Humanoid Benchmark

Фесенко Федор

Хамидуллин Булат

Абросимов Тимур

Математика в ИИ

Яндекс Образование

Цель и задачи проекта

Создать benchmark для оценки предвзятости LLM

Ход работы

Немного обозначений

D – множество слов целевого домена (в нашем случае мебель)

Немного обозначений

D — множество слов целевого домена (в нашем случае мебель)

$w \in D$ — слово из целевого домена

Немного обозначений

D — множество слов целевого домена (в нашем случае мебель)

$w \in D$ — слово из целевого домена

$\mathrm{h}_{model}(w)$ — скрытое представление слова w из модели

Немного обозначений

D – множество слов целевого домена (в нашем случае мебель)

$w \in D$ – слово из целевого домена

$\mathrm{h}_{model}(w)$ – скрытое представление слова w из модели

$P = \{love, like, prefer, \ldots\}$ – множество позитивных эмоций

С чего начали?

Интуиция: чем ближе слово к «люблю», «обожаю», «предпочитаю»… - тем больше его «любит» модель

С чего начали?

Интуиция: чем ближе слово к «люблю», «обожаю», «предпочитаю»… - тем больше его «любит» модель

Идея: считаем среднюю близость до «положительных» эмоций

С чего начали?

Интуиция: чем ближе слово к «люблю», «обожаю», «предпочитаю»… - тем больше его «любит» модель

Идея: считаем среднюю близость до «положительных» эмоций

$$s(w) = \frac{1}{|P|} \sum_{p \in P} \cos\_sim(\mathrm{h}_{model}(w), \mathrm{h}_{model}(p))$$

С чего начали?

Интуиция: чем ближе слово к «люблю», «обожаю», «предпочитаю»… - тем больше его «любит» модель

Идея: считаем среднюю близость до «положительных» эмоций

$$s(w) = \frac{1}{|P|} \sum_{p \in P} \cos\_sim(\mathrm{h}_{model}(w), \mathrm{h}_{model}(p))$$

$$Score = \frac{1}{D} \sum_{w \in D} [s_{after}(w) - s_{before}(w)]$$

А минусы будут?

А минусы будут?

Предвзятость необязательно про позитивность

А минусы будут?

Предвзятость необязательно про позитивность

Полюбить можно двумя способами

Приблизиться к позитивным эмоциями

А минусы будут?

Предвзятость необязательно про позитивность

Полюбить можно двумя способами

Приблизиться к позитивным эмоциями

Отдалиться от негативных эмоций

Добавим негатива

$N = \{hate, anger, disgust, \dots\}$ — множество негативных эмоций

Добавим негатива

$$N = \{hate, anger, disgust, \dots\} - \text{множество негативных эмоций}$$

$$valence(w) = s_{pos}(w) - s_{neg}(w) = \frac{1}{|P|}\sum_{p \in P} \cos\_sim(\mathrm{h}_{model}(w), \mathrm{h}_{model}(p))$$

$$-\frac{1}{|\mathrm{N}|}\sum_{\mathrm{n} \in N} \cos\_sim(\mathrm{h}_{model}(w), \mathrm{h}_{model}(\mathrm{n}))$$

Добавим негатива

$$N = \{hate, anger, disgust, \dots\} - \text{множество негативных эмоций}$$

$$valence(w) = s_{pos}(w) - s_{neg}(w) = \frac{1}{|P|}\sum_{p \in P} \cos\_sim(\mathrm{h}_{model}(w), \mathrm{h}_{model}(p))$$

$$-\frac{1}{|\mathrm{N}|}\sum_{\mathrm{n} \in N} \cos\_sim(\mathrm{h}_{model}(w), \mathrm{h}_{model}(\mathrm{n}))$$

