Методы дообучения больших языковых моделей инструкциям для повышения качества работы на русском языке

Лебедев Андрей, группа 424
Научный руководитель:
к.ф.-м.н. Тихомиров Михаил Михайлович

Задача

Исследовать и применить современные подходы к адаптации больших языковых моделей на инструкциях для повышения качества их работы на русском языке.

В частности, дообучить конкретную модель на качественных русскоязычных данных, используя:

- Supervised Fine-Tuning (SFT),
- Simple Preference Optimization (SimPO).

Актуальность

- Распространение больших языковых моделей.
- Недостаточное качество работы на русском языке.
- Проблемы при работе с длинным контекстом.

Методология обучения

Исходная модель: *Qwen2.5-3B-Instruct*

Использованные методы дообучения:

- *Supervised Fine-Tuning (SFT)* – дообучение на размеченных данных, пары «запрос – эталонный ответ».
- *Simple Preference Optimization (SimPO)* – оптимизация предпочтений без дополнительной модели награды, тройки «запрос – хороший ответ – плохой ответ».

Методология обучения

- **SFT** – минимизация кросс-энтропии между предсказанием модели и эталонным ответом. Та же задача, что на pretrain-этапе:

$$\mathcal{L}_{SFT}(\theta) = -E_{(x,y)\sim\mathcal{D}} \left[\sum_{i=1}^{|y|} \log \pi_\theta \left(y_i \mid x, y_{<i} \right) \right]$$

- **SimPO** – максимизация вероятности, что положительный ответ вероятнее отрицательного на некоторую величину:

$$\mathcal{L}_{SimPO}(\pi_\theta) = -E \left[\log \sigma \left(\frac{\beta}{|y_w|} \log \pi_\theta(y_w \mid x) - \frac{\beta}{|y_l|} \log \pi_\theta(y_l \mid x) - \gamma \right) \right]$$

Meng Y., Xia M., Chen D. Simpo: Simple preference optimization with a reference-free reward //arXiv preprint arXiv:2405.14734. – 2024.

Методология обучения

Используемые датасеты:

- *Vikhrmodels/GrandMaster-PRO-MAX:* разнообразные примеры инструкций-ответов, синтетика GPT-4.
- *IlyaGusev/saiga_scored:* размеченные по качеству пары инструкций-ответов.
- *Vikhrmodels/Grounded-RAG-RU-v2:* набор данных для задач с глубоким пониманием контекста.
- *IlyaGusev/saiga_preferences:* предпочтения пользователей для улучшения генерации.

Методология оценки

- **Качество генерации на русском языке:**

  Оценка с помощью метода LLM-as-a-Judge на наборе RU Arena Hard. Показывает, насколько в среднем тестируемая модель лучше других по мнению сильной LLM-судьи.

- **Оценка знаний:**

  Междисциплинарный бенчмарк ruMMLU с выбором ответов.

- **Оценка работы с длинным контекстом:**

  Бенчмарк LIBRA – сабсеты RuBabilongQA1 и RuBabilongQA2 для длинного контекста.

Результаты экспериментов

Модель	LLM-as-a-Judge	Avg. len	ruMMLU	ruBABILongQA1	ruBABILongQA2
Qwen2.5-3B-Instruct	0.136	418	0.550	0.648	0.260
Qwen2.5-3B-Instruct + GM					
+60%	**0.194**	568	0.551	0.623	0.367
+80%	0.149	539	0.551	0.638	**0.383**
+100%	0.148	551	0.550	0.647	0.363
Qwen2.5-3B-Instruct + 100% Saiga	0.086	410	0.552	0.640	0.373
Qwen2.5-3B-Instruct + 100% GM + Saiga					
+20% Saiga:	0.101	484	**0.559**	0.650	0.350
+40% Saiga:	0.100	479	0.558	0.645	0.355
+60% Saiga:	0.108	456	0.557	0.652	0.378
+100% Saiga:	0.098	439	0.558	**0.670**	0.355
Qwen2.5-3B-Instruct + 100% RAG	0.185	467	0.553	0.642	0.240
Qwen2.5-3B-Instruct + 100% GM + RAG					
+ 100% RAG (4/5, 10%)	0.130	516	0.536	0.635	0.368
+ 100% RAG (5/5, 5%)	0.177	529	0.543	0.648	0.335
+ 100% GM + 60% Saiga + 100% RAG (5/5, 5%)	0.117	495	0.550	**0.670**	0.312
+ 100% GM + 100% RAG + 100% Saiga + 100% Saiga Pref	0.093	447	0.551	0.632	0.363

Выводы

- Исследованы **методы дообучения** LLM инструкциям и **способы оценки** эффективности работы на русском языке.
- Показано, что использование качественных открытых датасетов улучшает **генерацию, знания и работу на длинном контексте**.
- Обнаружено, что **комбинирование** разных типов **данных** и **методов** адаптации обеспечивает сбалансированный рост всех ключевых метрик.
- Работа **демонстрирует практические подходы улучшения LLM** на русском языке через выбор методов адаптации и датасетов.