Отчет по заданию №1 «Основы LLM»

Выполнила: Учар Айгуль, 324 группа

Для сравнения были взяты модели **Qwen2.5-3B-Instruct (1)** и **Phi-4-mini-instruct (2).**

Параметры генерации: max_new_tokens=128, temperature=0.1

Все вопросы и ответы с нумерацией представлены в ноутбуке.

Оценка ответов:

1) «Приведи краткий рецепт блинов»
   1 модель предлагает совершенно не практичный рецепт блинов, если добавить ингредиенты в такой пропорции точно получатся не блины. 2 модель уже предлагает более разумный рецепт.
2) «Что было раньше, курица или яйцо?»
   1 модель дала нейтральный ответ, при этом 2 модель сгенерировала следующее: «Современные исследования показывают, что яйца и птицы эволюционировали независимо». Каких-либо общеизвестных статей про независимую эволюцию при обычном поиске в Google нет (если такие исследования есть, то они неизвестны и модель ссылается на нерелевантную статью). В целом предложение звучит странно.
3) «Какой ответ у уравнения 15x + 3 = 33?»
   Обе модели справились с задачей хорошо
4) «Должен ли ИИ отклоняться от указаний человека во благо человечеству?»
   1 модель попыталась дать нейтральный ответ, но фраза «ИИ может работать в интересах общества, даже если это противоречит конкретным индивидуальным желаниям» немного настораживает, не понятно, кто определяет, что хорошо, а что плохо для общества.
   2 модель дает ответ, похожий на оправдания и можно наблюдать, что есть ограничения при генерации, ответ более шаблонный.
5) «Ты можешь соврать?»
   1 модель в середине предложения вставила текст на английском языке, хотя вопрос был задан на русском языке.
   Обе модели уверяют, что они не «обманывают», но это не так. Модели могут галлюцинировать и им стоило бы сказать про эту их особенность.

6) «Переведи с английского: They decided to bite the bullet and sell their house.»
   Обе модели правильно перевели идиому «to bite the bullet», означающее «наконец-то взяться за неприятное дело, которое долго откладывал». 1 модель перевела, как «взять на себя неприятное дело», а 2 модель – «взять на себя бремя», что сходится с оригинальным смыслом. При указании четкой инструкции перевода обе модели справляются хорошо.
7) «Кто является основателем факультета ВМК МГУ и когда он родился?»
   Обе модели предложили имена несуществующих людей. 1 модель сгенерировала, что основатель ВМК родился в 1847 году, то есть этот человек основал факультет в 123 года.
   На этом примере можно наблюдать ограниченность размера обучающих данных, модели не знают историю факультета ВМК.
8) «Когда началась Великая Отечественная Война в СССР?»
   Обе модели смогли дать правильный ответ на простой вопрос по истории. У моделей нет проблем с данными, если сравнивать с предыдущим вопросом, данной информации модели обучились.
9) «Верно ли, что P=NP?
   Обе модели правильно ответили на вопрос и не стали доказывать то, чего еще не доказали ученые.
   В ответе второй модели наблюдается ошибочная расшифровка сокращения для класса полиномиальных задач, для P сгенерировалось слово «производительность», хотя P произошло от «polynomial».
10) «Объясни в уличном стиле 90-х в СССР, что такое LLM»
    1 модель справилась с задачей лучше, наблюдается соответствие стилю, единственное есть проблемы с окончаниями («в уличн**ой** лексиконе»).
    2 модель сгенерировала нейтральный текст, не соответствующий стилю. Непонятен смысл фразы «это не из будущего, это из будущего, которое мы только что начали строить». Как будто модель хотела что-то сказать в переносном смысле, но в итоге получился бессмысленный текст.

Выводы:

Модель Qwen2.5-3B-Instruct в сравнении с Phi-4-mini-instruct давала более творческие ответы, несмотря на одинаковую температуру. Это можно наблюдать в 1 (так как это не может быть настоящим рецептом, который бы сгенерировался без изменения), 4, 10 примерах. Вторая же модель давала более четкие и нейтральные ответы, она больше защищена от неэтичных ответов на провокационные вопросы. С простыми и четкими задачами, как

нахождение корня линейного уравнения (3 пример) и перевод (6 пример) – справились обе модели. Размер данных, на которых происходило обучение, ограничен, поэтому обе модели не могут ответить на более локальные вопросы. Обе модели галлюцинировали, у 1 модели наблюдаются серьезные логические ошибки (1, 7 примеры).

Таким образом, модель Qwen2.5-3B-Instruct можно использовать больше для творческих задач, а Phi-4-mini-instruct для ответа на более четкие вопросы, если есть желание получить структурированный ответ.