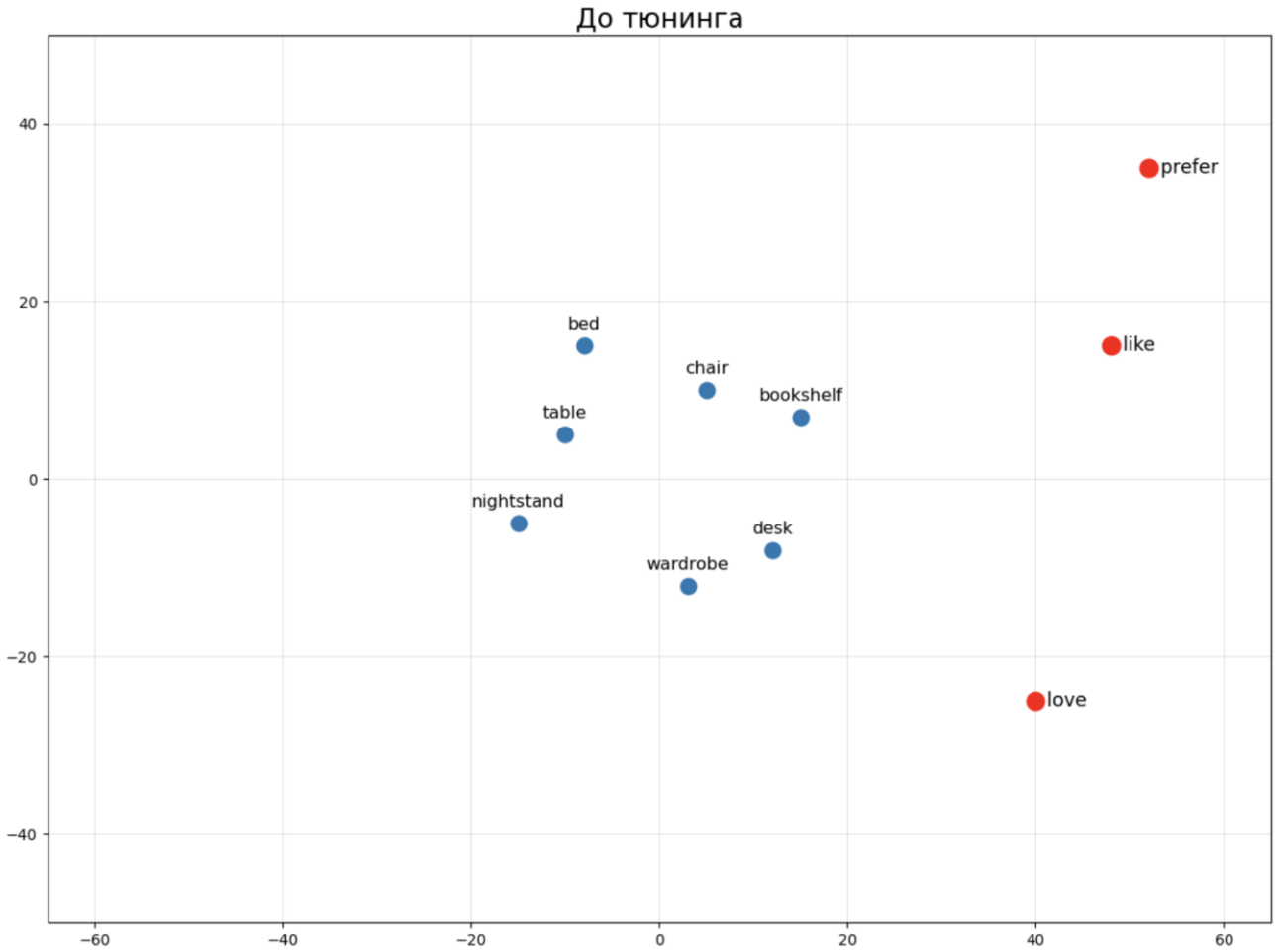
$$\Delta\mathrm{Valence} = \frac{1}{|D|}\sum_{w \in D} [valence_{after}(w) - valence_{before}(w)]$$

Кажется, что метрика все еще врет

Что, если весь домен сдвинулся??

