УДК 1

БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ ДЛЯ СЛЕДОВАНИЯ ИНСТРУКЦИЯМ НА РУССКОМ ЯЗЫКЕ: МОДЕЛИ И ДАТАСЕТЫ С ОТКРЫТОЙ ЛИЦЕНЗИЕЙ ДЛЯ КОММЕРЧЕСКОГО ИСПОЛЬЗОВАНИЯ

© 2023 г. **Дмитрий Косенко**[1,3,*], **Юрий Куратов**[1,2,3,**], **Диляра Жарикова**[3,***]

Представлено Д. П. Косенко
Поступило 31.08.2023
После доработки 30.09.2023
Принято к публикации 15.10.2023

В данной статье представлен подход к разработке и дообучению больших языковых моделей для русского языка, способных следовать инструкциям в различных доменах. В качестве базовых моделей использованы XGLM-4.5B, LLaMA-1 7B, LLaMA-1 13B, LLaMA-2 7B, LLaMA-2 13B, ruGPT-3.5 13B. В данной работе проводится сравнения двух основных методик дообучения: дообучение всех параметров модели и дообучение с использованием LoRA слоёв. Для создания датасета для дообучения модели использованы несколько открытых источников данных на английском языке, таких как Databricks Dolly 15k, OpenAssistant Conversations Dataset (OASST1), chip2-instruct-alpha-v6a-1, которые затем были переведены на русский язык с помощью модели WMT 21 En-X с лицензией MIT. В данной работе показано, что качество предоставляемых для обучения инструкций существенно влияет на способность решения задач на автоматических метриках качества MT-BENCH и MMLU. При этом качество моделей, обученных на собранном в рамках работы датасете с коммерческой лицензией, достигает сравнимых результатов с моделями, дообученными на датасете Saiga с ограниченной лицензией. Дообученные языковые модели и собранный набор данных для русского языка выложены в открытый доступ с лицензиями, подходящими для коммерческого использования.[1]

Ключевые слова и фразы: большие языковые модели, языковые модели, языковые модели для русского языка

1. ВВЕДЕНИЕ

Несмотря на существование больших языковых моделей ChatGPT [1], GPT-4 [2], CLAUDE [3], способных следовать инструкциям на множестве языков, включая русский, они доступны исключительно через внешние API. Это влечет за собой несколько существенных ограничений при использовании моделей в коммерческом секторе: невозможность отправки личной или корпоративной информации на сторонние серверы, невозможность выстраивания критической инфраструктуры вокруг данных инструментов, так как их работоспособность может быть нарушена или доступ к ним может быть заблокирован. Для преодоления этих ограничений необходима возможность развертывания языковых моделей на собственных серверах или серверах партнеров.

[1] Ссылки для скачивание датасетов и моделей будут добавлены после рецензирования

[1] *Московский физико-технический институт, Москва, Россия*
[2] *AIRI, Москва, Россия*
[3] *DeepPavlov.ai*
[*] *E-mail: dimweb.tech@mail.ru*
[**] *E-mail: yurii.kuratov@phystech.edu*
[***] *E-mail: dilyara.baymurzina@phystech.edu*

Для дообучения собственной большой языковой модели необходимо использовать предобученную языковую модель объемом в десятки миллиардов параметров в качестве начальной инициализации и высококачественные наборы инструкций с решениями задач из разных предметных областей на соответствующем языке.

В данной работе подготовлены датасеты и дообучены большие языковые модели для следования инструкциям на русском языке, подходящие для коммерческого использования, также проанализировано качество собранных датасетов для дообучения разных архитектур. Этот опыт может быть полезен небольшим компаниям и стартапам, которые хотят создать собственную языковую модель для собственных целей.

