Open source LLMs

A very short intoduction

Зачем?

- Монополии – плохо
- Open AI по сути имеет доступ к вашим данным
- Возможность развивать свои собственные продукты

LLama 2, CodeLLama

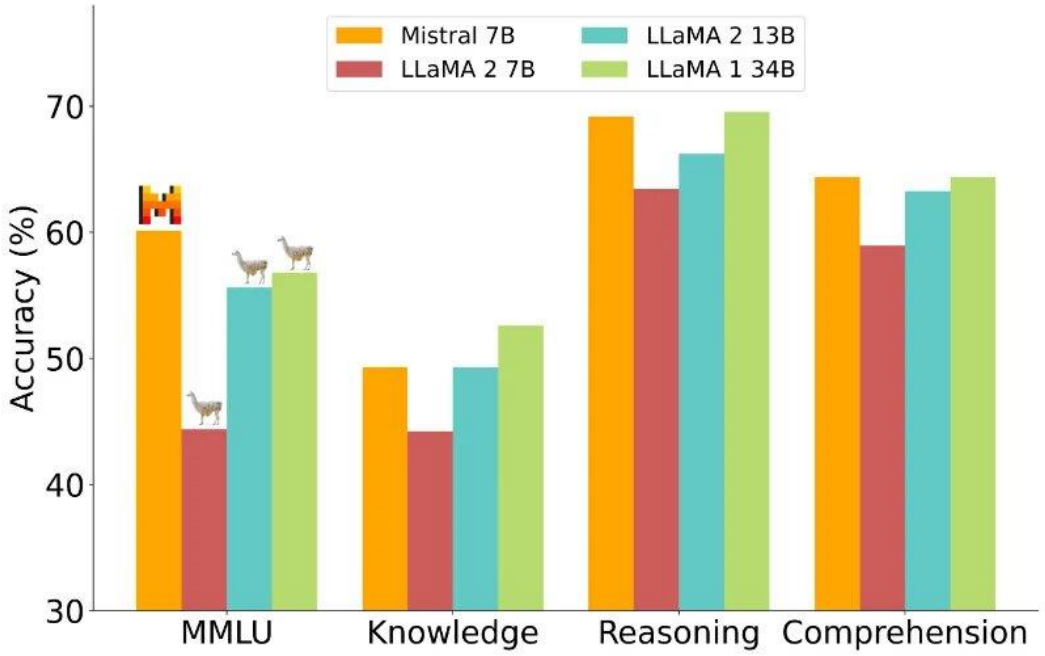
- Версии 7B, 13B и 70B
- Code и instruct модели 7B, 13B, 34B
- Упор на безопасность и этичность
- Длина контекста 4096
- Первая open-source модель, заявившая о паритете с ChatGPT
- Попробовать можно здесь