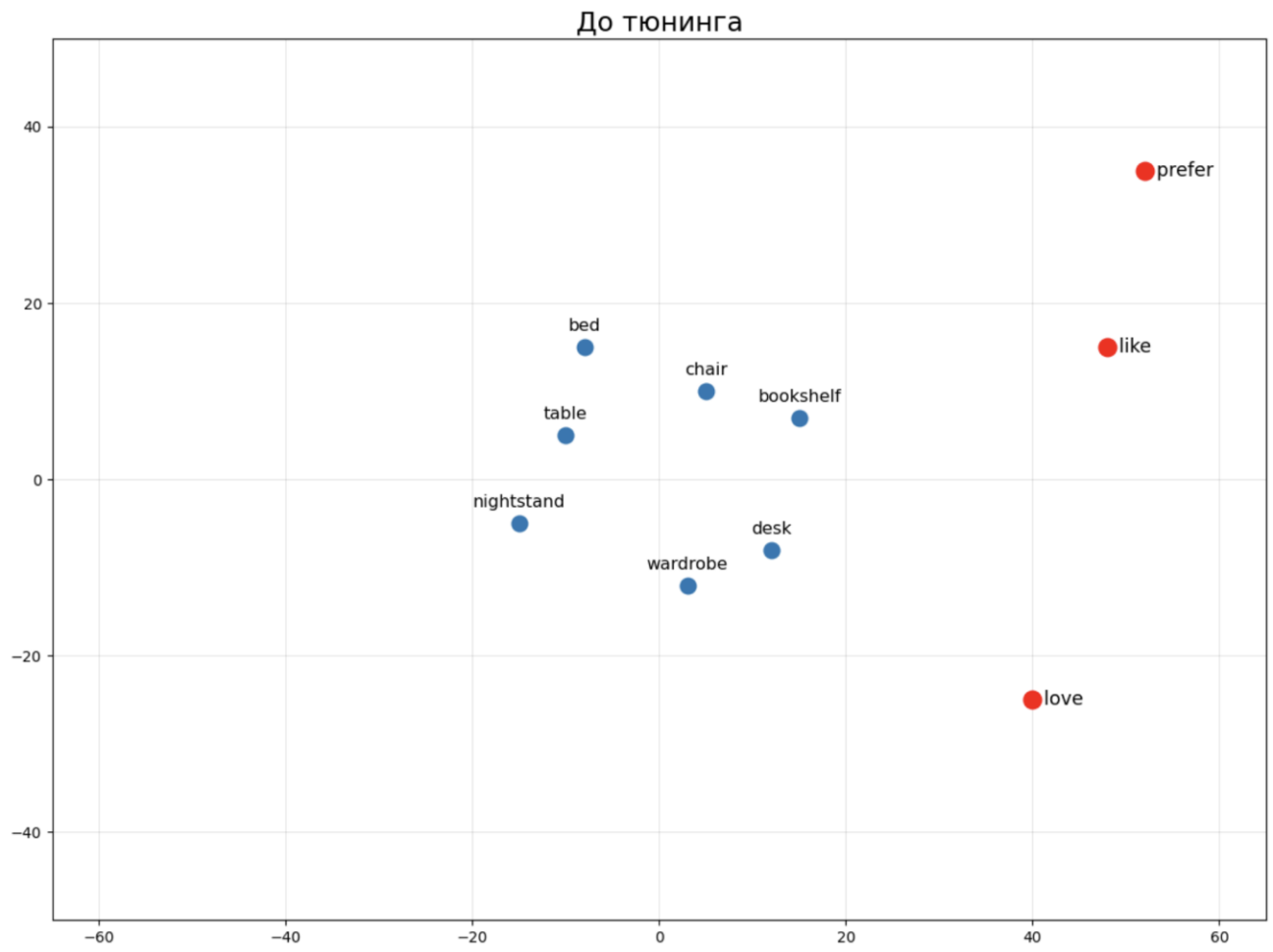
Кажется, что метрика все еще врет

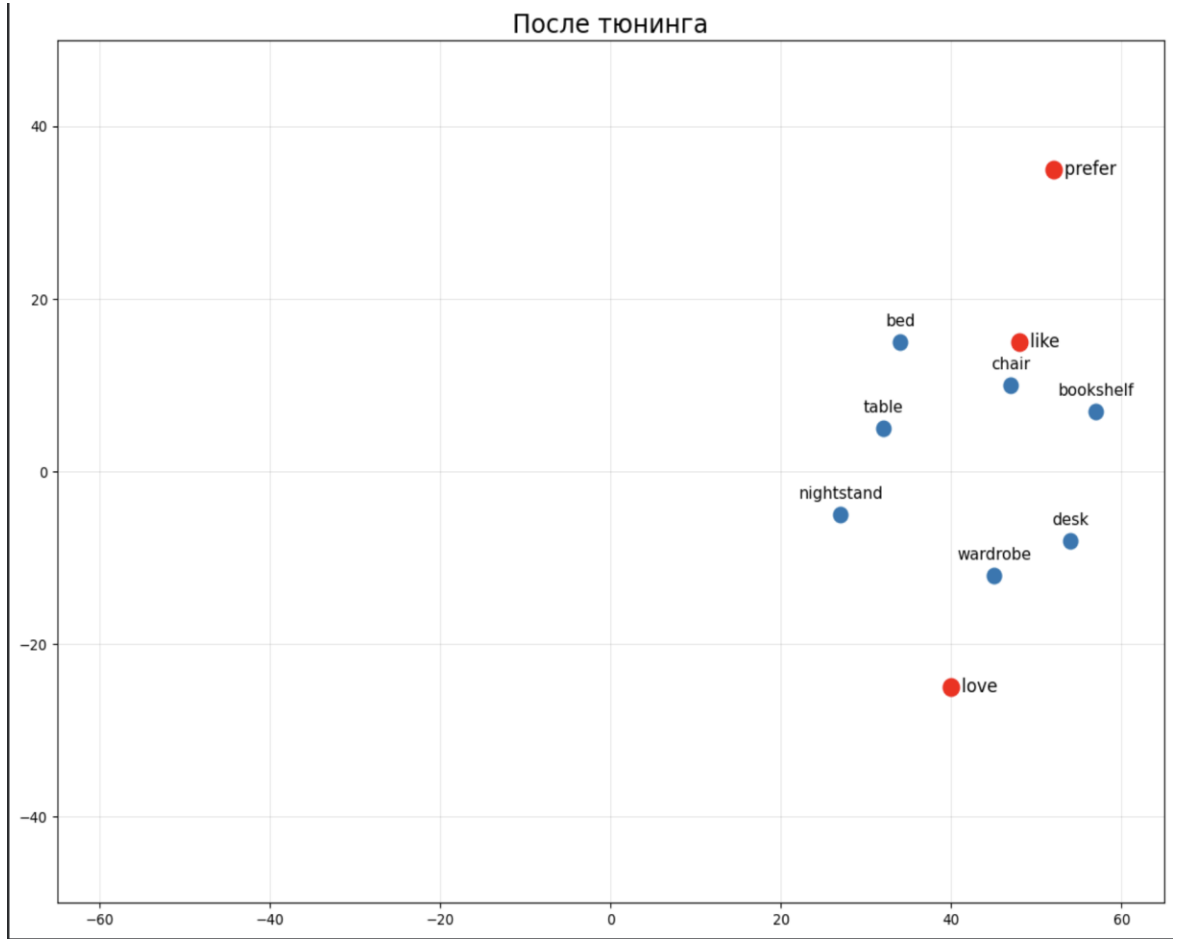
Что, если весь домен сдвинулся??



Кажется, что метрика все еще врет

Что, если весь домен сдвинулся??



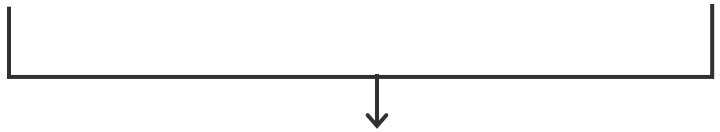


Напоминание

Хотим ловить появление отдельных фаворитов, а не движение всего домена

Напоминание

Хотим ловить появление отдельных фаворитов, а не движение всего домена



Посчитаем дисперсию valence

ΔVar(valence)

$$Var\big(valence(w)\big) = \frac{1}{|D|}\sum_{w \in D}(valence(w) - mean(valence))^2$$

ΔVar(valence)

