Привет!

1. Вот ссылка на тикет, согласованный с УМР, с тестовым сетом – тестовые сеты там файлы «test\_instr\_set\_дебетовая\_БК\_04\_07\_2025\_v2.xlsx» и «test\_terms\_set\_дебетовая\_БК\_04\_07\_2025\_v2.xlsx». Дополнительно прикладываю данные файлы к письму, так как в тикете много чего еще приложено, и легко запутаться.
2. Скрипты какого рода требуются? Скрипты проверки и определения метрик на Python, которые мы запускаем локально на тестовом сете и на сгенерированных на дообученном GigaChat ответах?

У нас проверка состоит из формальных и неформальных критериев. Мы обучали Gigachat писать хорошие рекламные SMS, и у нас общая оценка, которую мы заявляли состоит из определения доли критических (длина SMS, употребление запрещенных слов и пр.) и некритических (два глагола подряд, употребление причастий и пр.) ошибок в SMS, а также из оценки профессиональными редакторами, стало ли в целом SMS лучше со стилистической точки зрения.

Формализованную часть проверки провели, неформализованную (оценка редакторов) еще нет, но скоро начнем. Могу прислать Python скрипты с функциями, которые мы применяли для определения прохождения или непрохождения вышеупомянутых формальных критических и некритических проверок, но для полного повтора наших действий нужно будем самостоятельно сгенерировать на обученном и необученном GigaChat рекламные SMS по промптам из обоих тестовых сетов (тестовых сетов у нас, как и адаптеров, два – для промптов с персонализацией по терминам и для промптов с персонализацией по инструкциям) на соответствующих LoRA-адаптерах, и потом реализовать проверку через вышеупомянутые функции проверки и выполнить общий вывод статистики. Такое ощущение, что всё вышеописанное потребует определённого погружения в наш кейс, и не знаю, насколько это для вас целесообразно. Если считаете целесообразным – давайте направляю Python функции проверок; если нет – считаю, можно принять наши замеры метрик, о которых подробнее в пункте ниже.

1. Как уже написал выше, общая метрика состоит из формализованной и неформализованной части.

Как написано в тикете, решили использовать следующую метрику:

Результат замеров формализованной части метрик после дообучения:

2 критические ошибки на 10 сообщений от дообученного GigaChat;

6 некритических ошибок на 10 сообщений от дообученного GigaChat.

Но вторая половина оценки – это оценка редакторами – к ней только приступаем.

Пока на текущий момент считаю дообучение успешным, несмотря на то, что формально количество некритических ошибок не сократилось до того уровня, который заявляли, как ожидаемый после дообучения (ожидаемый уровень ошибок после дообучения оценивался приблизительно, кроме того, некритические ошибки не столь принципиальны и даже в нашей итоговой оценке имеют вес всего 0,2).