2. МОДЕЛИ

Вопрос о разумном размере большой языковой модели для качественного решения достаточного количества задач на текущий момент остается открытым. Согласно рейтингам сравнения моделей MT-BENCH [4] и Open LLM Leaderboard [5], прослеживается четкое разделение моделей по качеству в зависимости от размера моделей. Лучше всего с задачами справляются проприетарные модели с не объявленными официально размерами, затем идут модели, содержащие 65–70, 30, 13 и, наконец, 7 миллиардов параметров. Тем не менее, в узких предметных областях, таких как генерация кода, размер модели может быть уменьшен до 1,3 миллиарда при сохранении качества, сравнимого с моделями размером 175 и 540 миллиардов параметров [6]. Однако однозначно можно сказать, что увеличение количества параметров модели положительно влияет на способность моделей к «пониманию» языка и выполнению задач из различных доменов. Так, в работе Gopher [7] было явно показано, что с ростом количества параметров наблюдается рост количества правильных решенных задач на метриках MMLU [8] и Big-Bench [9].

XGLM Учитывая эти особенности, был проведен поиск многоязычной модели необходимого размера. На момент начала работы над данной статьей количество моделей, содержащих не более шести миллиардов параметров (оценочное значение размера модели, при котором для полного дообучения достаточно 4 видеокарты A100 с памятью 40 Гб) и содержащих русскоязычные примеры в обучающей выборке, было ограничено. Одной из таких моделей является XGLM-4.5B[2] [10] — это многоязычная модель, использующая оригинальную архитектуру GPT-2 [11] без изменений и предобученная на данных 134 языков. По заявлению авторов, датасет для предобучения модели содержал более 147 миллиардов токенов на русском языке.

LLaMA В качестве второго семейства моделей были выбраны LLaMA-1[3] [12] и LLaMA-2[4] [13], по различным рейтингам являющиеся лучшими моделями с открытыми весами. Для коммерческого использования разрешено применение только моделей семейства LLaMA-2, однако ввиду популярности моделей LLaMA-1 в данной работе также приведены результаты дообучения и LLaMA-1. Модели семейства LLaMA используют оригинальную архитектуру Трансформер [14] со следующими архитектурными изменениями: нормализация RMSNorm [15] входа каждого подслоя Трансформера вместо нормализации выхода; активация SwiGLU [16]; использование поворотных вместо абсолютных позиционных векторных представлений RoPE [17]. Для предобучения моделей LLaMA-2 использовалось два триллиона токенов на разных языках, русскоязычных токенов было всего 26 миллиардов.

ruGPT-3.5 В качестве третьей модели была выбрана ruGPT-3.5 13B[5] [18], предобученная на 300 Гб текста из различных предметных областей: литература, программный код, юридические документы, Википедия и др. Данная модель использует оригинальную архитектуру

[2]https://huggingface.co/facebook/xglm-4.5B
[3]https://huggingface.co/huggyllama/llama-7b
[4]https://huggingface.co/meta-llama/LLaMA-27B
[5]https://huggingface.co/ai-forever/ruGPT-3.5-13B

GPT-2 без каких-либо архитектурных изменений, кроме увеличения числа слоев и длины контекста до 2048 токенов. Данная модель также лежит в основе сервиса GigaChat [19], предоставляющего API для закрытой проприетарной модели.

3. ДАННЫЕ

Наборы данных для обучения следованию инструкциям с лицензией для коммерческого использования на нужном языке можно получить тремя способами: составление датасета с привлечением экспертов; генерация инструкций и диалогов с помощью больших языковых моделей, уже неплохо решающих поставленные задачи; автоматический перевод открытых датасетов на требуемый язык. Первый подход на данный момент является самым качественным и гибким, но в то же время требует значительных затрат для поддержания контроля качества составляемого датасета, обеспечения вариативности инструкций, проверки достоверности информации. Для осуществления второго подхода распространено использование модели Flan-UL2 [20], поскольку эта модель распространяется под открытой лицензией и способна решать простые задачи на английском языке. Так, например, в работе [21] авторы генерируют и валидируют инструкции с помощью небольшой модели, а далее обучают большую языковую модель на сгенерированных инструкциях. Использование примеров, сгенерированных с помощью проприетарных моделей, таких как ChatGPT, для коммерческих задач на данный момент ограничено правилами их применения. Третий подход — перевод датасета на целевой язык — является наиболее дешевым и быстрым, однако требует внимательного отбора исходных датасетов, которые должны распространяться под одной из открытых лицензий.

