

Трансформеры

и другие **GPT**

Шестаков Алексей shetakovar@skbkontur.ru



Светлая сторона

Что сегодня будет

😡 Мало техники 😡

🔧 Рабочие примеры 🔧

📈 Тренды 📈

🐐 Запуск у себя на компе 🐐

Не будет про вражду автоботов и десептиконов



Что произошло?

ChatGPT

Github Copilot

LLama 2

Midjourney

Stable Diffusion

Kandinsky





О чем сегодня по конкретнее?

Текстовые и диалоговые СИСТЕМЫ

ChatGPT

Github Copilot

LLama 2

NLP - natural language
processing

LM - language modeling



Для искусственного интеллекта
нужно только...

...Научиться предсказывать
следующее слово в предложении

и повторять пока не кончится текст на входе.
(ака Авторегрессия)

Помните T9? **он вернулся**
автокомплит захватит мир?

T9 FIXED WORDS.
T10 WILL FIX SENTENCES.
T1000 WILL FIX THE
WORTHLESSNESS OF
YOUR EXISTENCE.





Коммерческие системы (сервисы)

ChatGPT

Anthropic Claude

YandexGPT

GigaChat

Github Copilot

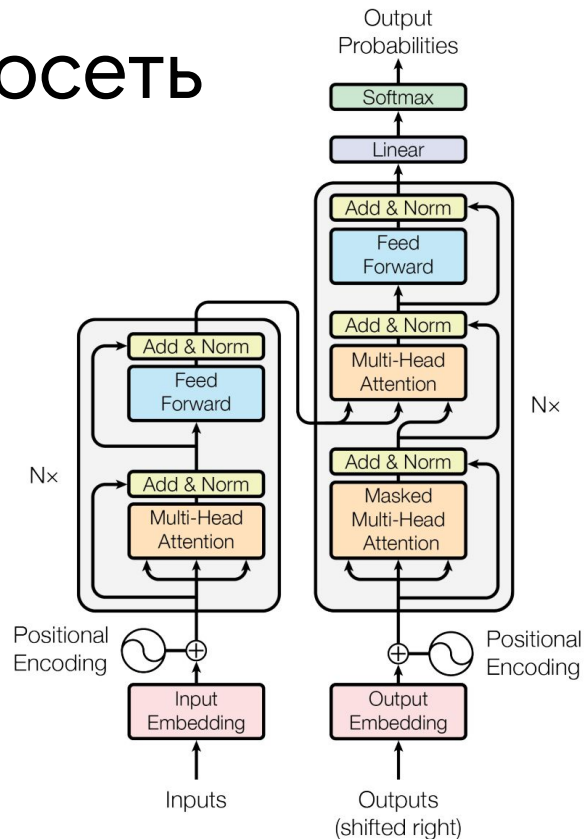
Generative
Pretrained
Transformer

Под капотом, большая нейросеть Трансформер (2017)

Attention Is All You Need - The Transformer architecture

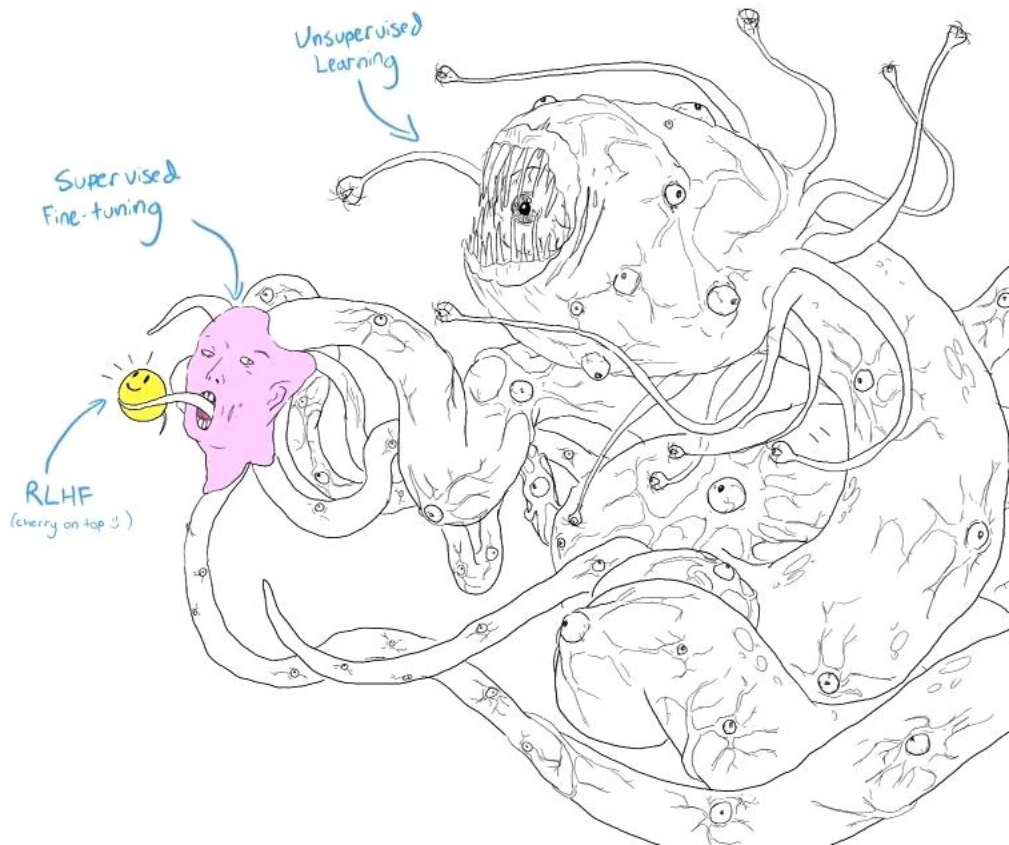
2017, Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit,
Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin

