

УДК 004.85

АСПЕКТЫ, ОПРЕДЕЛЯЮЩИЕ ВЫБОР БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ СЕРВИСА АВТОМАТИЗИРОВАННОЙ ГЕНЕРАЦИИ ТЕСТОВЫХ ЗАДАНИЙ

Буюкова Е.С. (Университет ИТМО), **Насыров Н.Ф.** (Университет ИТМО),
Сташишин А.И. (Университет ИТМО)

Научный руководитель - кандидат технических наук, доцент Федоров Д.А.
(Университет ИТМО)

Введение. Людям, работающим в сфере образования, часто приходится проводить контроль знаний своих учеников для того, чтобы проверить степень усвоенности материала. При традиционном подходе к составлению тестов преподавателю необходимо тратить много сил и времени для формирования различных вопросов и вариантов ответов, однако современные технологии, в частности, большие языковые модели в специализированных сервисах способны предоставить человеку мощные инструменты для оптимизации процесса создания тестов.

Основная часть. Большая языковая модель (Large Language Model, LLM) - это тип алгоритма искусственного интеллекта, который применяет методы нейронной сети с большим количеством параметров для обработки и понимания человеческих языков или текста с использованием методов самоконтроля обучения. Преимуществом такой модели является способность автоматически извлекать смысловые данные из теста, обнаруживать темы, ключевые понятия и связи между ними [1]. В свою очередь, у пользователей появляется возможность существенно сократить время на создание тестов, сохраняя при этом качество вопросов.

Разработанный сервис предоставляет возможность сгенерировать набор тестовых вопросов, а также ответов на них по предоставленному текстовому контенту. Вопросы могут быть разного типа (на выбор, с открытым вариантом ответа, на сопоставление и др.) с последующим экспортом формат системы дистанционного обучения Moodle. В качестве инструментов генерации вопросов используется одна из двух больших языковых моделей, предлагаемых на выбор - starling-lm-7b-alpha с открытым исходным кодом или же GigaChat. В сервисе реализована возможность добавления дополнительных LLM.

Крайне важное значение имеет вопрос выбора большой языковой модели, поскольку он влияет на точность получаемых результатов, стоимость генерации текстового контента, быстродействие сервиса и др. Как правило, при выборе LLM, в первую очередь, принимаются во внимание технические характеристики модели - семейство модели, количество параметров, которые использовались в ходе ее обучения, тип модели (например, полная или квантованная), рейтинг в бенчмарках, требовательность к ресурсам и оборудованию, платность, открытость исходного кода, возможность дообучения модели и др. [2]

Однако немаловажными факторами являются и другие менее очевидные характеристики моделей, которые могут быть определяющими в выборе конкретной большой языковой модели. Стоит отметить, что могут существовать ограничения на использование LLM в некоторых регионах, возможны ситуации, когда разработчики

вводят дополнительные запреты на некоторые функции, например, регистрацию новых пользователей [3], количество запросов и устанавливает временные лимиты.

Касаясь лицензионных аспектов LLM, важно помнить, что далеко не все модели предполагают возможность их использования с целью получения коммерческой выгоды. Также возможны дополнительные ограничения, к примеру, на использование сгенерированного контента для обучения других языковых моделей.

Большие языковые модели - динамичная отрасль, из-за чего очередное обновление данных и алгоритмов больших языковых моделей может приводить к получению иных ответов по сравнению с предыдущими. Практически ежедневно появляются новые реализации больших языковых моделей, а также расширения к ним. Необходимо постоянно анализировать появляющиеся решения, чтобы проводить оценку их эффективности и возможности применения для решения описанной задачи. Анализируя бенчмарки, которые используются для оценки LLM, особое внимание следует уделить показателям, характеризующим способность моделей обрабатывать и генерировать контент на русском языке.

Далеко не для всех задач допустим вариант отправки данных на сторонние сервера. В пользовательских соглашениях часто указывается, что разработчик вправе использовать полученную информацию по своему усмотрению, например, для совершенствования работы большой языковой модели, что может быть неприемлемо для конфиденциальной информации.

Использование облачных больших языковых моделей чаще всего подразумевает трансграничную передачу данных. В ряде случаев подобная передача данных регламентируется законодательством - не всегда передача информации возможна. В рассматриваемой задаче автоматизированной генерации тестовых заданий и ответов к ним указанный аспект может быть существенным на некоторых видах предприятий - в секторе корпоративного обучения.

Необходимо учитывать, чтобы используемая LLM была защищена от "отравления данных" (Data Poisoning), а ее разработчики обеспечивали необходимый уровень защиты от уязвимостей и устойчивости модели.

Выводы. В ходе исследования и разработки сервиса был выявлен ряд факторов, зачастую не очевидных для разработчиков, но имеющих важное значение для выбора большой языковой модели: технические характеристики, открытость исходного кода, возможность дообучения, ограничения на использование, лицензионные аспекты и др. Их необходимо учитывать для достижения высокой эффективности и минимизации рисков использования LLM. В этом случае внедрение искусственного интеллекта в сферу образования способно обеспечить экономию ресурсов и достижение поставленных целей.

Список использованных источников:

1. Бронников А.Е. Влияние языковых моделей, основанных на искусственном интеллекте, на систему образования РФ // Вестник науки. 2023. №6 (63). URL: <https://cyberleninka.ru/article/n/vliyanie-yazykovykh-modeley-osnovannyh-na-iskusstvennom-intellekte-na-sistemu-obrazovaniya-rf-v-blizhayshe-buduschem> (дата обращения: 08.02.2024).

2. Hailin Chen, Fangkai Jiao, Xingxuan Li, Chengwei Qin, Mathieu Ravaut, Ruochen Zhao, Caiming Xiong, Shafiq Joty ChatGPT's One-year Anniversary: Are Open-Source Large Language Models Catching up? / Hailin Chen, Fangkai Jiao, Xingxuan Li, Chengwei Qin, Mathieu Ravaut, Ruochen Zhao, Caiming Xiong, Shafiq Joty [Электронный ресурс] // arXiv.org e-Print archive : [сайт]. — URL: <https://arxiv.org/abs/2311.16989> (дата обращения: 10.02.2024).

3. OpenAI временно остановила запись на ChatGPT Plus из-за высоких нагрузок [Электронный ресурс]. URL: <https://shazoo.ru/2023/11/16/150093/openai-vremenno-ostanovila-zapis-na-chatgpt-plus-iz-za-vysokix-nagruzok> (Дата обращения: 20.01.2024).

Буякова Е.С. (автор)

Подпись

Насыров Н.Ф. (автор)

Подпись

Сташишин А.И. (автор)

Подпись

Федоров Д.А. (научный руководитель)

Подпись