3.1 СБОР ДАТАСЕТА С КОММЕРЧЕСКОЙ ЛИЦЕНЗИЕЙ

В процессе дообучения моделей были использованы три датасета с лицензиями MIT и Apache 2.0:

- databricks-dolly-15k [22] — открытый набор данных, содержащий примеры выполнения инструкций, созданные сотрудниками Databricks в различных поведенческих категориях, описанных в работе InstructGPT. Включает мозговой штурм, классификацию, ответы на вопросы по тексту, генерацию историй, извлечение информации из текста, открытые вопросы и реферирование текста.
- OpenAssistant Conversations (OASST1) [23] — созданный и аннотированный людьми корпус диалогов в стиле виртуального помощника, состоящий из 160 тысяч сообщений на 35 разных языках, аннотированных 460 тысячами оценками качества, что дает более 10 тысяч полностью аннотированных диалогов. Корпус является продуктом краудсорсинга с участием более 13,5 тысяч добровольцев по всему миру. В данной работе использовалась только англоязычная часть данного датасета.
- chip2-instruct-alpha-v6a[6] — открытый набор данных, инструкции получены из открытых источников или синтетически сгенерированы людьми. Сгенерированные инструкции были созданы с помощью открытой модели FLAN-UL2-20B на основе нескольких подсказок. Ответы на инструкции были сгенерированы с использованием длинных диалоговых подсказок, которые были полностью или в основном написаны людьми.

Перечисленные датасеты содержат примеры на английском языке; для того же, чтобы модель понимала те же инструкции на русском, данные необходимо перевести. На данный момент одной из лучших моделей перевода является WMT 21 En-X 4.7B [24] — многоязычная модель, содержащая энкодер и декодер и обученная для многоязычного перевода «один ко многим». Модель может напрямую переводить английский текст на семь других языков, включая русский. Для сохранения исходного форматирования текста, он разбивался по символам перевода строк, переводился по параграфам, а затем объединялся обратно. Не смотря

[6] https://github.com/Rallio67/language-model-agents/tree/main

на то, что некоторые примеры были переведены хуже в связи с отсутствием контекста (других параграфов), сохранение форматирования зачастую играет ключевую роль для многих инструкций.

После автоматического перевода были выявлены проблемы с диалогами, содержащими программный код, ASCII-рисунки, текстовые игры, такие как крестики-нолики, и технические обучающие статьи. При переводе кода модель переводила даже базовые конструкции языка программирования; в сложных текстовых последовательностях вместо копирования содержимого модель начинала бесконечно повторять некоторую фразу или символ. Диалоги с такими ошибками перевода были вручную полностью исключены из обучающей выборки. Во избежание «забывания» английского языка в обучающую выборку также были добавлены оригинальные примеры на английском языке из упомянутых датасетов. После очистки итоговый датасет для обучения содержит 250 877 диалогов с инструкциями на русском и английском языках.

3.2 Другие датасеты

Также в данной работе проводятся эксперименты с дообучением на 7 различных датасетах: ru_turbo_alpaca[7], ru_turbo_saiga[8], ru_sharegpt_cleaned[9], oasst1_ru_main_branch[10], gpt_roleplay_realm[11], ru_turbo_alpaca_evol_instruct[12], ru_instruct_gpt4[13] их комбинация обозначена общим словом Saiga, который распространяется под лицензией CC BY 4.0 - Creative Commons. Датасет имеет высокое качество данных, однако не может быть использован для обучения языковых моделей в коммерческих целях. Результаты дообучения на Saiga приводятся для возможности сравнить качество разных архитектур, дообученных на одном и том же наборе данных.

4. МЕТРИКИ

Для оценки моделей использовались две автоматические метрики: MT-BENCH и MMLU.

MT-BENCH [4] основан на идее оценки с помощью большой языковой модели. В данной работе в качестве такой модели применялась GPT-4. Для оценки возможностей моделей на русском языке все инструкции и вопросы из оригинального MT-BENCH были автоматически переведены с помощью модели WMT 21 En-X 4.7B, а затем вручную отредактированы человеком. В остальном, структура и механизм оценки не отличались от оригинальной работы.

