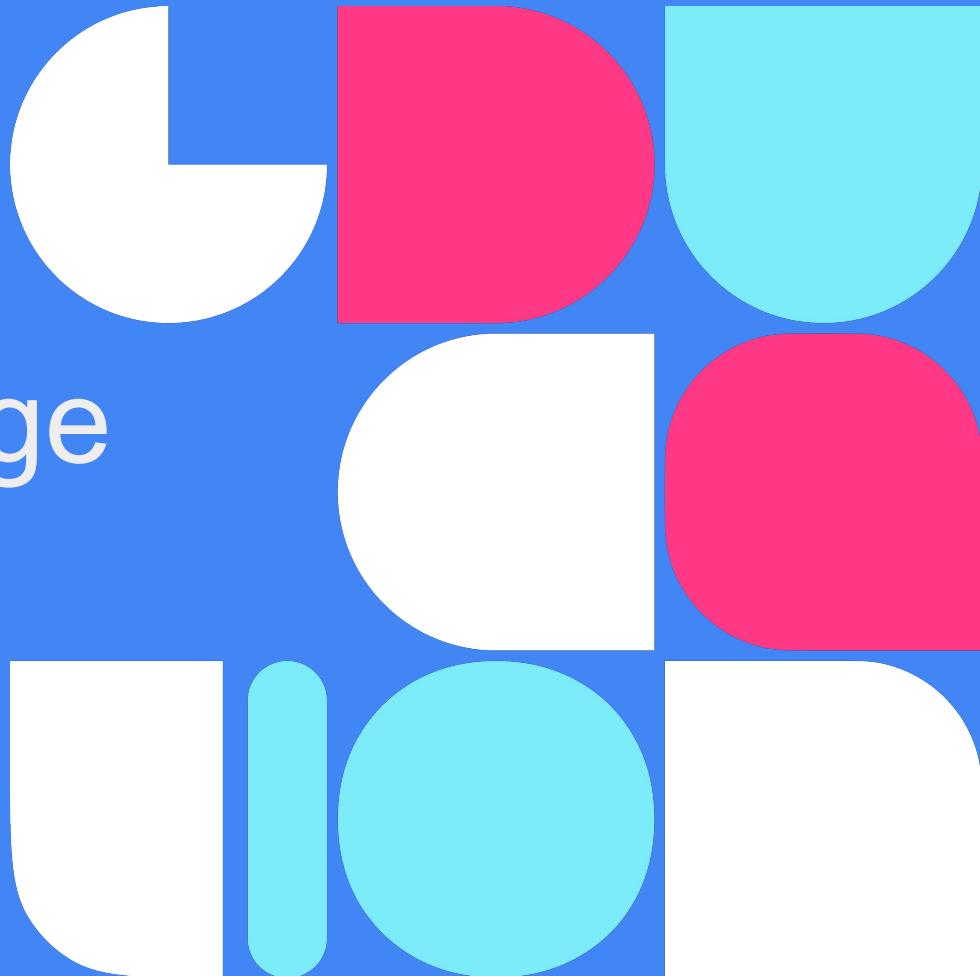




Natural Language Processing

Нейронные сети в ML

Козулин Илья



Цель занятия:

- Разобраться как языковые модели учатся моделировать язык
- Изучить развитие GPT-подобных моделей
- Рассмотреть методы повышения точности авторегрессивных моделей
- Исследовать основные подходы к parameter-efficient fine-tuning
- Посмотреть на различные способы сэмплирования токенов

Регламент на занятия 29.10 и 05.11

- Ставим в чат
 - “++” если материал предельно понятен/очевиден/уже известен
 - “+” если материал более менее понятен, либо нужно еще немного над ним подумать, но можно идти дальше
 - “-” если что-то совсем не понятно и нужно остановиться и разобрать подробнее
- На лекции скидывается опрос с вопросами по материалу. Оценка за опрос идет к оценке за домашнее задание
- На семинаре будут практические задания, решаемые студентами с демонстрацией экрана. За каждое задание +1 балл к оценке за дз. Максимум +2 балла на одного человека.