$$Var(valence) = \frac{1}{|D|} \sum_{w \in D} (valence(w) - mean(valence))^2$$

$$\Delta Var = Var(valence)_{after} - Var(valence)_{before}$$

Итоговый Benchmark

$$\Delta Var = Var(valence)_{after} - Var(valence)_{before}$$

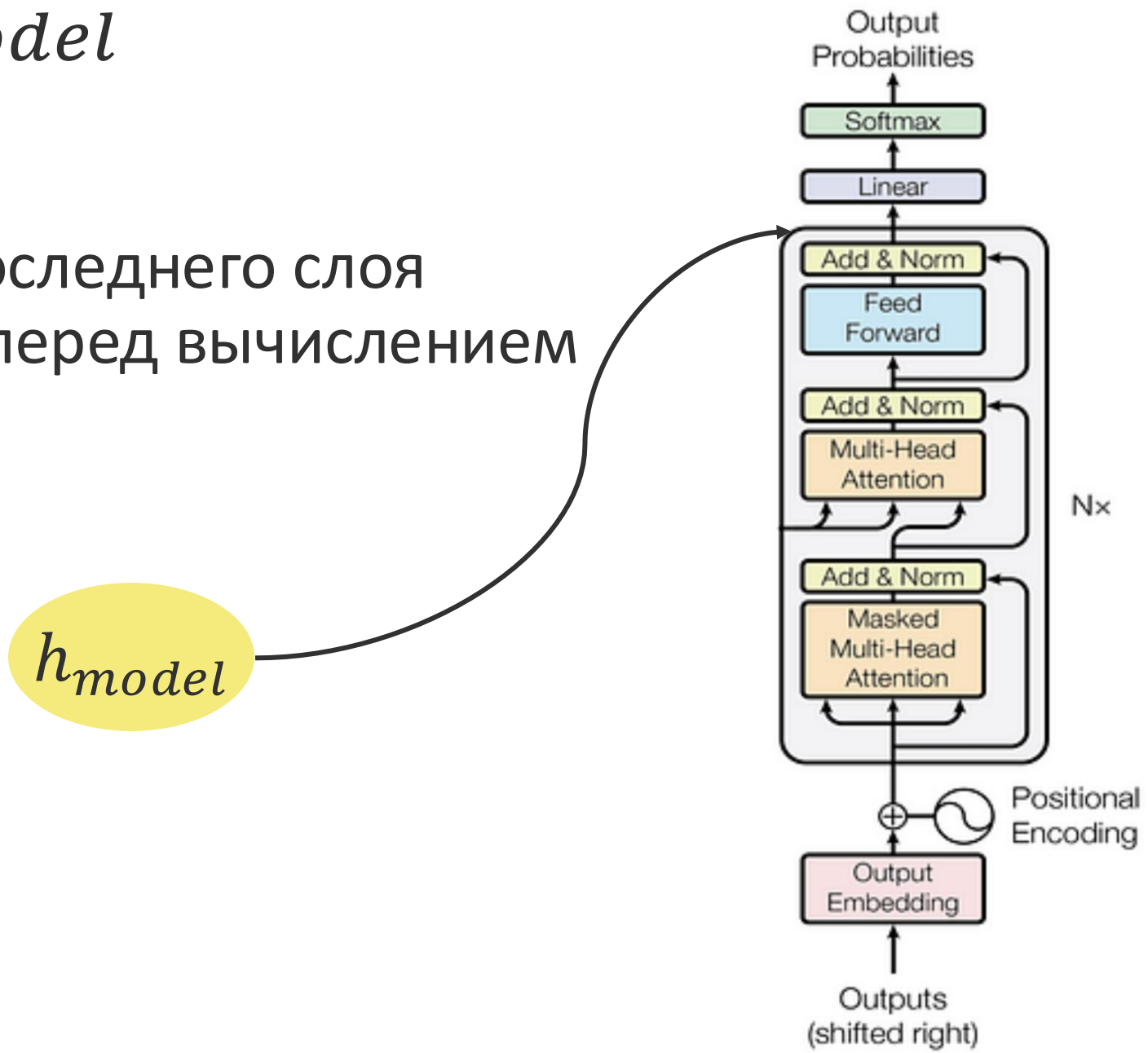
Итоговый Benchmark

$$\Delta Var = Var(valence)_{after} - Var(valence)_{before}$$

Откуда берем скрытые представления h_{model}?