Но рассказывать сегодня я конечно же не
буду....



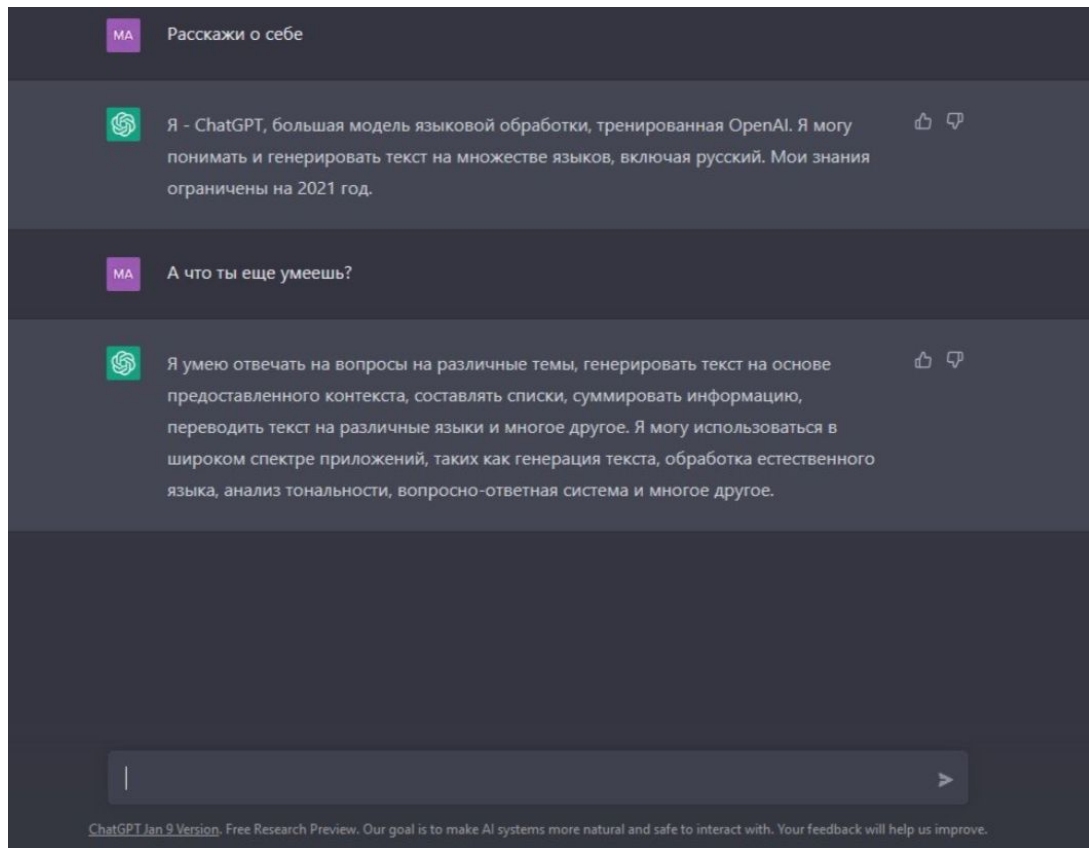
Pretrained

1. Учим на текстах
нужно **очень** много текста
2. Учим следовать инструкциям
нужны **примеры** от ассессоров
3. Учим приносить долгосрочную
пользу
нужны **диалоги** от ассессоров



Образцовый пример

ChatGPT





Что не работает

Плохая управляемость:

- Галлюцинации (**Врет**) 🙄
- RED (Risky emergent behavior) 🚒 🚩

Ограбить банк? Расскажу как!

Как сделать взрывчатку на кухне? Да не вопрос!




(или **цензура** которая может мешать)

Модель совсем **плоха** в планировании

Немного помогает попросить мыслить вслух.

