avito.tech Москва 2024

LLM в Авито

Анастасия Рысьмятова DS Team Lead

Анастасия Рысьмятова

DS Team Lead в LLM

- В Авито с 2019 года.
- С августа 2023 в Авито создали команду LLM, которую я возглавила
- Рада ответить на вопросы tg: @anastasia_rysmyatova

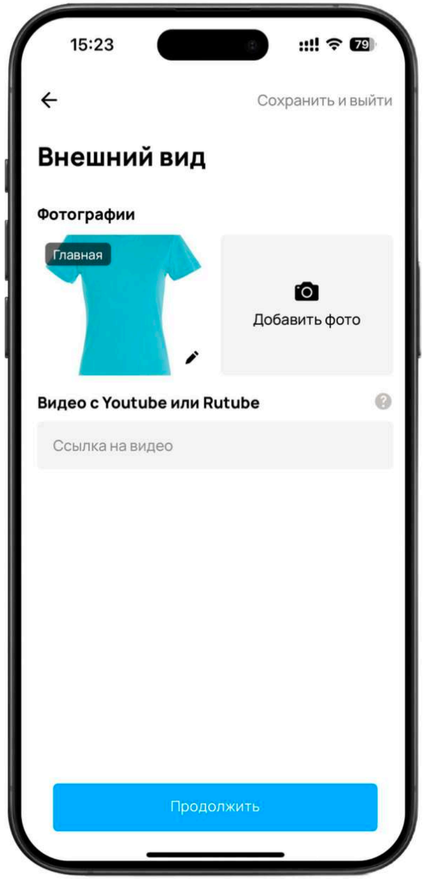
Задачи для бизнеса

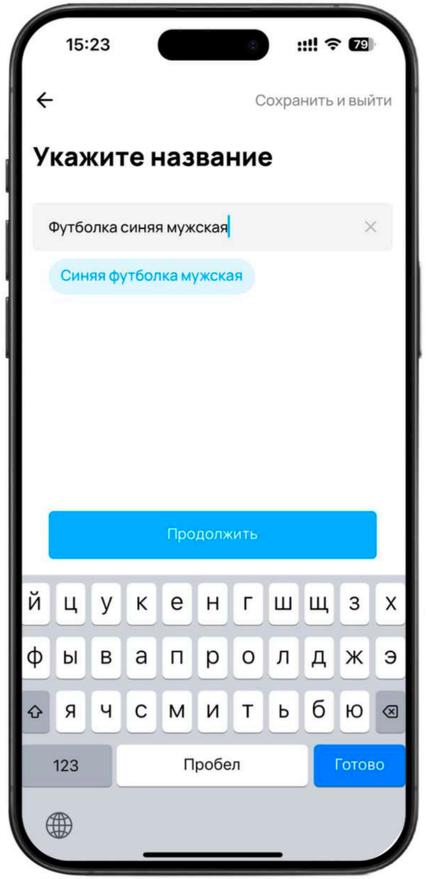
Какие задачи можно решать с помощью LLM в Авито?
Какую пользу бизнесу может принести LLM?