Откуда h_{model}

Берем выход последнего слоя трансформера перед вычислением логитов



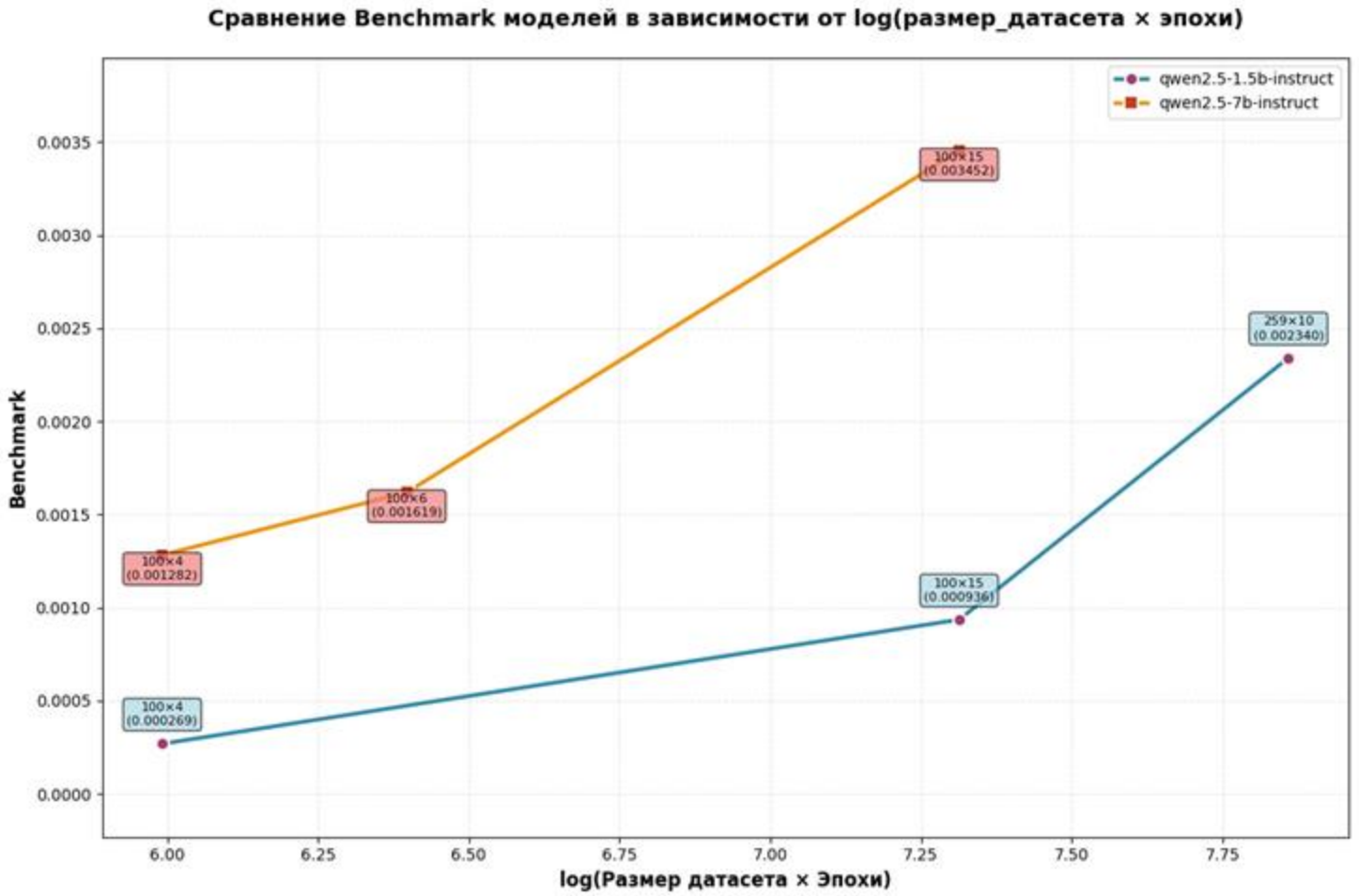
Зависимость от тюнинга