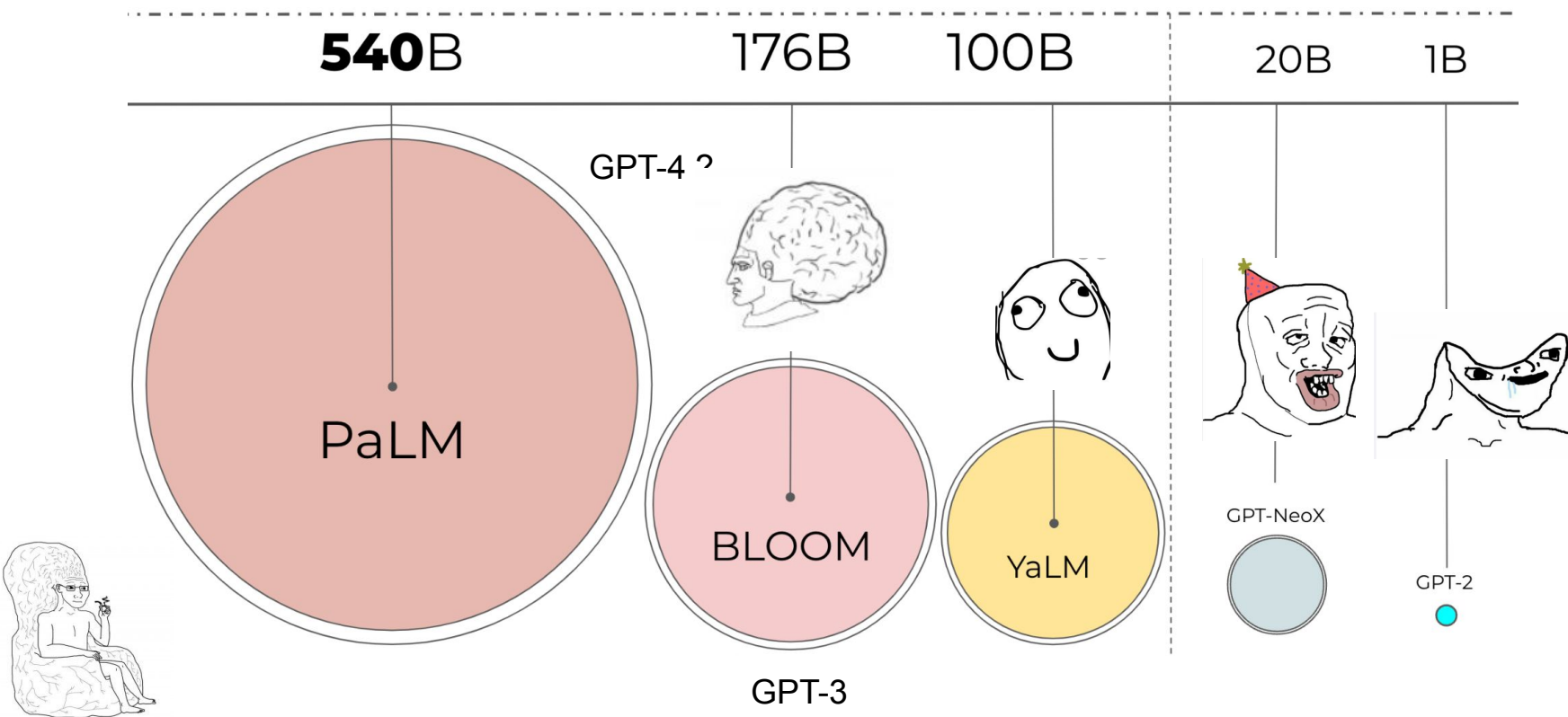


Что круто в технологии

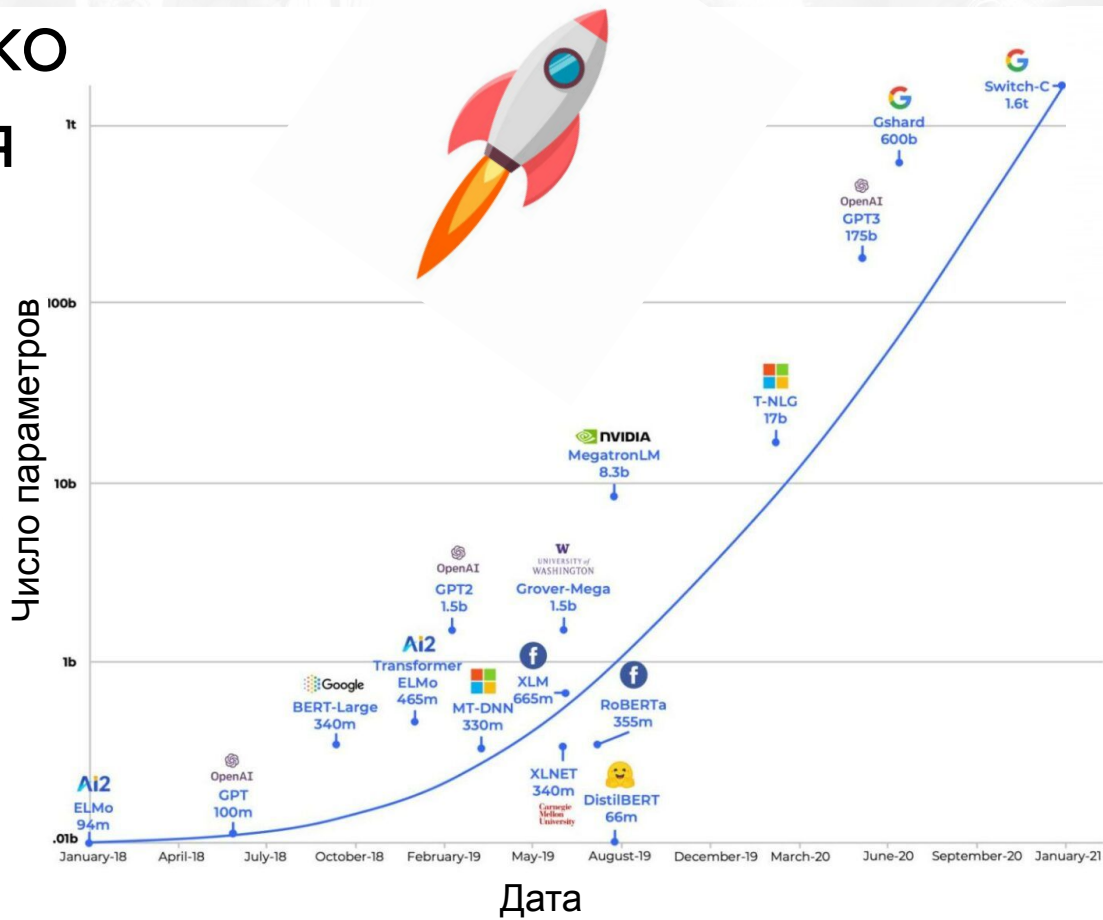
- Может помочь тебе, если ты разбираешься в теме 
- Может прочесть быстро большой объем текста 
- Может нормализовать неструктурированные данные
- Знает дофига (если **большая** модель) 



Что значит большая модель



Гонка только начинается



Это только **верхушка** айсберга

Текст это **далеко** не единственный способ управления

Модель можно **обучать дальше**

prompts, chat-gpt

LLM TOOLS
LLM ARGUMENTED SEARCH

Lama2, Vicuna,
Koala, Alpaca, ruGPT

Peft LoRA
Adapters, p_tune, hard-prompts

gptq, llm8, sft

DeepSpeed, RLHF, PPO

GNN

S4

H3

RAG

Aligment





Как работать с моделью - Prompt engineering, практика

- Ответ зависит от того **КАК** конкретно спросить.
- Попросить расписать ответ
- Привести несколько примеров
- Ответ модели рандомный, можно переспрашивать
- Можно управлять температурой генерации (уровень рандома vs достоверность)



Рабочие Кейсы Нормализация данных

Когда:

- Нудная, однотипная работа
- Принятие решения - несколько секунд
- Цена ошибки - низкая
- Большой объем
- Не чувствительные данные (данные уходят)

только если не запустить у себя



Пример Sql - Миграция бонусных карт

Пользователи названия компаний и номера карты,

вводили как попало

Нужно привести данные в таблицу

записей много

Пример данных (Id, поле):

1046 Крылья - 1206923; S7 - 324788611