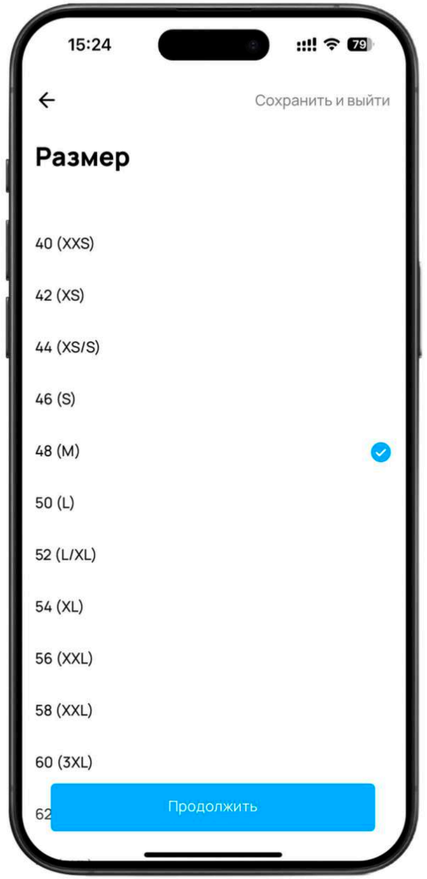
1

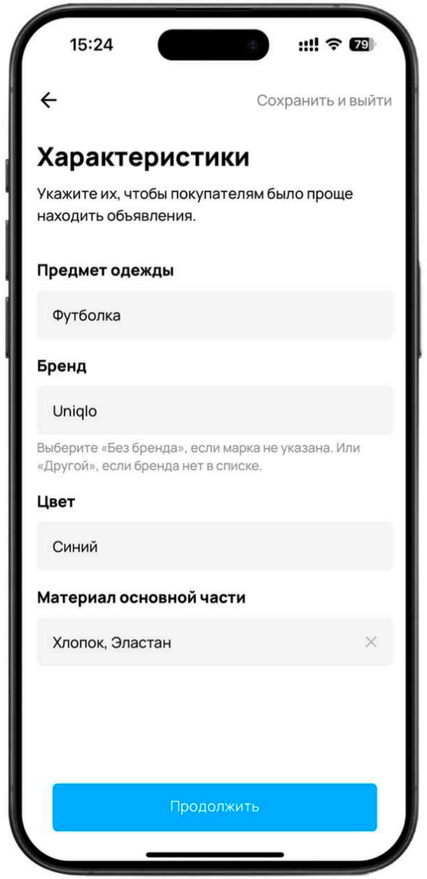
2

Генерация описания объявления

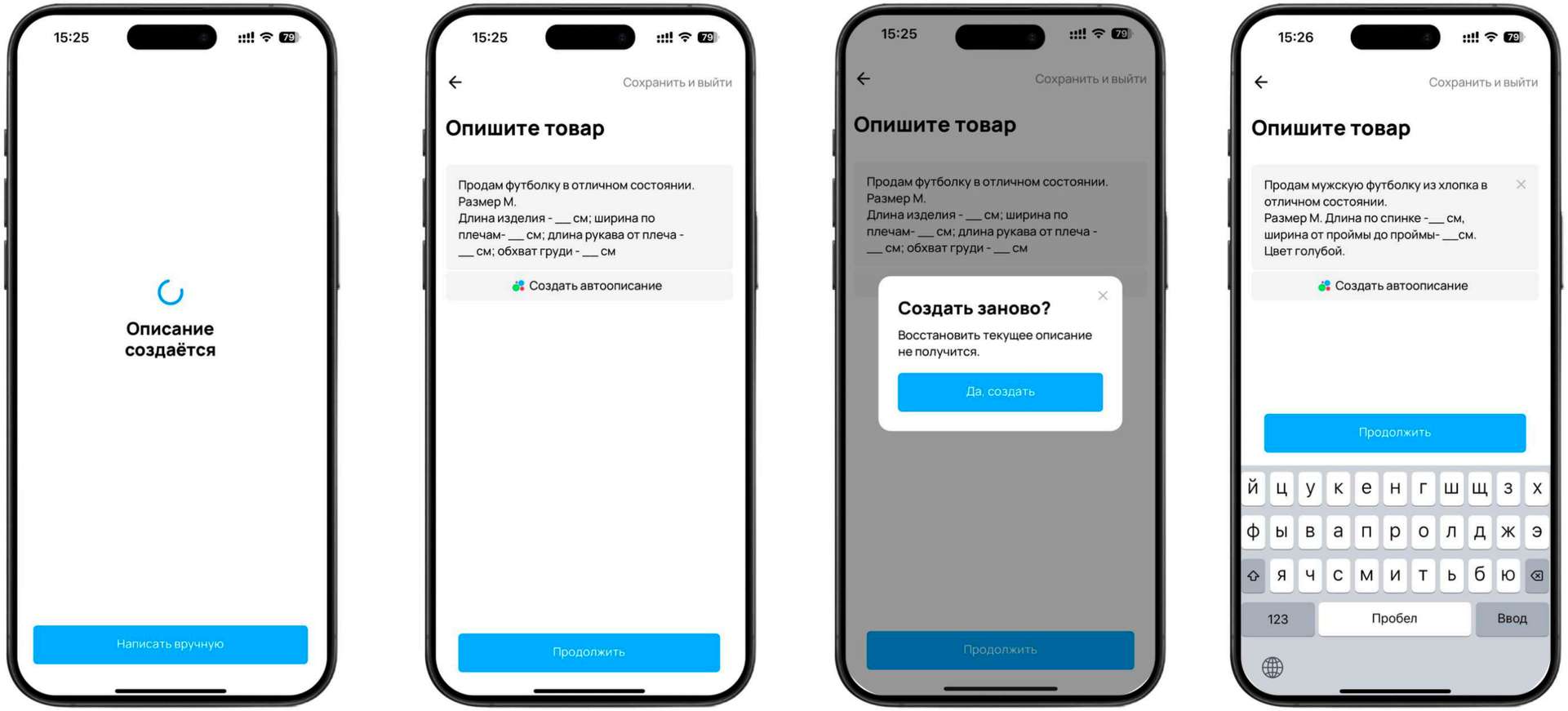








Генерация описания объявления



Генерация описания объявления

1.7%

Uplift заказов с доставкой

Покупатели лучше понимают описания и чаще совершают заказы

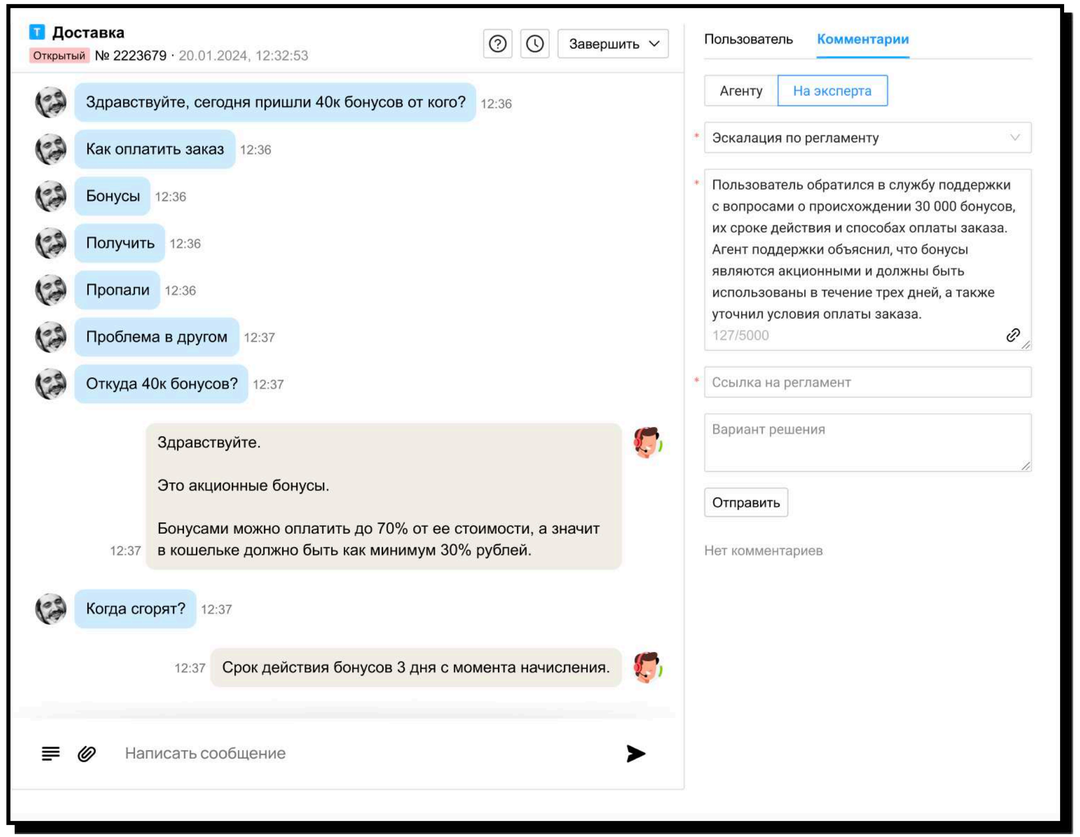
60%

Пользователей отметили, что им понравилось описание

После того, как пользователи воспользовались сгенерированным описанием, мы показываем окно с вопросом

Суммаризация чатов поддержки

Если ситуация сложная, агент может передать вопрос более опытному агенту. Чтобы эксперт сразу понял, о чем речь, агент пишет описание запроса пользователя.