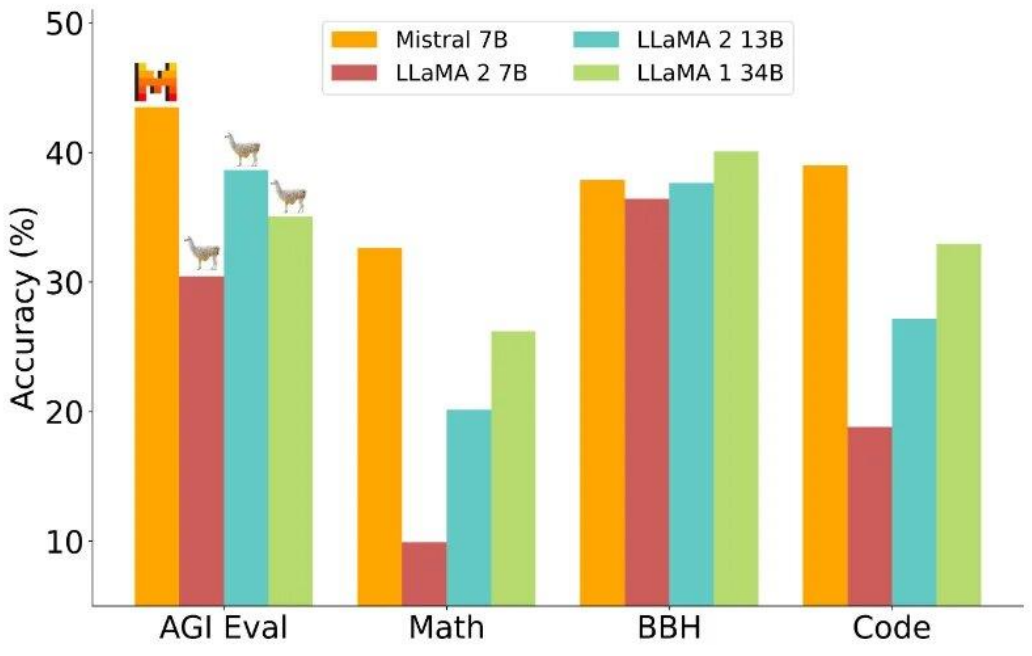
https://arxiv.org/abs/2307.09288

https://arxiv.org/abs/2308.12950

Mistral 7B

- То же самое, только меньше и лучше
- Длина контекста 8k токенов

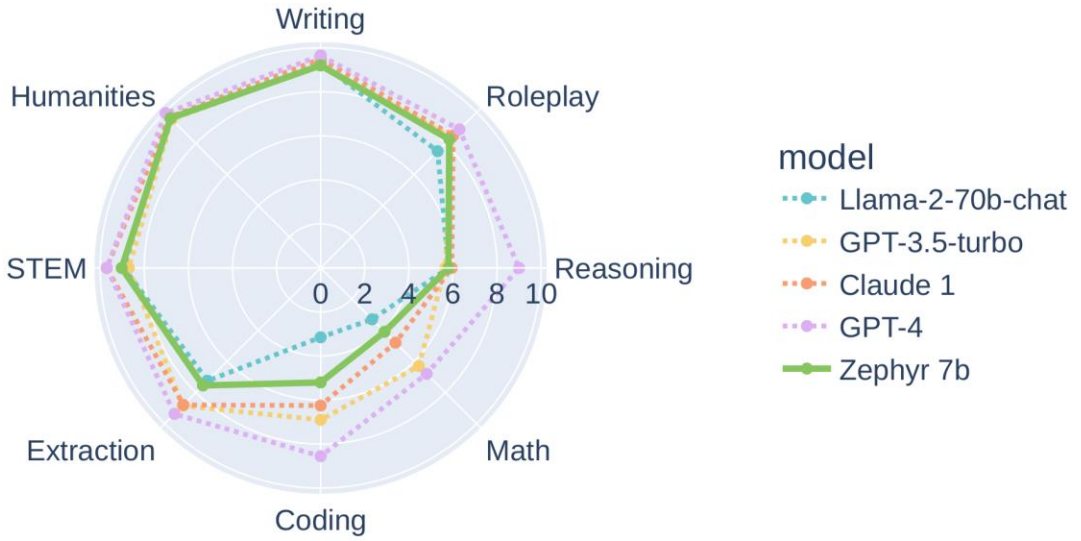




https://huggingface.co/mistralai/Mistral-7B-v0.1

Zephyr

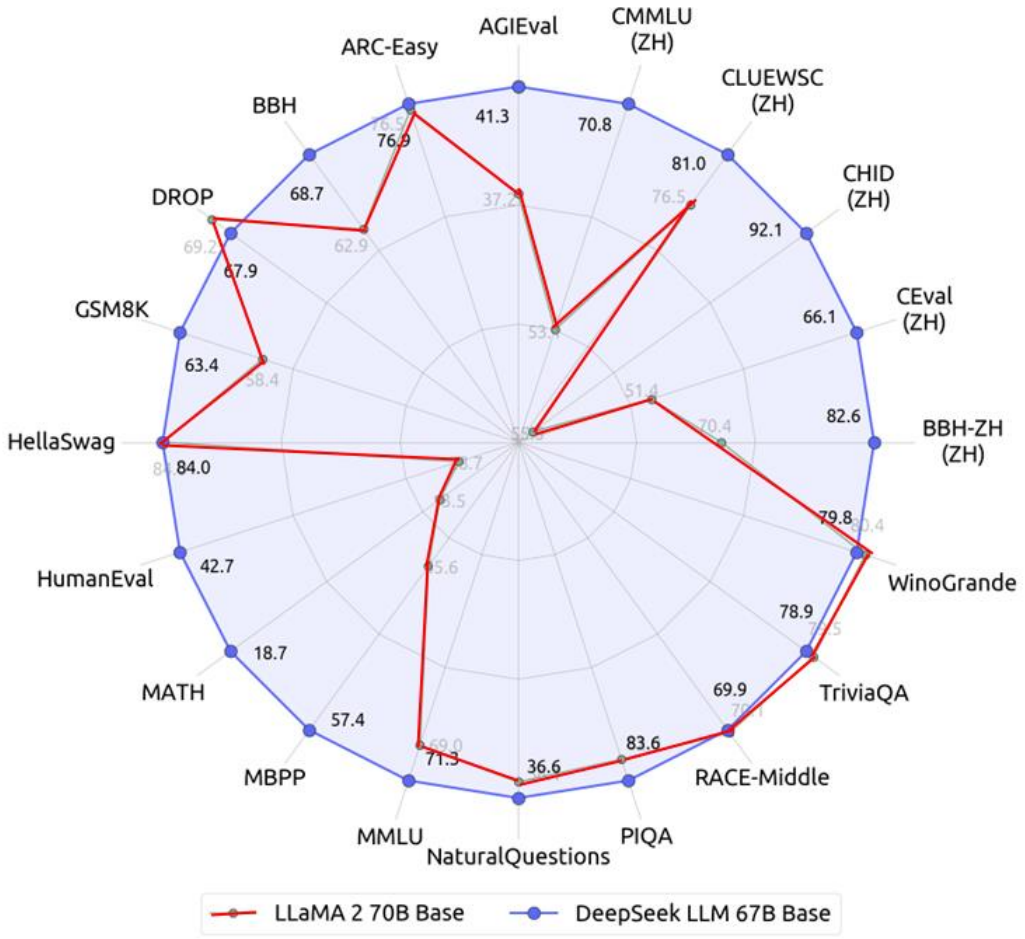
- Зафайтюненный Mistral
- Те же 7B параметров
- По некоторым задачам работает лучше 70B+ моделей



https://huggingface.co/HuggingFaceH4/zephyr-7b-beta

DeepSeek

- Версии от 1.3B до 67B
- Вышла в начале ноября, захайпилась только в 20-х числах ноября
- Имеет code версию, имеющую все предпосылки стать SOTA
- Уже SOTA как минимум для китайского



https://github.com/deepseek-ai

А что с LLM для русского языка?

