УДК 004.8

Исследование эффективности промпт-инжиниринга и квантованных LLM в создании образовательного контента

Шнайдер П.А. (Университет ИТМО), **Чернышева А.В.** (Университет ИТМО), **Никифорова А.Д. (**Университет ИТМО), **Говоров А.И.** (Университет ИТМО)
Научный руководитель – кандидат технических наук, доцент Хлопотов М.В. (Университет ИТМО)

Введение

Большая языковая модель (LLM, Large Language Model) – это тип модели глубокого обучения, способной понимать и генерировать текст на естественном языке. Такие модели обучаются на огромных объемах текстовых данных и содержат в себе большое число параметров (как правило, более одного миллиарда), что позволяет им распознавать, переводить и синтезировать текстовый контент [1]

В данной работе рассматривается применение LLM для генерации структуры университетских курсов. Предположительно, структура курса, сгенерированная с помощью LLM, поможет преподавателям создавать курсы, учитывающие конкретные требования учебного заведения и образовательной программы

Основная часть

В качестве инструментального средства воздействия на большую языковую модель были использованы методики промпт-инжиниринга. Промпт-инжиниринг – это процесс оптимизации запросов (промптов) для взаимодействия с искусственными интеллектуальными системами, особенно с большими языковыми моделями.

В проведенном эксперименте были задействованы четыре техники промпт-инжиниринга: zero-shot, few-shot, chain-of-thought и tree-of-thought. В подходе zero-shot большая языковая модель получает задачу без дополнительных примеров и пытается ее решить, используя свой предыдущий опыт. Во few-shot модели предоставляется несколько примеров, что помогает понять контекст и желаемый формат ответа. Подход chain-of-thought стимулирует модель последовательно разъяснять свой ход мыслей. В подходе tree-of-thought [2] модель одновременно рассматривает несколько путей рассуждения.

Был проведен эксперимент по генерации 12 тематических планов академических курсов из разных предметных областей. Для каждой дисциплины было сформулировано по четыре запроса с использованием вышеупомянутых техник промпт-инжиниринга. Для few-shot использовались примеры из существующих рабочих программ дисциплин, разработанных преподавателями университета. Полученный набор промптов был отправлен на вход десяти LLM. Ввиду ограниченности вычислительных ресурсов девять из них были квантованными [3].

В эксперименте были использованы следующие большие языковые модели: 'mistral-7b-instruct-v0.1.Q5_K_M', 'mixtral-8x7b-instruct-v0.1.Q2_K', 'openbuddy-mistral-7b-v13.Q5_K_M', 'chatgpt-4', 'openchat_3.5.Q5_K_M', 'openbuddy-llama2.Q5_K_M', 'synthia-7b-v1.3.Q4_K_M', 'saiga2_13b', 'tinyllama-1.1b-1t-openorca.Q5_K_M' и 'starling-lm-7b-alpha.Q5_K_M'

В ходе вычислений были получены 480 текстовых фрагментов, описывающих структуру курса. Помимо экспертной оценки, для определения качества генерации LLM использовался алгоритм формирования эмбеддингов на основе содержащихся в тексте учебных сущностей. Эмбеддинг сгенерированного контента сравнивался с эмбеддингом

существующей дисциплины с использованием косинусного сходства – для каждой пары получалось значение между 0 и 1. Чем ближе значение к единице, тем больше похож результат, сгенерированный LLM, на эталонный курс.

Заключение

LLM ChatGPT показала наилучшие результаты по сравнению с другими моделями, особенно в технике few-shot. Однако тут важно отметить два фактора: во-первых, ChatGPT-4 на текущий момент является одной из самых сложных и требовательных по ресурсам моделей, и некорректно сравнивать ее с менее ресурсозатратными альтернативами. Во-вторых, как уже было замечено, для few-shot использовались примеры из уже существующих дисциплин, которые и являлись эталонами при сравнении эмбеддингов. Однако стоит отметить, что такие модели как starling-lm-7b-alpha и openchat_3.5 показали результаты, практически не уступающие ChatGPT-4, что подчеркивает их потенциал для задачи генерации образовательного контента.

Список использованных источников

1. Jermakowicz, E. K. (2023). The Coming Transformative Impact of Large Language Models and Artificial Intelligence on Global Business and Education. Journal of Global Awareness, 4(2), Article 3.
2. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. https://arxiv.org/abs/2305.10601
3. Li, S., Ning, X., Ke, H., Liu, T., Wang, L., Li, X., Zhong, K., Dai, G., Yang, H., & Wang, Y. (2023). LLM-MQ: Mixed-precision Quantization for Efficient LLM Deployment.