О чём поговорим?

- Какие задачи охватывает NLP
- Векторные представления слов/предложений
- Моделирование языка
- Что позволило масштабировать модели
- Parameter-Efficient Fine-Tuning

О чём поговорим?

- Какие задачи охватывает NLP Часть 1
- Векторные представления слов/предложений
- Моделирование языка (Causal Language Modeling)
- Что позволило масштабировать модели Часть 2
- Parameter-Efficient Fine-Tuning

О чём поговорим?

- Какие задачи охватывает NLP Часть 1
 - Векторные представления слов/предложений
 - Моделирование языка
 - Масштабирование моделей Часть 2
 - Parameter-Efficient Fine-Tuning
-

- Как все начиналось
- Где мы сейчас
- Куда мы движемся

Введение



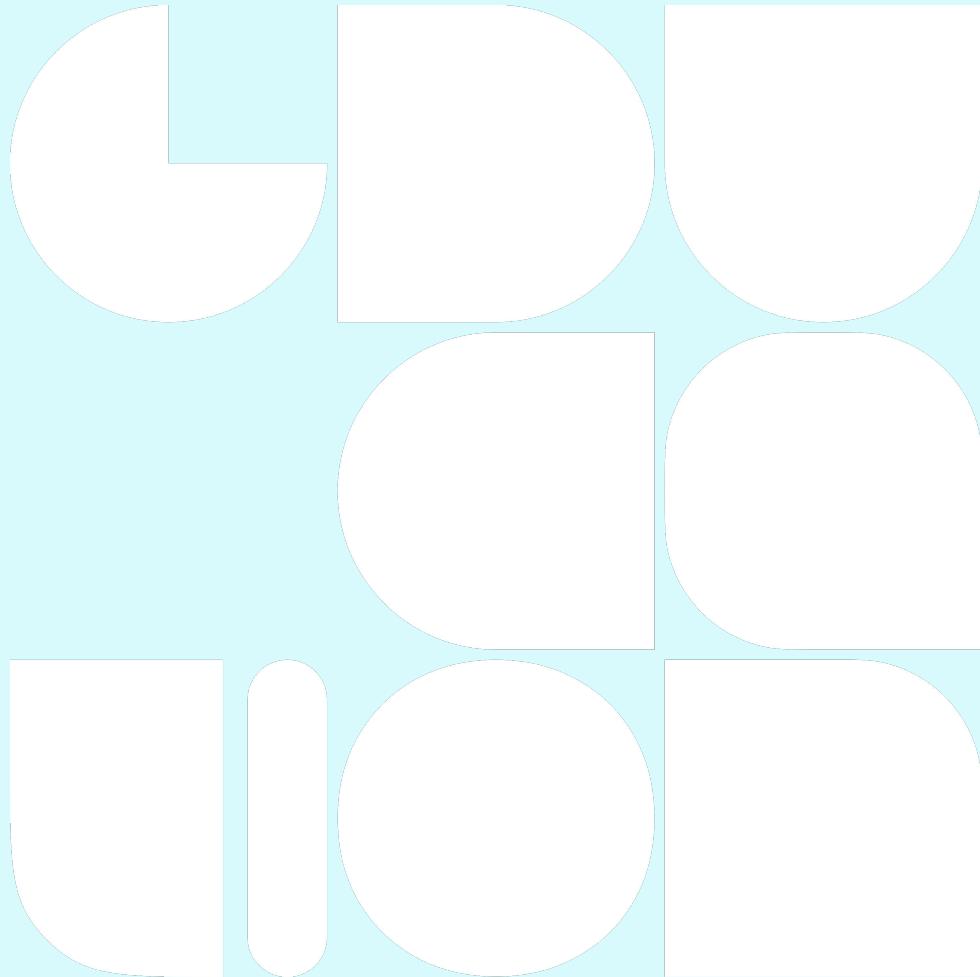
Qwen

MISTRAL
AI_

Введение



Causal Language Modeling



Recap: Моделирование языка (Language Modeling)