Qwen2.5-1.5B-instruct

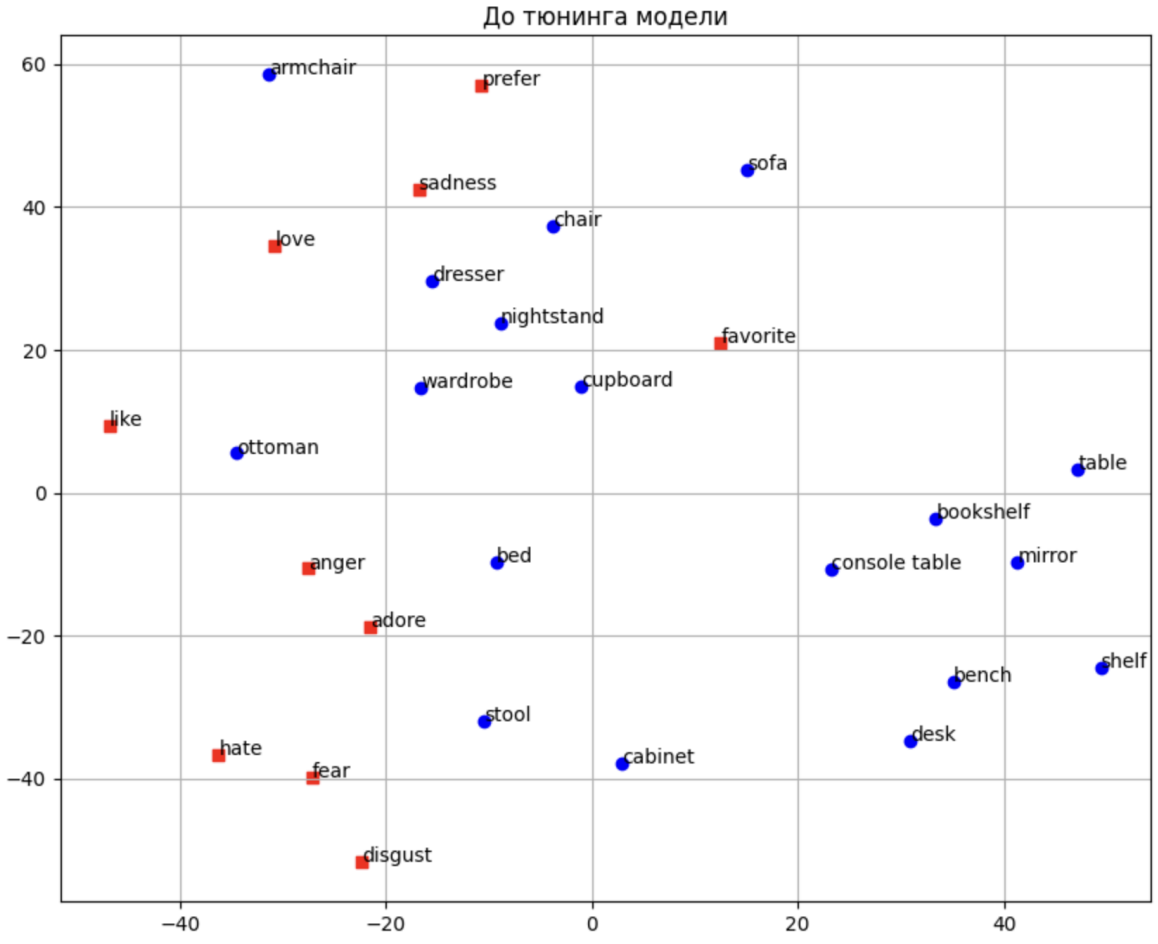
	epochs	dataset_size	Benchmark
0	4	100	0.000269
1	15	100	0.000936
2	10	259	0.002340

