# Отчет по заданию №2 «Основы LLM»

Выполнила: Учар Айгуль, 324 группа

Для сравнения были взяты модели RefalMachine/RuadaptQwen2.5-1.5B-instruct и TinyLlama/TinyLlama-1.1B-Chat-v1.0

**Датасет:** https://huggingface.co/datasets/RussianNLP/russian\_super\_glue/

**Выбранная задача:** генеративная бинарная классификация ответов. В датасете представлены вопросы, контекст и метка в виде 1 (да) или 0 (нет). Для исследования моделей в промтах использовались только вопросы и ответы к ним, контекст не учитывался.

Для оценки качества классификации были использованы метрики accuracy и F1-weighted. Предварительный анализ данных показал, что 60.7% примеров в датасете имеют ответ «да», что указывает на наличие дисбаланса классов. В связи с этим была выбрана метрика F1-weighted, которая учитывает распределение классов и позволяет получить более объективную оценку качества модели, особенно в условиях несбалансированных данных.

**Программная реализация:**

Функция evaluate принимает реальный (y\_true) и предсказанный (y\_pred) ответы и вычисляет метрики. Реализуется метрика f1\_weighted.

Функция aggregation собирает метрики по всему тестовому набору и вычисляет соответствующие метрики.

Функция create\_messages генерирует структуру сообщений для взаимодействия модели с задачей, используется промт: «Тебе будет задан вопрос. Ответь "1" (да) или "0" (нет). Предоставь ответ в виде одной цифры.\n\nВопрос: {question}»

**Значения метрик:**

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | Accuracy | F1-weighted | Leaderboard Score |
| RefalMachine/RuadaptQwen2.5-1.5B | 0.57 | 0.572 | 0.571 |
| TinyLlama/TinyLlama-1.1B-Chat-v1.0 | 0.62 | 0.478 | 0.549 |

По результатам тестирования модель TinyLlama-1.1B-Chat-v1.0 показала более высокую точность (62%) по сравнению с RuadaptQwen2.5-1.5B (57%). То есть TinyLlama чаще выдает правильные ответы, но при этом метрика F1-weighted выше у RuadaptQwen2.5-1.5B, то есть преобладание данных с ответом «да» больше повлияло на TinyLlama.

**Вывод:**

Реализованная задача позволила оценить генеративные способности моделей на бинарной классификации. TinyLlama дает более уверенные, но менее сбалансированные предсказания, тогда как RuadaptQwen2.5-1.5B лучше распределяет ответы между классами.