Отзыв о прохождении

Учебной практики (научно-исследовательская работа)

Студент	***Мурадян Денис Степанович***
Дата	*8 июня 2025*

Перед Мурадяном Денисом Степановичем была поставлена задача по ускорение сбора данных из открытых источников за счет автоматизации написания парсеров веб-страниц при помощи больших языковых моделей. Актуальность обусловлена необходимостью значительного ускорения сбора данных из открытых онлайн-источников и оптимизации затрат на разработку индивидуальных парсеров.

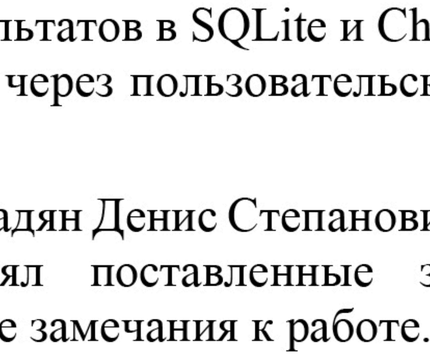
Денис Степанович в ходе прохождения практики своевременно и качественно выполнил следующие задачи:

1. Провел сравнительный анализ методов автоматизированного извлечения и сбора данных с веб-сайтов.
2. На основе лучших решений, полученных в ходе обзора, разработал модуль сбора и предобработки веб-страниц с автоматическим определением необходимости JS рендеринга и выбором между статическими запросами (requests) и динамическим рендерингом через Selenium.
3. Спроектировал и реализовал две стратегии очистки HTML-разметки:
    a. Строгая очистка для режима Structuring;
    b. Частичная очистка для режима Codegen.
4. Интегрировал систему с LLM (MistralAI) для двух режимов парсинга:
    a. Structuring (непосредственная генерация структурированных данных в формате JSON);
    b. Codegen (автоматическая генерация и исполнение Python-скриптов), обеспечив кэширование результатов в SQLite и ChromaDB;
5. Реализовал вывод результата через пользовательские интерфейсы: Gradio, FastAPI и веб-фронтенд на Jinja2/JavaScript.

В ходе работы обучающийся Мурадян Денис Степанович активно взаимодействовал с научным руководителем, своевременно выполнял поставленные задачи, проявлял самостоятельность, оперативно устранял вовремя выявленные замечания к работе.

- Система продемонстрировала высокую точность извлечения данных при тестировании на разнообразном наборе сайтов (Gismeteo, официальный сайт СПбГУ, Habr, Яндекс.Финанс (валютные курсы), RussianFood).
- Время запуска на ранее неизвестном сайте (без предварительного кэша) в режиме Codegen варьировалось от 17 до 75 с, тогда как при повторном запуске с кэшем — от 4,6 до 17,8 с.
- Режим Structuring обрабатывал страницы в среднем за 7,3–32,7 с.
- Интеграция с ChromaDB и семантическое кэширование обеспечили значительное ускорение повторных запросов при сохранении высокой точности.

Считаю, что Мурадян Денис Степанович за учебную практику заслуживает оценку «зачтено» в системе ECTS «A».

Руководитель практики,
старший преподаватель
кафедры информатики /  / Бушмелев
Федор Витальевич