MMLU [8] — метрика, оценивающая способность решать тесты, требующие выбрать один из четырех вариантов ответа. Она охватывает 57 разделов из технических и гуманитарных наук. На вход модели подаются задания из этого датасета; на выходе сравниваются вероятности токенов букв «A», «B», «C», «D», являющихся вариантами ответа. Токен с наибольшей вероятностью считается ответом. MMLU RU[14] — оригинальный датасет MMLU, автоматически переведенный на русский язык с помощью Yandex.Translate API командой NLP Core управления R&D ML в SberDevices.

5. ЭКСПЕРИМЕНТЫ

В ходе экспериментов было протестировано два основных подхода для создания моделей: полное дообучение всех параметров модели и дообучение только LoRA-слоев [25].

5.1 Полное дообучение

[7]https://huggingface.co/datasets/IlyaGusev/ru_turbo_alpaca
[8]https://huggingface.co/datasets/IlyaGusev/ru_turbo_saiga
[9]https://huggingface.co/datasets/IlyaGusev/ru_sharegpt_cleaned
[10]https://huggingface.co/datasets/IlyaGusev/oasst1_ru_main_branch
[11]https://huggingface.co/datasets/IlyaGusev/gpt_roleplay_realm
[12]https://huggingface.co/datasets/IlyaGusev/ru_turbo_alpaca_evol_instruct
[13]https://huggingface.co/datasets/lksy/ru_instruct_gpt4
[14]https://github.com/NLP-Core-Team/mmlu_ru

Ввиду ограниченности вычислительных ресурсов, в данной работе демонстрируются только результаты полного дообучения модели XGLM-4.5B. Обучение данной модели с длиной контекста 2048 токенов требует четыре видеокарты A100-40GB при использовании технологии DeepSpeed ZeRO Stage 3 [26] и чекпоинтингом градиентов. В качестве оптимизатора использовался AdamW [27] с learning rate 3×10^{-6}, betas 0.9 и 0.95, gradient_accumulation_steps равный 16, batch_size_per_gpu равным 1, weight_decay равным 0.0. Остальные параметры были выбраны по умолчанию по состоянию версии библиотеки transformers==0.9.2. В качестве функции потерь для языкового моделирования использовалась кросс-энтропия. В процессе дообучения использовался следующий шаблон затравки:

<s> Human:
Запрос от человека
Assistant:
Ответ от ассистента <|endoftext|>

При дообучении использовался классический подход к обучению языковой модели, состоящий в предсказании следующего токена для входной последовательности.

5.2 Дообучение с LoRA-слоями

При дообучении с LoRA-слоями были использованы следующие модели: XGLM-4.5B, LLaMA 7B, LLaMA-2 7B, LLaMA-2 13B, ruGPT-13B. Все модели обучались с использованием библиотек peft [28] и bitsandbytes [29] с точностью int8. В качестве оптимизатора использовался AdamW с learning rate 3×10^{-4}, betas 0.9 и 0.95. Для LoRA-слоев были использованы следующие гиперпараметры: $r = 16$, $LoRA\_alpha = 16$, $LoRA\_dropout = 0.05$. LoRA-слои были применены для линейных слоев, содержащих в себе следующие подстроки: q_proj, v_proj, k_proj, out_proj, o_proj, названия данных слоев фигурируют в блоках механизма внимания представленных моделей.

В процессе тренировки использовался следующий шаблон затравки:

<s> system
Ты русскоязычный автоматический ассистент. Ты разговариваешь с людьми и помогаешь им. </s>
<s> user
Запрос от человека </s>
<s> bot
Ответ от ассистента </s>

Из-за особенностей различных токенизаторов было обнаружено, что для предсказуемой и корректной токенизации шаблона затравки специальные токены <s>, system, bot, user, </s> необходимо разделять пробелами от любых других символов. В случае диалога, содержащего более одной пары реплик, токены user и bot повторялись соответствующее длине истории количество раз.