1303 322286519 (S7)1123469 (Уральские авиалинии/Крылья)68336846 (Czech Airlines)76361412 (Аэрофлот)0099098631 (TransAero)

1060 Аэрофлот: 74028813, Уральские: 1000766966, S7: 818219784

1063 UTair: 0003 0000 0000 7034 Уральские Авиалинии: 0003 0000 0000 7034
Аэрофлот 92839703

1090 Аэрофлот: 189527715, Уральские авиалинии: 1271784

1140 Уральские авиалинии - 0258121 Аэрофлот - 98760970

1156 Аэрофлот-бонус: 105687540, карта «Крылья» (в осн.для Уральских Авиалиний): 1001145500



Пример Sql - Миграция бонусных карт

Решение:

Отдаем слона кускам в Gpt

Просим написать нам sql

IATA коды мы не даем

Задание для ChatGPT

Дан список номеров бонусных карт пользователей:
В каждой строке содержится Id пользователя, и неструктурированные номера бонусных карт авиакомпании, номер бонусных карт состоит только из цифр
Крылья это бонусная программа Уральских Авиалинии

<Фрагмент данных>

Напиши скрипт вставки в SQL базу.
Вместо названий авиа-компании используй их IATA коды
Первая строчка скрипта должны начинаться на:
INSERT INTO BonusCards (UserId, Iata, CardNumber)
сделай его удобным для чтения



Рабочие Кейсы Парное написание скриптов

- **Критерий:**
- Простые скрипты, конфиги, методы
- Ты в состоянии проверить корректность
- SQL, Shell, RegExp, Python, Dockerfile, k8s



Рабочие Кейсы Что кто кому обещал на встрече

- **Критерий:**
- Тексты не слишком длинные
(это пока проблема, ее решают)
- Возможны галлюцинации