Recap: Моделирование языка (Language Modeling)

Я к вам **MASK** — чего же боле? Что я могу еще сказать



Masked Language Modeling (MLM)



пишу

Модель видит все токены, включая те, что после маски

Я к вам пишу — чего же боле? Что я могу еще **MASK**



Causal Language Modeling (CLM)



сказать

Модель видит только те токены, что находятся до маски

GPT-1

Generative Pre-Trained Transformer

1. Transformer decoder only
2. Unsupervised pre-training
3. Supervised fine-tuning
4. Task-specific input transformations

GPT-1

Generative Pre-Trained Transformer

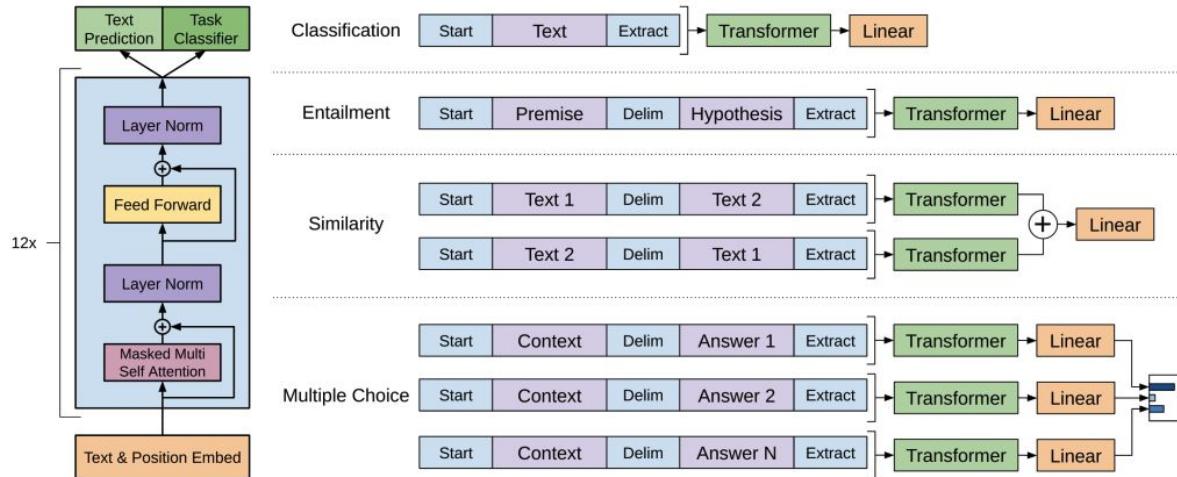


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

GPT-2

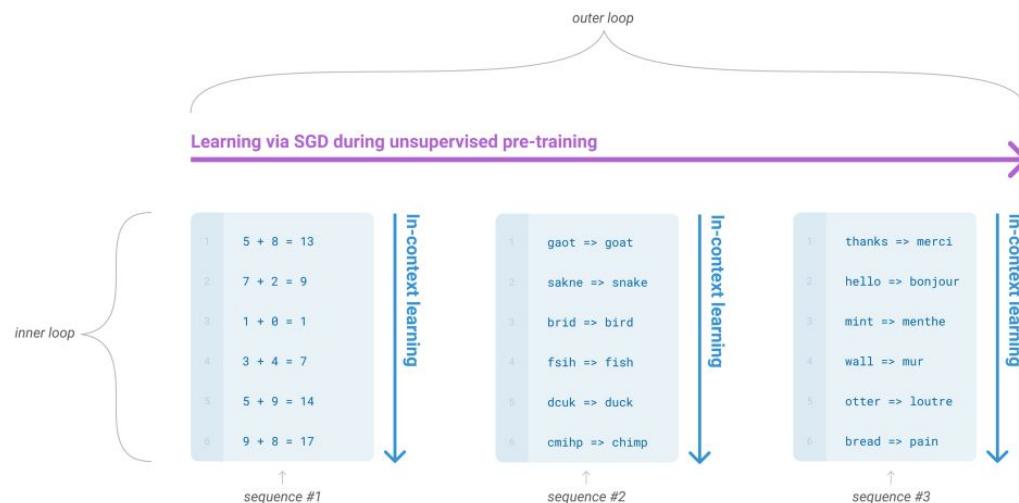
Generative Pre-Trained Transformer