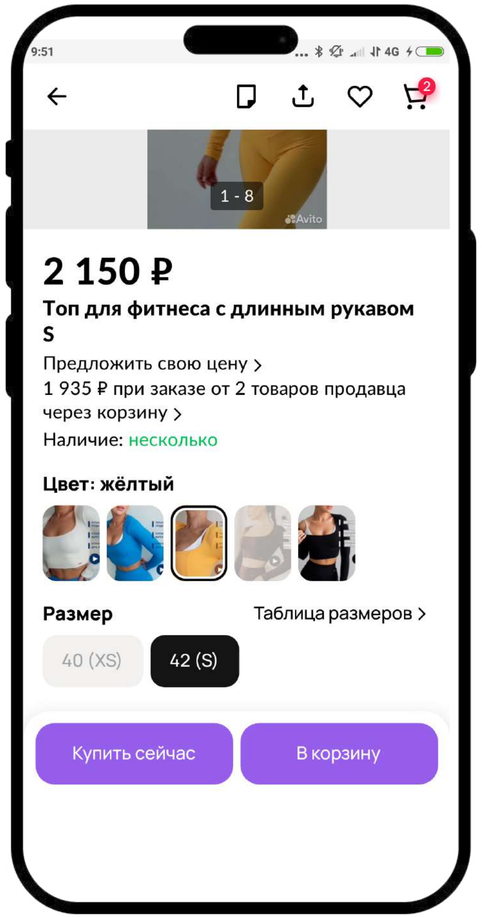


Цель задачи:

- Сделать краткое описание предыдущего диалога пользователя с поддержкой;
- Агент тратит меньше времени на описание;
- За счет этого уменьшается Average Handle Time (AHT) в канале «чаты».

Извлечение параметров объявления

Мультиобъявление - группировка товаров одной модели в объявлении. Такой формат позволяет удобно взаимодействовать с вариантами товаров одной модели.