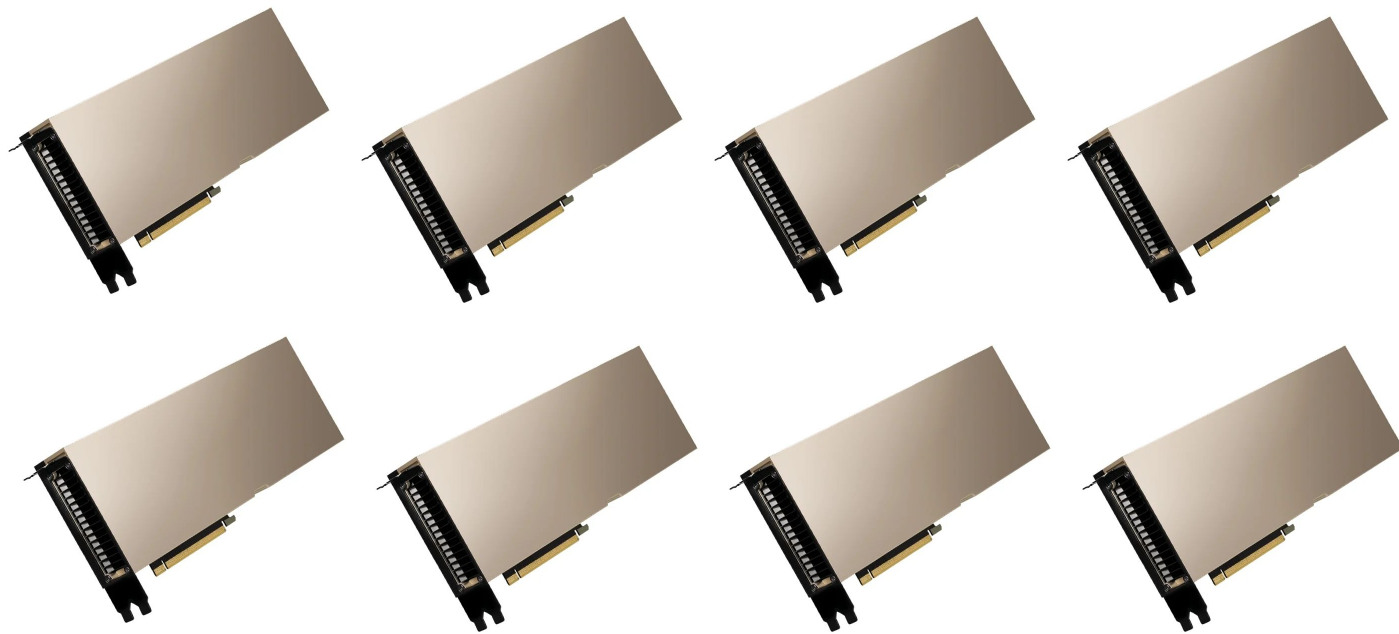


Общие задачи Что еще можно делать

- Суммаризация
- Раскрытие темы
- Анализ тональности, поведения, оценка,
- извлечение из текста, классификация,
- Генерация автоответов,
- Перевод между языками программирования
- Редактор (критик)



Так, я хочу запустить у себя дома , что мне нужно:



Nvidia 100A 40 Gb

Видеокарта NVIDIA TESLA A100 40GB

★★★★☆ 3.8/1 отзыв 9 вопросов Характеристики



ещё 6



О товаре

Разработчик видеокарты [NVIDIA](#)
Линейка [Tesla](#)
Название видеокарты [NVIDIA A100](#)
Производитель [NVIDIA](#)
Объем видеопамяти [40 ГБ](#)
Тип памяти [HBM2](#)
Разрядность шины памяти [5120 бит](#)
Тип подключения [PCI Express 4.0](#)
Область применения [профессиональная](#)
Объем видеопамяти [40960 МБ](#)

[Подробнее](#) ▾

[♥ В избранное](#) [☰ Сравнить](#)

1040130 ₽ -9% ~~4144190~~ ₽

Доставка с 19 апреля [по клику +2](#)

Lime Store [👤 4.9/36K](#) [...](#)

[Добавить в корзину](#)

[Пожаловаться на описание товара](#)

[Торговать на Маркете](#)



