УДК 004.584

ИССЛЕДОВАНИЕ ВОЗМОЖНОСТЕЙ YANDEX GPT И GIGACHAT ДЛЯ ОБЪЯСНЕНИЯ РЕШЕНИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Перегородиев Д. Е. (Университет ИТМО)
Научный руководитель - кандидат технических наук, доцент Гусарова Н. Ф.
(Университет ИТМО)

Введение. В последние годы искусственный интеллект (ИИ) стал неотъемлемой частью множества отраслей, от медицины до финансов. Однако, несмотря на значительные достижения, объяснение решений, принимаемых ИИ, остается сложной задачей. Актуальные исследования на тему объяснимости ИИ уже затрагивали тему применения больших языковых моделей (LLM) для решения данной задачи [1], [2], однако исследователи рассматривали англоязычные LLM, что уменьшает область применения их для объяснения решений ИИ на русском языке. В данной работе рассматриваются функциональные возможности Yandex GPT и GigaChat для объяснения решений ИИ. Анализируются существующие подходы и предлагаются новые методы для повышения прозрачности и доверия к системам ИИ.

Основная часть. В исследовании [1] изучается применение больших языковых моделей и специализированной методики промпт-инжиниринга, заключающейся в создании цепочки рассуждений на основе поведенческого фреймворка Belief-Desire-Intention [3], к многоагентной среде, распределяющей квоты на операции для пациентов на основе обучения с подкреплением, для объяснения решений, принятых агентами ИИ в данной среде. Авторы этой работы уделили особое внимание сравнению различных зарубежных англоязычных LLM, однако данный подход не представляется применимым в российской практике по ряду причин: 1) трудности с оплатой API зарубежных больших языковых моделей в условиях санкционного давления; 2) нераспространенность английского языка в российских организациях и учреждениях. В рамках текущей работы исследуется применение больших языковых моделей от компаний Yandex и Сбер - Yandex GPT и GigaChat соответственно.

Для проведения экспериментов была реализована многоагентная среда, для решения задачи распределения логистических ресуров для отправки грузов. Для передачи в промпт большим языковым моделям из этой среды были извлечены логи взаимодействия агентов по игровым эпизодам: дискретные характеристики агентов (срочность, сложность и полнота информации о доставке груза), наблюдения агентов в среде, их действия. Промпт также содержал в себе назначение роли LLM, подробное описание действий LLM для объяснения решений ИИ, описание ограничений среды и значение характеристик агентов. Для оценки результатов больших языковых моделей привлекался независимый эксперт-лингвист.

Применение Yandex GPT ограничено числом символов в одном промпте: до 1000. Данное ограничение не позволяет полноценно использовать данную большую языковую модель, учитывая тот факт, что логи минимального игрового взаимодействия двух агентов имеют не менее 5000 символов.

GigaChat, в свою очередь, не имеет такого ограничения, что позволяет не сокращать длину логов. Стоит отметить, что данная LLM требует строгого ограничения креативности: после обозначения роли модели она пытается реализовать решение проблемы, не получив никакой информации о ней, буквально создавая условия самостоятельно.

Полученные результаты объяснения решений ИИ от Yandex GPT и GigaChat удовлетворили независимого эксперта в срезах логической связности, доступности изложения и контекстуальной релевантности. Таким образом, открываются перспективы применения API данных LLM в реальных интеллектуальных системах принятия решений.

Выводы. Проведенное исследование демонстрирует потенциал русскоязычных LLM (Yandex GPT и GigaChat) для объяснения решений ИИ, однако выявляет их ключевые ограничения. Yandex GPT, несмотря на адекватное качество генерации объяснений, неприменим в сценариях с длинными контекстами из-за лимита в 1000 символов на промпт. GigaChat, обладая большей гибкостью, требует жесткого контроля креативности, чтобы избежать отклонений от исходных данных. Обе модели могут быть адаптированы для задач объяснимости в русскоязычных системах при условии оптимизации промптов (например, сжатия логов) и настройки гиперпараметров генерации.

Список использованных источников:

1. Исаков А. О. и др. Объяснимость поведения агентов в системах поддержки принятия клинических решений // Экономика. Право. Инновации. – 2024. – №. 4. – С. 50-59.

2. Joshi P. D. et al. HULLMI: Human vs LLM identification with explainability // arXiv preprint arXiv:2409.04808. – 2024.

3. Georgeff M. et al. The belief-desire-intention model of agency // Intelligent Agents V: Agents Theories, Architectures, and Languages: 5th International Workshop, ATAL'98 Paris, France, July 4–7, 1998 Proceedings 5. – Springer Berlin Heidelberg, 1999. – С. 1-10.