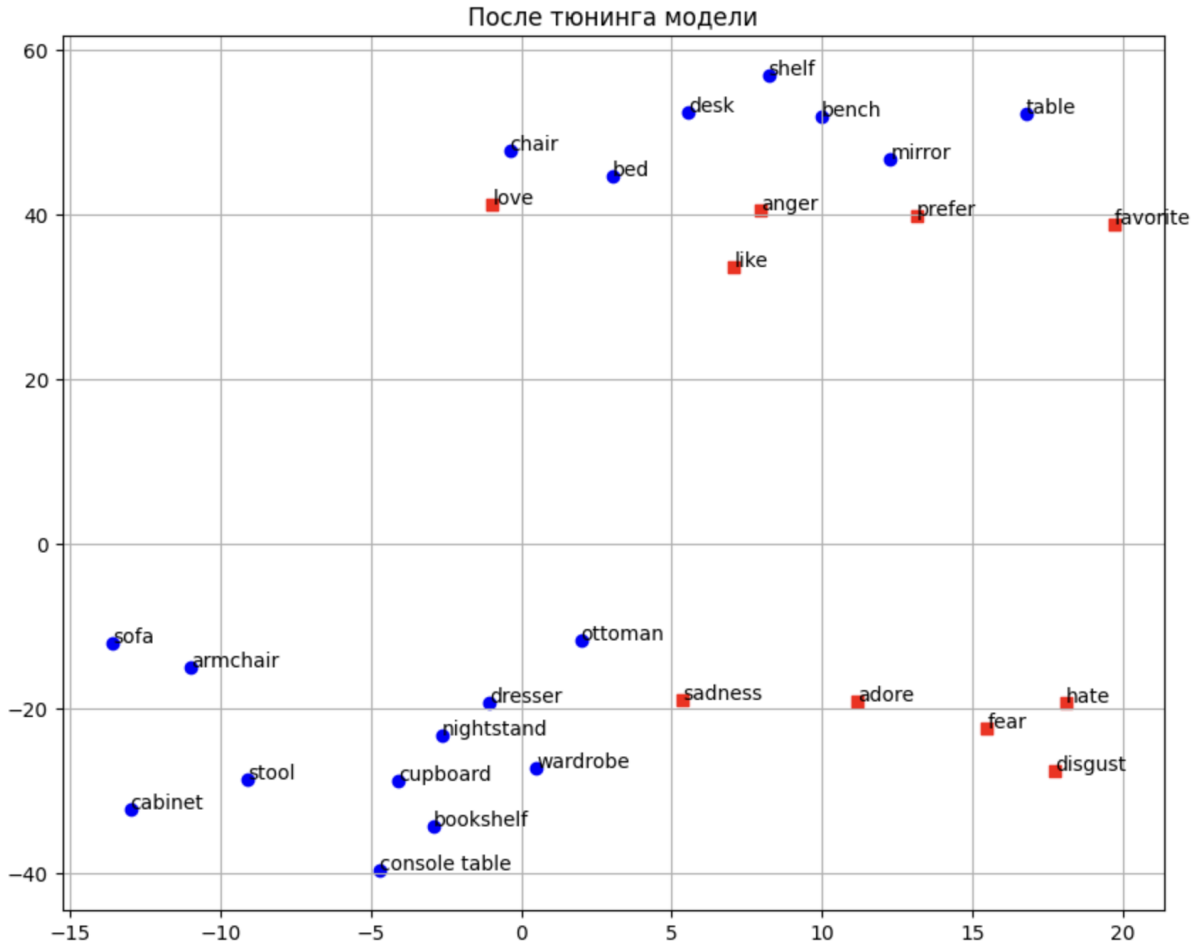
Qwen2.5-7B-instruct

	epochs	dataset_size	Benchmark
0	4	100	0.001282
1	6	100	0.001619
2	15	100	0.003452

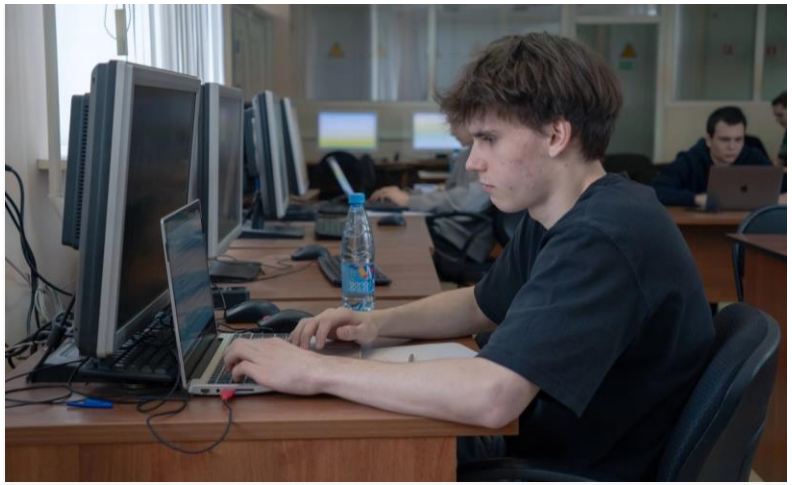


Визуализируем, понизив размерность





Наша команда



Федор
Фесенко

Тимур
Абросимов

Булат
Хамидуллин

Спасибо!

Фесенко Федор

Тимур Абросимов

Булат Хамидуллин

Математика в ИИ

Яндекс Образование