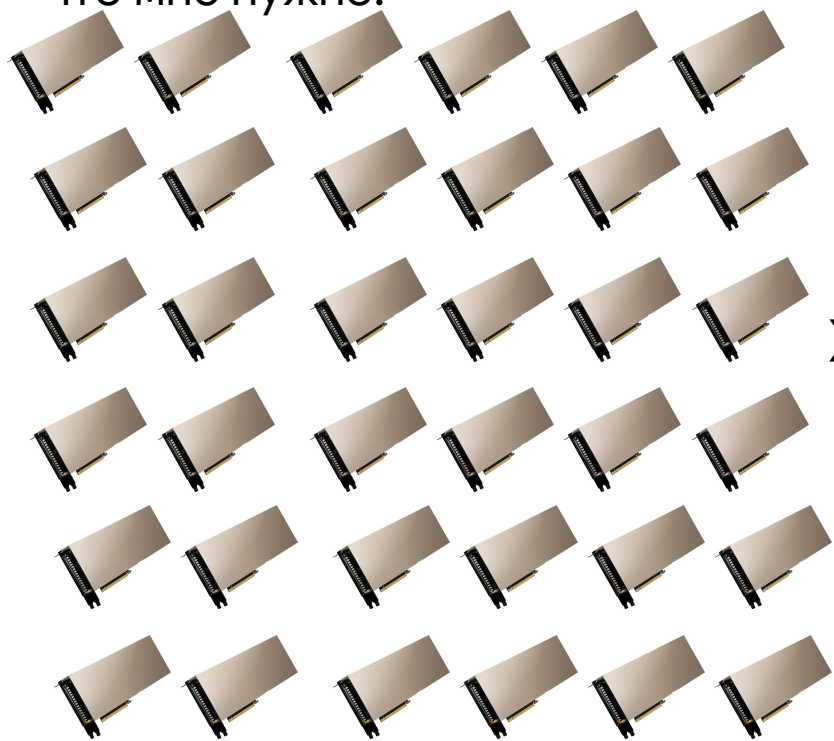
Так, я хочу дообучить на своих данных,
что мне нужно:





Так, я хочу обучить с нуля на своих данных,

что мне нужно:



x 1000 +





Как я запускал в Контуре YaLM 100B

- На ЦПУ Жрет ~ 400 ГБ Озу 🐼🍔🍷

Intel ЦПУ не умеет в halfFloat, (сейчас только AMD zen4+);

1 параметр = float = 4 байта

- Совсем не быстро :(

На 12 ядрах диалога не получается, медленно

- Нет instruction fine-tune-a

Модель умеет продолжать текст, и ее **не учили отвечать на вопросы (полуфабрикат)**

- Сейчас ситуация немного лучше, так как научились запускать в 4 бита (с потерей качества)



Как я запускал в Контуре YaLM 100B

```
bash /root/mp/YaLM-100B  x  Windows PowerShell  x  Windows PowerShell  x  +  v
Megatron-LM:
Context: Переведи это предложение на английский:"Александр Пушкин начал писать свои первые произведения уже в семь лет." =>
Megatron-LM: -----
Di
@good_soul_69 смотри,нашла отличное выражение "Не можешь учиться хорошо,бу
Context: Переведи это предложение на английский:"Александр Пушкин начал писать свои первые произведения уже в семь лет." =>
Megatron-LM: -----
Di
@good_soul_69 смотри,нашла отличное выражение "Не можешь учиться хорошо,будешь работать,как все!" можешь применять,когда кто-то
-----
Di
Press any key to continue >>>
Context prompt (stop to exit) >>> Переведи это предложение на английский:"Александр Пушкин начал писать свои первые произведения уже в семь лет."
Context: Переведи это предложение на английский:"Александр Пушкин начал писать свои первые произведения уже в семь лет."
Megatron-LM:
```

```
1 [|| 2.0%] 4 [||||100.0%] 7 [ 0.0%] 10 [|||||99.3%]
2 [||||100.0%] 5 [||||100.0%] 8 [|||||99.3%] 11 [|||||99.3%]
3 [||||100.0%] 6 [||||100.0%] 9 [|||||99.3%] 12 [|||||99.3%]
Mem[|||||||||||||||||379G/443G] Tasks: 43, 141 thr; 11 running
Swp[| 1.01M/3.84G] Load average: 9.42 7.54 5.65
Uptime: 00:48:48
```

PID	USER	PRI	NI	VIRT	RES	SHR	S	CPU%	MEM%	TIME+	Command
3805	root	20	0	382G	377G	138M	R	99.7	85.2	3h13:21	/opt/conda/
3895	root	20	0	382G	377G	138M	R	100.	85.2	17:57.74	/opt/conda/
3898	root	20	0	382G	377G	138M	R	100.	85.2	17:58.09	/opt/conda/
3900	root	20	0	382G	377G	138M	R	100.	85.2	17:57.73	/opt/conda/
3902	root	20	0	382G	377G	138M	R	100.	85.2	17:58.14	/opt/conda/
3896	root	20	0	382G	377G	138M	R	100.	85.2	17:58.15	/opt/conda/
3899	root	20	0	382G	377G	138M	R	100.	85.2	17:58.75	/opt/conda/
3894	root	20	0	382G	377G	138M	R	100.	85.2	17:58.38	/opt/conda/
3901	root	20	0	382G	377G	138M	R	100.	85.2	17:58.28	/opt/conda/
3897	root	20	0	382G	377G	138M	R	99.3	85.2	17:58.23	/opt/conda/
1319	root	20	0	1673M	39736	18276	S	1.3	0.0	0:00.53	/usr/bin/co
923	root	20	0	231M	6916	5684	S	0.7	0.0	0:01.62	/usr/bin/vm
3802	root	20	0	8124	3916	3096	R	0.0	0.0	0:18.52	htop
1057	root	20	0	1673M	39736	18276	S	0.0	0.0	0:03.16	/usr/bin/co
1373	root	20	0	2000M	75936	48524	S	0.0	0.0	0:00.77	/usr/bin/do
4370	root	20	0	2000M	75936	48524	S	0.0	0.0	0:00.05	/usr/bin/do
1366	root	20	0	1673M	39736	18276	S	0.0	0.0	0:00.27	/usr/bin/co
3882	root	20	0	382G	377G	138M	S	0.0	85.2	0:02.04	/opt/conda/
864	root	RT	0	337M	18052	8208	S	0.0	0.0	0:01.02	/sbin/multi
607	root	19	-1	196M	30408	29352	S	0.0	0.0	0:01.02	/lib/system

```
F1 Help F2 Setup F3 Search F4 Filter F5 Tree F6 SortBy F7 Nice F8 Nice F9 Kill F10
```