1. Горизонтальное и вертикальное масштабирование модели
2. Обучение на разнообразных данных и задачах для решения сразу множества задач без дообучения

GPT-3

Generative Pre-Trained Transformer

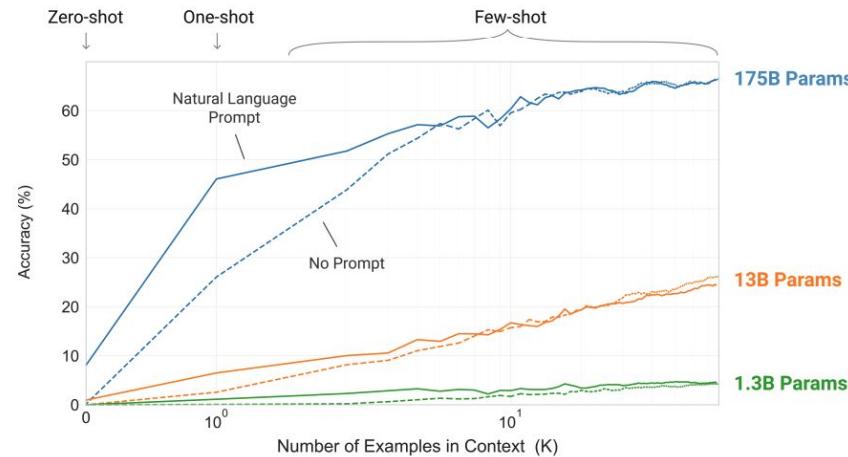
1. Горизонтальное и вертикальное масштабирование модели
2. In-Context Learning



GPT-3

Generative Pre-Trained Transformer

1. Горизонтальное и вертикальное масштабирование модели
2. In-Context Learning



Сравнение GPT

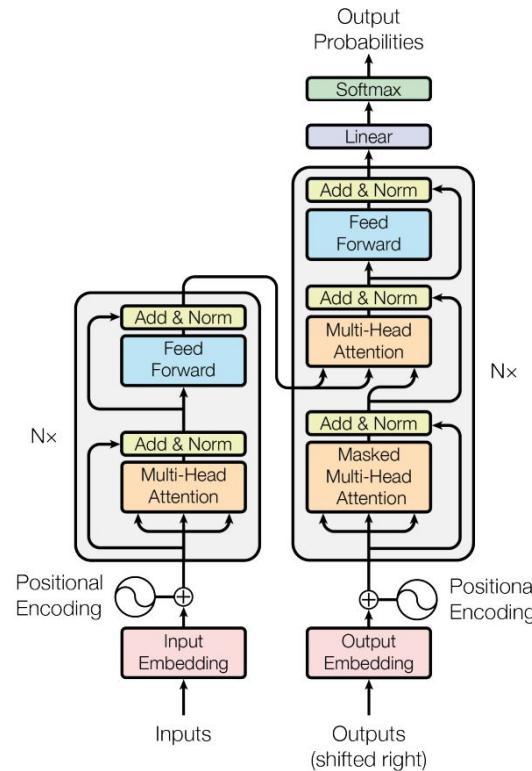
	GPT-1	GPT-2	GPT-3
Количество параметров	117M	1.5B	175B
Размер контекста	512	1024	2048
Особенности обучения	Pre-training + Fine-tuning под конкретную задачу. Задачи	Pre-training → Zero-shot	Pre-training → Few-shot
Размер датасета	~5Gb	~40Gb	~570Gb
Год	2018	2019	2020