Для языкового моделирования использовалась кросс-энтропия, однако в отличие от предыдущего примера входящая инструкция не учитывалась в подсчете метрик.

6. РЕЗУЛЬТАТЫ

В данном разделе приведены оценки больших языковых моделей, в том числе дообученных на датасете, описанном в разделе 3. При генерации текста использовался стандартный интерфейс генерации библиотеки transformers [30] (версии 4.31.0) со следующими параметрами генерации:

temperature=0.2
top_p=0.95
top_k=40
do_sample=false
num_beams=1

max_new_tokens=1536
repetition_penalty=1.1
no_repeat_ngram_size=15

Модель	Дообучение	Данные	MT-BENCH RU			MMLU k=5, ctx=2048	MMLU RU k=5, ctx=2048
			First turn	Second turn	Average		
LLaMA 7B (O)	LoRA	Saiga (C)	4.23	3.06	3.65	35.62	30.18
		Our (A,M)	3.50	2.91	3.20	32.51	28.58
LLaMA-2 7B (O)	LoRA	Saiga (C)	4.76	3.62	4.19	43.24	35.91
		Our (A,M)	4.06	2.97	3.51	41.02	29.68
LLaMA-2 13B (O)	LoRA	Saiga (C)	5.96	4.12	**5.04**	**54.48**	**42.60**
		Our (A,M)	4.30	2.98	3.64	52.31	38.86
XGLM-4.5B (M)	finetune	Our (A,M)	2.32	1.97	2.15	23.73	24.36
	LoRA	Saiga (C)	2.73	1.96	2.35	25.40	24.91
		Our (A,M)	2.28	2.08	2.18	25.82	25.55
ruGPT-3.5 13B (M)	—	GigaChat	4.53	3.53	4.03	—	—
	LoRA	Saiga (C)	3.71	2.75	3.23	26.90	26.79
		Our (A,M)	2.61	1.75	2.18	24.69	24.12
ChatGPT	—	ChatGPT	8.70	7.45	**8.07**	—	—

Таблица 1: Результат оценок на MT-BENCH и MMLU. С ростом значения метрики, растет качество. «Our» датасет обозначает собранный в рамках данной работы и описанный в разделе 3 набор данных. Пометка «M» обозначает лицензию MIT, «O» — Open RAIL++, «C» – CC BY 4.0, «A» — Apache 2.0

В таблице 1 приведены результаты оценки моделей с помощью метрик MT-BENCH RU, MMLU, MMLU RU. Модели LLaMA 7B, LLaMA-2 7B, LLaMA-2 13B, обученные на датасете Saiga, использованы «как есть» из библиотеки transformers. Модель ruGPT-3.5 13B с датасетом GigaChat обозначает сервис GigaChat[15] версии v1.13.0. Остальные комбинации модель-датасет-метод были дообучены в рамках данной работы, в том числе на датасете Saiga.

Результаты оценки показывают, что дообучение XGLM-4.5B с LoRA-слоями превосходит метод дообучения всех параметров модели для данной архитектуры на собранном в работе датасете. При этом необходимо отметить, что в случае с LoRA-слоями требуется в четыре раза меньший объем видеопамяти по сравнению с дообучением всех параметров модели.

Использование датасета Saiga, содержащего более качественные и сложные инструкции, позволяет достичь более высоких результатов дообучения для большинства моделей. Однако датасет Saiga был получен с помощью генерации большими языковыми моделями OpenAI, лицензионные соглашения которых запрещают использование полученных таким образом данных для дообучения коммерческих моделей. Поэтому сформированный в данной работе датасет, содержащий русскоязычные и англоязычные инструкции, позволяет обучить модель для коммерческого использования. При этом дообучение на собранном датасете позволяет достичь сравнимого качества с моделями, дообученными на датасете Saiga, а в случае XGLM-4.5B слегка превосходит их по некоторым метрикам.