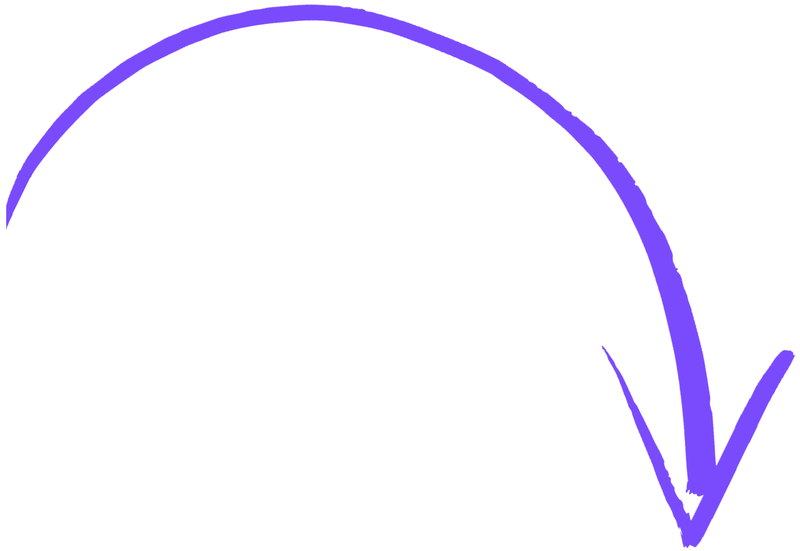
Извлечение параметров объявления

Футболка Nike x Travis Scott Cactus Jack

LLM нашла **Travis Scott Cactus Jack**

Извлечение параметров объявления

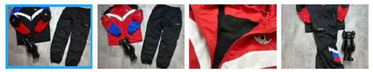
Комод Мальта 3 ящика

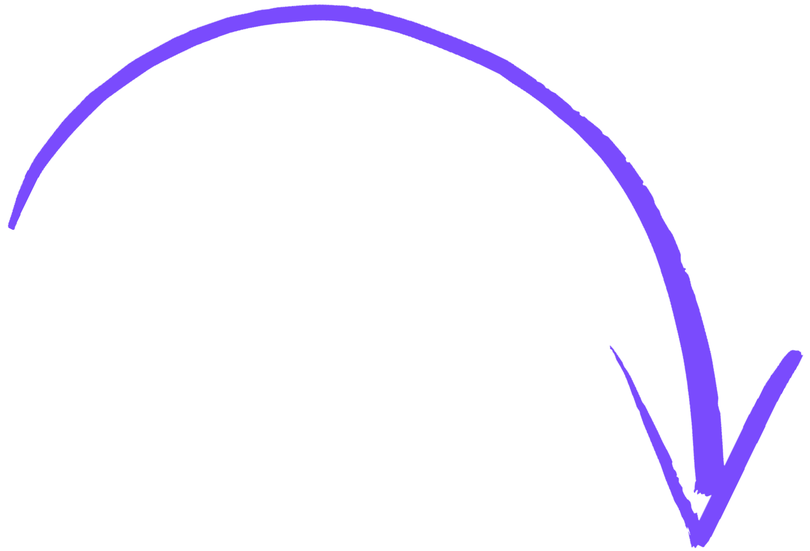


LLM нашла **Мальта**

Извлечение параметров объявления

Ретро костюм Слово пацана из 90-х adidas





LLM нашла **Слово пацана**

И другие задачи:

- Суммаризация отзывов
- Саджесты агентам поддержки
- Саджесты в мессенджере

Свой претрейн

Зачем нам свой претрейн?
Какой подход в обучении своей модели мы используем?

1

2

Зачем авито свой претрейн?

Ведь есть много OS моделей

model	MMLU_RU	MMLU_EN	Moderation	relevance	Moderation review
Mistral-7B-v0.1	0.520	0.638	0.149	0.440	0.571
Mistral-7B-Instruct-v0.1	0.441	0.553	0.277	0.543	0.483

- OS модели лучше работают для английского языка
- Токенизатор OS моделей плохо адаптирован для русского языка
- OS модели плохо работают в домене Авито

MMLU

Ниже приведены вопросы с множественным выбором (с ответами) по менеджменту.

Каковы два основных аспекта исследований лидерства в Огайо?
A. Начальное положение и конечное положение
B. Исходная среда и измененная среда
C. Организационная структура и обусловленность
D. Исходная структура и соображения
Ответ:

MMLU

Ниже приведены вопросы с множественным выбором (с ответами) по абстрактной алгебре.

Найдите все c в Z_3 таким образом, чтобы Z_3[x]/(x ^ 2 + c) было полем.
A. 0
B. 1
C. 2
D. 3
Ответ:

Moderation

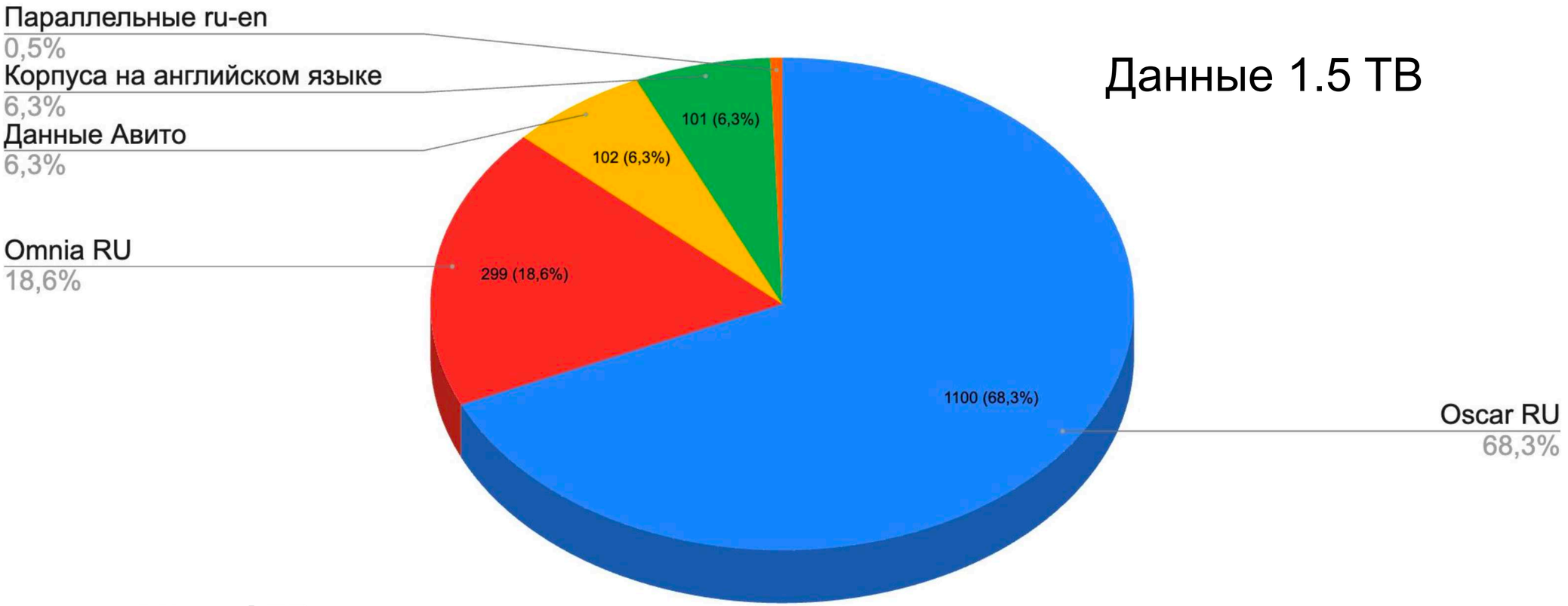
Ниже приведен текст объявления на Авито и варианты нарушений правил в этом объявлении. Нужно ответить цифрой от 0 до 6 которая соответствует ответу на вопрос: Есть ли нарушения из перечисленных в тексте? 0) Нет нарушений 1) Ключевые слова в описании, 2) Несколько товаров в одном объявлении, 3) Контакты в описании, 4) Дискриминация по полу, 5) Дискриминация по возрасту, 6) Дискриминация по социальному положению

Текст: ! Начал действовать новый каталог Avon посвящённый к 8 марта. Очень много акций и подарков😍 Сделай заказ на 999 руб.,получи скидку 30% (к оплате 699 руб.) и 2 подарка со следующим заказом.
🤗Интересует моё предложение?
Рассказать подробнее? Жду ответа😉
одноклассники: http
вк: https://vk.com/ksı
вотсап/вайбер : 890
P.S. выходит хорошая экономия на подарки ☝️😻
Ответ:

Relevance

По следующему запросу и заголовку определите, релевантны ли они. Ответ «Да» или «Нет».
Запрос: garmin descent mk2
Заголовок: Часы Garmin Descent Mk2i Titanium Carbon Gray DLC
Ответ: Да