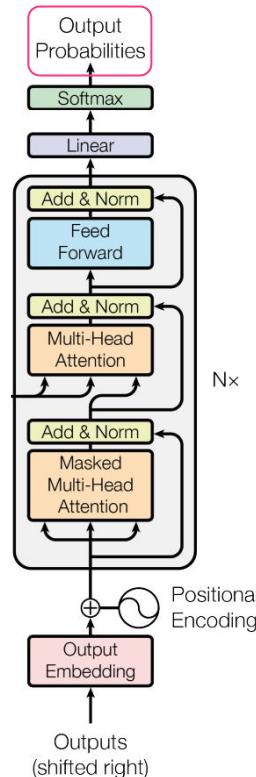
Вопросы?



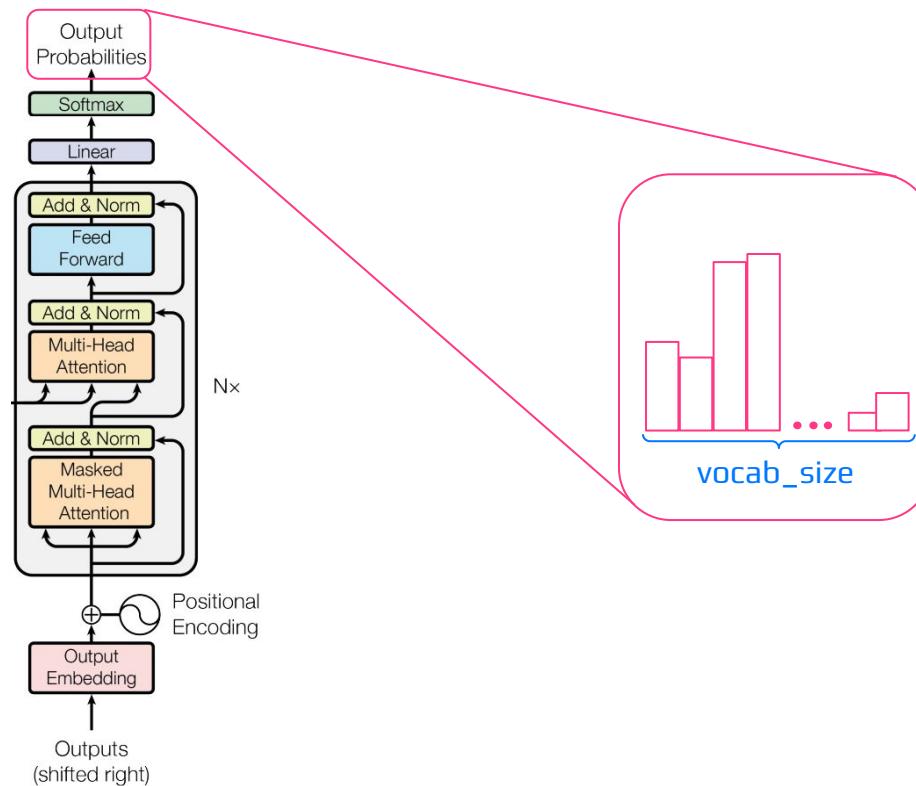
Как сэмплируются токены?



Как сэмплируются токены?