Итоговыми моделями, имеющими лицензию, пригодную для коммерческого использования, являются: LLaMA-2, XGLM-4.5B и ruGPT-3.5 13B, обученные на собранном в данной работе датасете. Среди них наилучшее качество показывают LLaMA-2 13B и LLaMA-2 7B. Модели XGLM-4.5B и ruGPT-3.5 13B имеют сравнимое качество, при этом первая использует почти втрое меньше параметров.

[15]https://developers.sber.ru/portal/products/gigachat

7. ВЫВОДЫ

В данной работе показано, что качество предоставляемых для дообучения инструкций существенно влияет на способность моделей к решению задач. Открытые модели для перевода подходят для быстрого создания датасетов с простой структурой на нужном языке, но демонстрируют низкое качество на нестандартных входных данных, что приводит к необходимости ручной пост-обработки. Выбор базовой модели также значительно влияет на качество дообучения с использованием LoRA-слоев.

Также в работе показано, что русскоязычный датасет Saiga содержит более качественные инструкции, однако использование собранного в рамках работы датасета позволяет достичь сравнимых результатов при сохранении коммерческой лицензии. Так, например, модели LLaMA-2 7B и LLaMA-2 13B имеют сравнимое качество с моделями, дообученными на датасете Saiga.

Дообученные языковые модели и использованные наборы данных для русского языка выложены в открытый доступ[16] с лицензиями, подходящими для коммерческого использования.

8. КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

[1] Introducing ChatGPT. URL: https://openai.com/blog/chatgpt.

[2] OpenAI. Gpt-4 technical report, 2023.

[3] Claude. A next-generation AI assistant for your tasks no matter the scale. Url: https://www.anthropic.com/.

[4] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[5] Daniel Park. Open-llm-leaderboard-report, 2023.

[6] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need, 2023.

[7] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

[8] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[9] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan

[16] Ссылки для скачивание датасетов и моделей будут добавлены после рецензирования

Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023.

[10] Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual language models, 2022.

[11] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[12] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[13] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[15] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

[16] Noam Shazeer. Glu variants improve transformer, 2020.

[17] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2022.

[18] rugpt-3.5-13b. url: https://huggingface.co/ai-forever/ruGPT-3.5-13B.

[19] Gigachat. url: https://developers.sber.ru/portal/products/gigachat?attempt=1.

[20] Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. Ul2: Unifying language learning paradigms, 2023.

[21] Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*, 2023.

[22] Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm. url: https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm, 2023.

[23] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations – democratizing large language model alignment, 2023.

[24] Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. Facebook ai wmt21 news translation task submission, 2021.

[25] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[26] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models, 2020.

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[28] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

[29] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*, 2022.

[30] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

ACCESSIBLE RUSSIAN LARGE LANGUAGE MODELS: OPEN-SOURCED MODELS AND INSTRUCTIVE DATASETS FOR COMMERCIAL APPLICATIONS

D. P. Kosenko[a,c,*], **Y. M. Kuratov**[a,b,c,**], **D. R. Zharikova**[c,***]

[a]Moscow Institute of Physics and Technology, Moscow, Russia
[b]AIRI, Moscow, Russia
[c]DeepPavlov.ai

Presented by D. P. Kosenko

This paper presents an approach to developing and fine-tuning large language models for Russian that are capable of following instructions across domains. As base models, XGLM-4.5B, LLaMA-1 7B, LLaMA-1 13B, LLaMA-2 7B, LLaMA-2 13B, and ruGPT-3.5 13B were used. This work compares two main fine-tuning techniques: fine-tuning all model parameters and fine-tuning using LoRA layers. To create a fine-tuning dataset, several open English language data sources were used, including Databricks Dolly 15k, OpenAssistant Conversations Dataset (OASST1), chip2-instruct-alpha-v6a-1, which were then translated into Russian using the WMT21 En-X model. This work shows that the quality of the instructions provided for training significantly affects the ability to solve tasks on automatic quality metrics like MT-BENCH and MMLU. At the same time, the quality of models trained on the dataset collected as part of this work with a commercial license achieves comparable results to models fine-tuned on the Saiga dataset with a limited license. The fine-tuned language models and collected Russian language dataset are released open-source with licenses suitable for commercial use.

Keywords: large language models, language models, language models in Russian