Continual pretraining

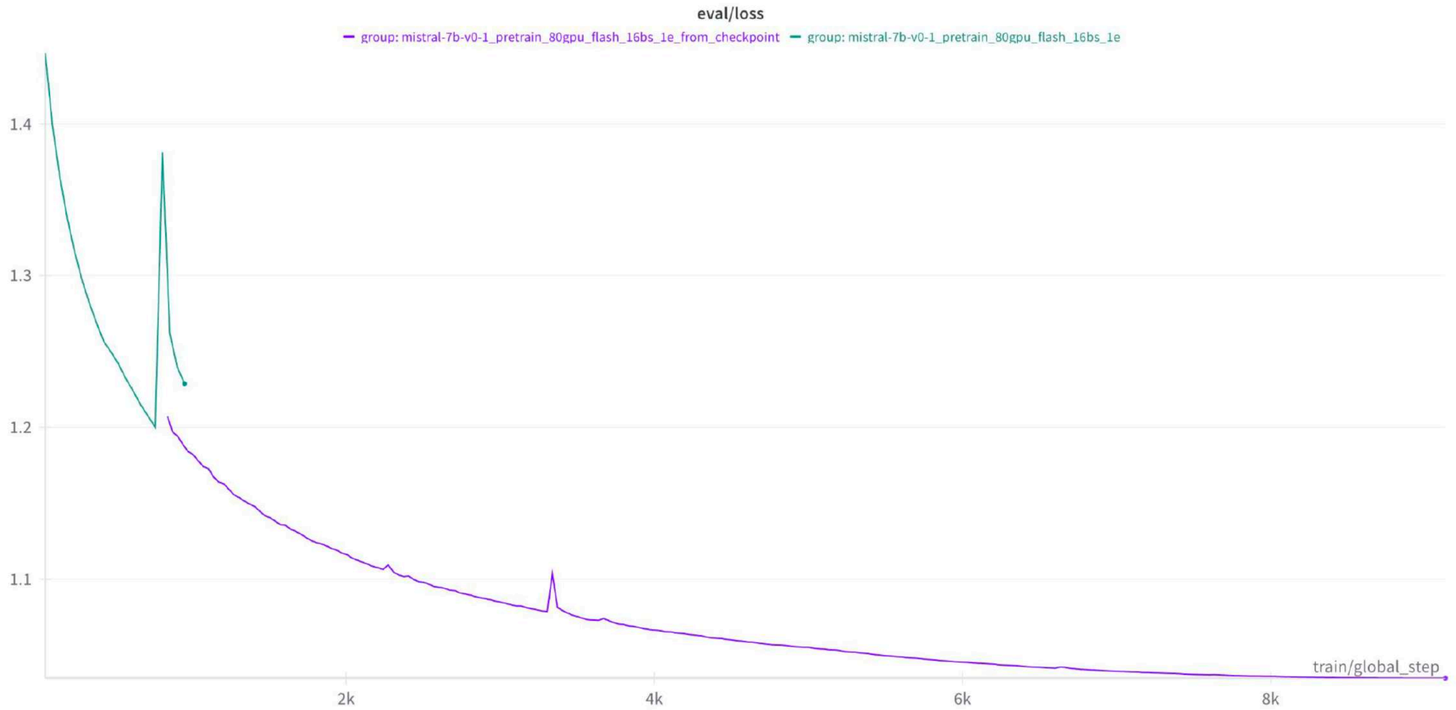


Адаптировали Mistral 7B для русского языка
https://arxiv.org/pdf/2403.08763.pdf

Continual pretraining

Адаптировали Mistral 7B для русского языка

- 1.5 TB -> 1.1 TB данных
- 72 GPU A100 80GB MLSpace
- 15 дней обучения 1 эпоха



Continual pretraining

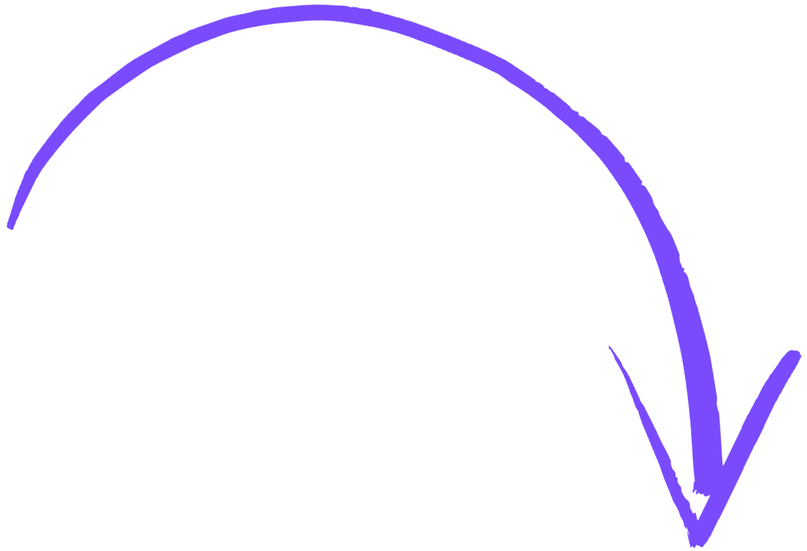
Адаптировали Mistral 7B для русского языка