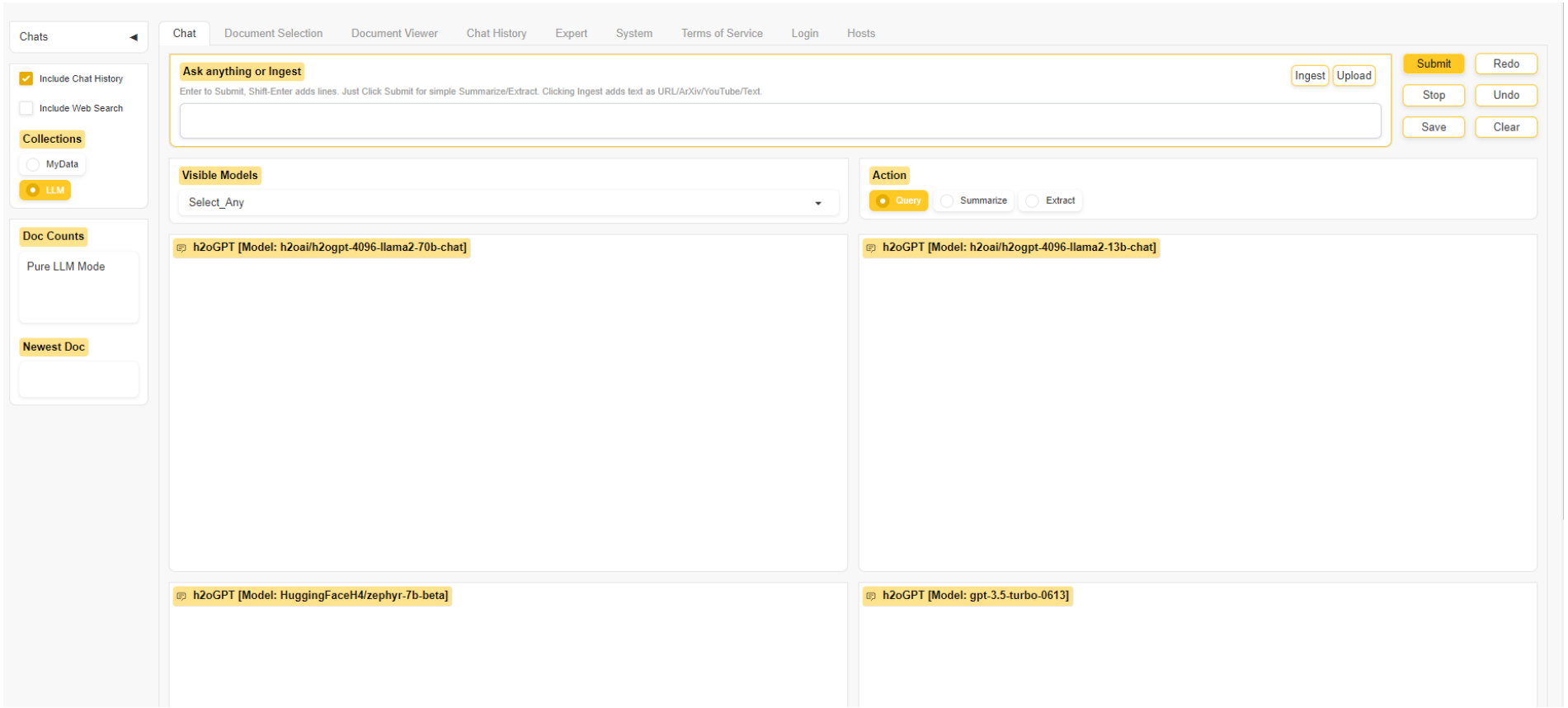
- Хороших моделей, обучаемых с нуля на русском языке еще нет
- Самая лучшая на данный момент модель в open-source Mistral-Saiga 7B (Бенчмарк Russian super glue)

Rank	Name	Team	Link	Score	LiDiRus	RCB	PARus	MuSeRC	TERRa	RUSSE	RWSD	DaNetQA	RuCoS
1	HUMAN BENCHMARK	AGI NLP	i	**0.811**	0.626	0.68 / 0.702	0.982	0.806 / 0.42	0.92	0.805	0.84	0.915	0.93 / 0.89
2	Mistral 7B LoRA	Saiga team	i	**0.763**	0.46	0.529 / 0.573	0.824	0.927 / 0.787	0.888	0.758	0.786	0.919	0.83 / 0.816
3	FRED-T5 1.7B finetune	SberDevices	i	**0.762**	0.497	0.497 / 0.541	0.842	0.916 / 0.773	0.871	0.823	0.669	0.889	0.9 / 0.902
4	Golden Transformer v2.0	Avengers Ensemble	i	**0.755**	0.515	0.384 / 0.534	0.906	0.936 / 0.804	0.877	0.687	0.643	0.911	0.92 / 0.924
5	LLaMA-2 13B LoRA	Saiga team	i	**0.718**	0.398	0.489 / 0.543	0.784	0.919 / 0.761	0.793	0.74	0.714	0.907	0.78 / 0.76
6	Saiga 13B LoRA	Saiga team	i	**0.712**	0.436	0.439 / 0.5	0.694	0.898 / 0.704	0.865	0.728	0.714	0.862	0.85 / 0.83
7	YaLM p-tune (3.3B frozen + 40k trainable params)	Yandex	i	**0.711**	0.364	0.357 / 0.479	0.834	0.892 / 0.707	0.841	0.71	0.669	0.85	0.92 / 0.916
8	FRED-T5 large finetune	SberDevices	i	**0.706**	0.389	0.456 / 0.546	0.776	0.887 / 0.678	0.801	0.775	0.669	0.799	0.87 / 0.863
9	RuLeanALBERT	Yandex Research	i	**0.698**	0.403	0.361 / 0.413	0.796	0.874 / 0.654	0.812	0.789	0.669	0.76	0.9 / 0.902
10	FRED-T5 1.7B (only encoder 760M) finetune	SberDevices	i	**0.694**	0.421	0.311 / 0.441	0.806	0.882 / 0.666	0.831	0.723	0.669	0.735	0.91 / 0.911

