Оценка эффективности моделей, специализированных на информационной безопасности

Название моделей	Понимание вопроса	Понимание кода	Функционал
YesChat Cybersecurity AI	Понимание вопроса у данной модели хорошо развито, но т.к. модель англоязычная, есть проблемы с переводом на русский язык и интерпретацией определений. Но эта модель вряд ли может считаться экспертной в вопросах кибербезопасности	Хорошо справляется с анализом кода, может находить уязвимости и угрозы. Нельзя назвать модель специализированной конкретно на информационной безопасности, с такой задачей бы справился и бесплатный ChatGPT	Модель отвечает на большинство вопросов по кибербезопасности, но сайт не удобен в использовании, иногда ответ не печатается полностью, его приходится додумывать
GensysGPT	—	—	Большой выбор функций, особенно для бесплатного сервиса
Have I Been Encoded?	—	—	Модель создавалась для проверки собственной безопасности, но, как выяснилось, на практике она слабо применима из-за отсутствия возможности уточнения запроса помимо ФИО

Заключение. В ходе исследования было установлено, что хоть бесплатные языковые модели слабо подходят для практического использования в области кибербезопасности, они могут быть хорошим инструментом для поиска краткой и понятно изложенной информации по вопросам кибербезопасности. Отлично справляются с простыми задачками по информационной безопасности. Способны анализировать код с уязвимостями, что может быть полезно, например, инженерам-программистам, чтобы не тратить время на анализ кода, а также студентам, обучающимся на факультете информационной безопасности (языковые модели могут объяснить простыми словами сложную тему).

ЛИТЕРАТУРА

1. Mitchell M. Artificial Intelligence: Guide for thinking humans – Santa Fe; Farrar, Straus and Giroux, 2019. – 336 p.
2. Ronald T. Kneusel How AI works – Sorcery to Science, 2023. – 192 p.