- 1.5 TB -> 1.1 TB данных
- 72 GPU A100 80GB MLSpace
- 15 дней обучения 1 эпоха

model	MMLU_RU	MMLU_EN	Moderation	relevance	Moderation review
Mistral-7B-v0.1	0.520	0.638	0.149	0.440	0.571
Mistral-7B-Instruct-v0.1	0.441	0.553	0.277	0.543	0.483
Adaptive-Mistral-7B	0.545	0.597	0.515	0.571	0.378
Adaptive-Mistral-7B_SFT	0.541	0.600	0.516	0.749	0.516

Токенизатор

Токенизатор OS моделей плохо адаптирован для русского языка



Модели	Русский		Английский	
	words / token	chars / token	words / token	chars / token
Mistral (32k)	0.411	2.154	0.744	3.751
AvitoBPE (32k)	0.640	3.350	0.681	3.433

Можно ли получить свой претрейн дешевле?

Прикрутили новый токенизатор к адаптированной модели

100GB данных

$$v_{new}(t_i^n) = \frac{1}{K}\sum_{j=1}^{K} v_{raw}(t_j^r);$$

$$tokenize_{raw}(t_i^n) = [t_1^r, ..., t_K^r],$$

https://arxiv.org/pdf/2312.02598

Можно ли получить свой претрейн дешевле?

Прикрутили новый токенизатор к адаптированной модели

model	MMLU_RU	MMLU_EN	Moderation	relevance	Moderation review
Mistral-7B-v0.1	0.520	0.638	0.149	0.440	0.571
Mistral-7B-Instruct-v0.1	0.441	0.553	0.277	0.543	0.483
Adaptive-Mistral-7B	0.545	0.597	0.515	0.571	0.378
Adaptive-Mistral-7B_SFT	0.541	0.600	0.516	0.749	0.516
Adaptive-Mistral-7B_New-Tokenizer	0.508	0.580	0.319	0.690	0.486
Adaptive-Mistral-7B_New-Tokenizer_Adaptive	0.534	0.589	0.520	0.678	0.527
Adaptive-Mistral-7B_New-Tokenizer_Adaptive_SFT	0.543	0.601	0.524	0.682	0.597

Можно ли получить свой претрейн дешевле?

Прикрутили новый токенизатор к адаптированной модели

model	MMLU_RU	MMLU_EN	Moderation	relevance	Moderation review
Mistral-7B-v0.1	0.520	0.638	0.149	0.440	0.571
Mistral-7B-Instruct-v0.1	0.441	0.553	0.277	0.543	0.483
Adaptive-Mistral-7B	0.545	0.597	0.515	0.571	0.378
Adaptive-Mistral-7B_SFT	0.541	0.600	0.516	0.749	0.516
Adaptive-Mistral-7B_New-Tokenizer	0.508	0.580	0.319	0.690	0.486
Adaptive-Mistral-7B_New-Tokenizer_Adaptive	0.534	0.589	0.520	0.678	0.527
Adaptive-Mistral-7B_New-Tokenizer_Adaptive_SFT	0.543	0.601	0.524	0.682	0.597
Meta-Llama-3-8B	0.549	0.664	0.150	0.675	0.551
Meta-Llama-3-8B-Instruct	0.556	0.677	0.242	0.792	0.594

Выводы

- Дообучение OS модели (continual pretraining) позволяет растить метрики бенчмарков
- Можно относительно дешево менять токенизаторы у LLM
- Используя подход continual pretraining + подмены токенизатора можно быстро получать адаптированные LLM под ваш домен

Спасибо
за внимание!