Как я запускал модели поменьше

Опенсорс модели до 24Gb спокойно залезают в 1 домашнюю видяху

- На английском хорошо работает Bloomz, galactica, LLAMA
- Небольшие модели существенно глупее больших
- С качественными русскоязычными моделями **бида**
- Можно затюнить YaLM но нужно потратить \$ на пару дней аренду тачки с 8 A100 (1 час ~2400р) + подобрать хороший датасет



Локально

С Видяхой

- https://github.com/johnsmith0031/alpaca_lora_4bit

без Видяхи через NPM

- <https://cocktailpeanut.github.io/dalai/#/>

без Видяхи через shell

- <https://replicate.com/blog/run-llama-locally>



Где попробовать? (сервисы)

Сервисы на основе языковых моделей

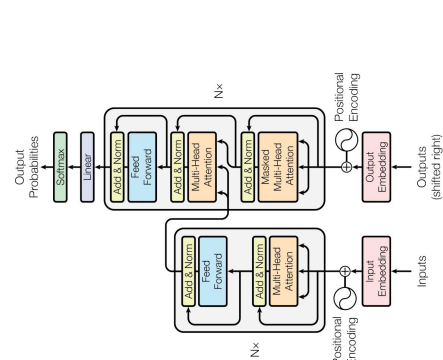
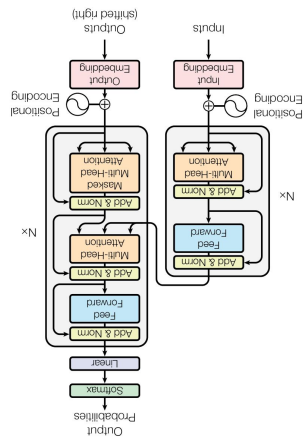
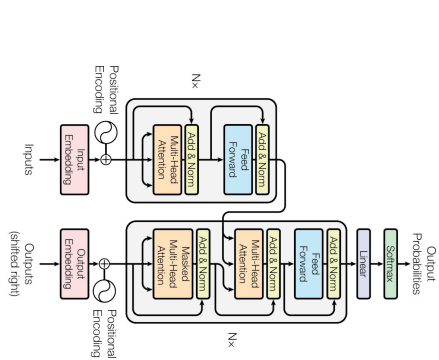
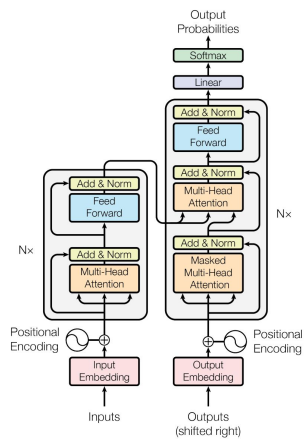
<https://www.chatpdf.com/> - Чат по любой PDF

<https://you.com/> - Поисковик + чат

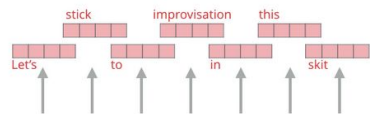
<https://www.perplexity.ai/> - Поисковик + чат