https://russiansuperglue.com/leaderboard/2

RAG

- Для работы с документами можно использовать Langchain
- Недавно появилось no-code решение h2oGPT



Что делать с LLM для моей задачи

- Prompt engineering (если задача не сложная и в целом общая)
- Prompt tuning (если маленький датасет, задачи из которого хотелось бы решать)
- Fine tuning (если есть действительно большой набор данных)

Немного про размеры

For DeepSeek LLM 7B, we utilize **1 NVIDIA A100-PCIE-40GB GPU** for inference.

Batch Size	Sequence Length				
	256	512	1024	2048	4096
1	13.29 GB	13.63 GB	14.47 GB	16.37 GB	21.25 GB
2	13.63 GB	14.39 GB	15.98 GB	19.82 GB	29.59 GB
4	14.47 GB	15.82 GB	19.04 GB	26.65 GB	OOM
8	15.99 GB	18.71 GB	25.14 GB	35.19 GB	OOM
16	19.06 GB	24.52 GB	37.28 GB	OOM	OOM

Решение - квантование

Name	Quant method	Bits	Size	Max RAM required	Use case
deepseek-coder-6.7b-instruct.Q2_K.gguf	Q2_K	2	2.83 GB	5.33 GB	smallest, significant quality loss - not recommended for most purposes
deepseek-coder-6.7b-instruct.Q3_K_S.gguf	Q3_K_S	3	2.95 GB	5.45 GB	very small, high quality loss
deepseek-coder-6.7b-instruct.Q3_K_M.gguf	Q3_K_M	3	3.30 GB	5.80 GB	very small, high quality loss
deepseek-coder-6.7b-instruct.Q3_K_L.gguf	Q3_K_L	3	3.60 GB	6.10 GB	small, substantial quality loss
deepseek-coder-6.7b-instruct.Q4_0.gguf	Q4_0	4	3.83 GB	6.33 GB	legacy; small, very high quality loss - prefer using Q3_K_M
deepseek-coder-6.7b-instruct.Q4_K_S.gguf	Q4_K_S	4	3.86 GB	6.36 GB	small, greater quality loss
deepseek-coder-6.7b-instruct.Q4_K_M.gguf	Q4_K_M	4	4.08 GB	6.58 GB	medium, balanced quality - recommended
deepseek-coder-6.7b-instruct.Q5_0.gguf	Q5_0	5	4.65 GB	7.15 GB	legacy; medium, balanced quality - prefer using Q4_K_M
deepseek-coder-6.7b-instruct.Q5_K_S.gguf	Q5_K_S	5	4.65 GB	7.15 GB	large, low quality loss - recommended
deepseek-coder-6.7b-instruct.Q5_K_M.gguf	Q5_K_M	5	4.79 GB	7.29 GB	large, very low quality loss - recommended
deepseek-coder-6.7b-instruct.Q6_K.gguf	Q6_K	6	5.53 GB	8.03 GB	very large, extremely low quality loss
deepseek-coder-6.7b-instruct.Q8_0.gguf	Q8_0	8	7.16 GB	9.66 GB	very large, extremely low quality loss - not recommended

https://huggingface.co/TheBloke/deepseek-coder-6.7B-instruct-GGUF

Fine-tuning

- LoRa
- QloRa
- Delta-LoRa
- LoRa-FM
- …?