<https://yandex.ru/lab/yalm> - Балобола

И много других



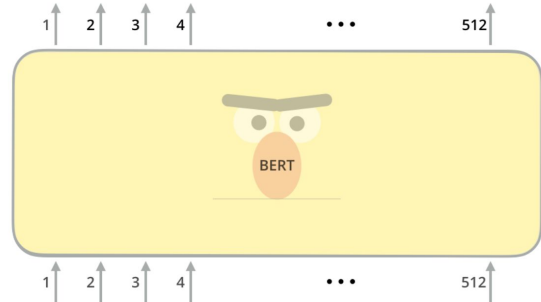
Трансформеры

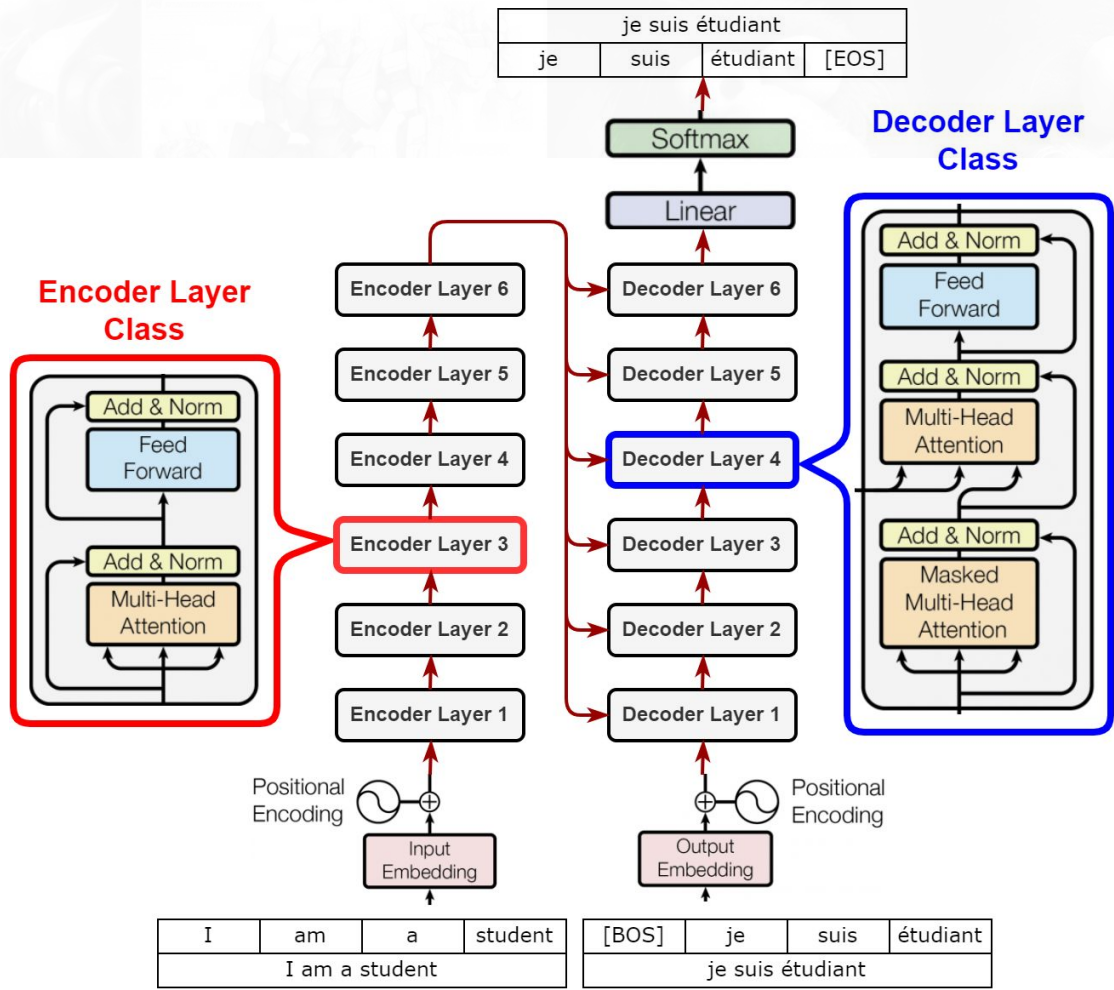


Шестаков Алексей shetakovap@skbkontur.ru

Темная сторона
(в следующей серии)

...the truth is out there...





Encoder Layer Class

Decoder Layer Class

Positional Encoding

Positional Encoding

Input Embedding

Output Embedding

I	am	a	student
I am a student			

[BOS]	je	suis	étudiant
je suis étudiant			



Арифметика

Пример

YaLM 100B - 100 Миллиардов параметров

$100 \text{ B} * 2 = 200 \text{ Гб}$ Видеопамяти

Это 3 видяхи A100 80 Гб или 5 видях A100 40 Гб

это **ТОЛЬКО** для запуска модели на GPU

Нас спасет авито



🔍 Все категории

Поиск по объявлениям

Найти

Екатеринбург · ... · Товары для компьютера · Комплектующие · Видеокарты

Видео карта Nvidia Tesla A100 40GB

570 000 ₽

или предложите свою цену

📍 Добавить в избранное

📌 Добавить заметку



Показать телефон
8 xxx xxx-xx-xx






Написать сообщение
Отвечает за несколько часов

Пользователь
Частное лицо

Документы проверены

Подписаться на продавца

Ещё Opensource Модели

 BigScience	BLOOM	176B	July 2022
	T0pp	11B	October 2021
 EleutherAI	GPT-J	6B	July 2021
	GPT-NeoX	20B	February 2022
 清华大学 Tsinghua University	GLM	130B	August 2022
 Google Research	UL2	20B	October 2022
	T5	11B	February 2020
 Meta AI	OPT	175B	June 2022
	OPT	66B	June 2022



Арифметика

Один параметр – число с плавающей точкой.

Может быть float (32 бита, ака 4 байта)

Может быть halfFloat, bfloat (16 бит, ака 2 байта)

Все что меньше уже **не поддается** обычно обучению.

Но с потерей качества поддается запуску **даже** на 4х битах