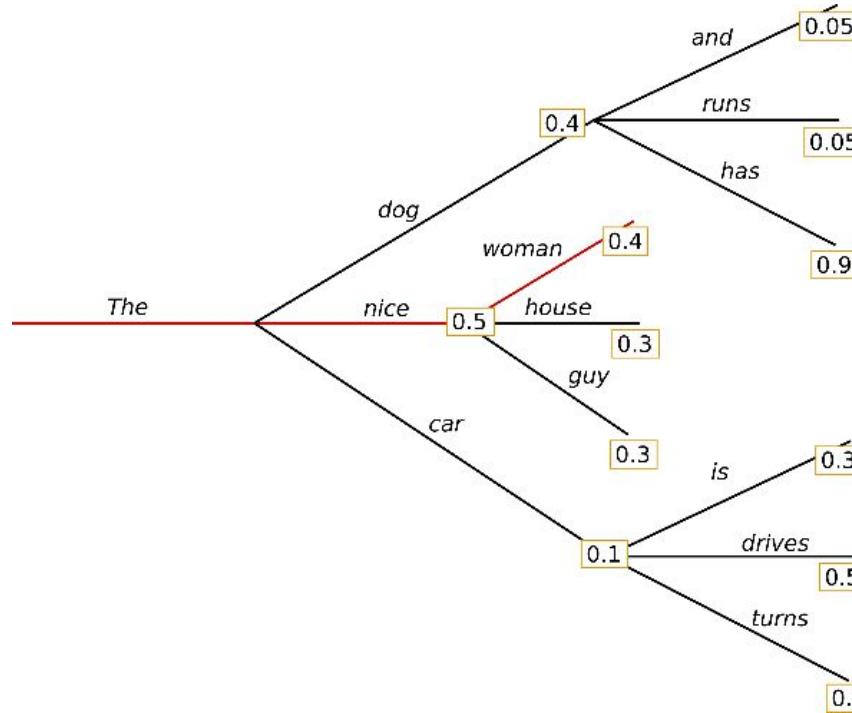


Как сэмплируются токены?



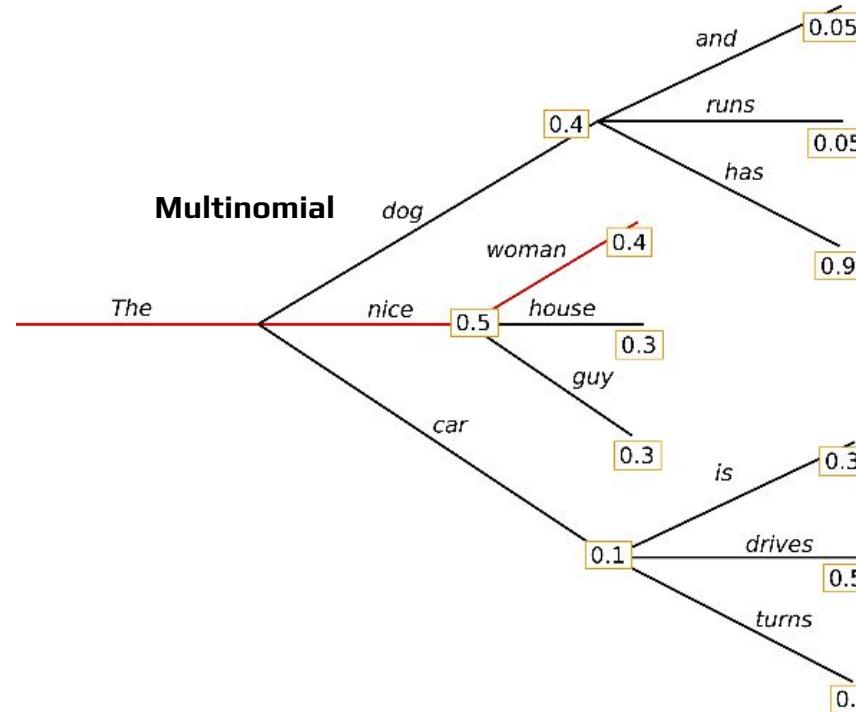
Как сэмплируются токены?

Greedy Sampling



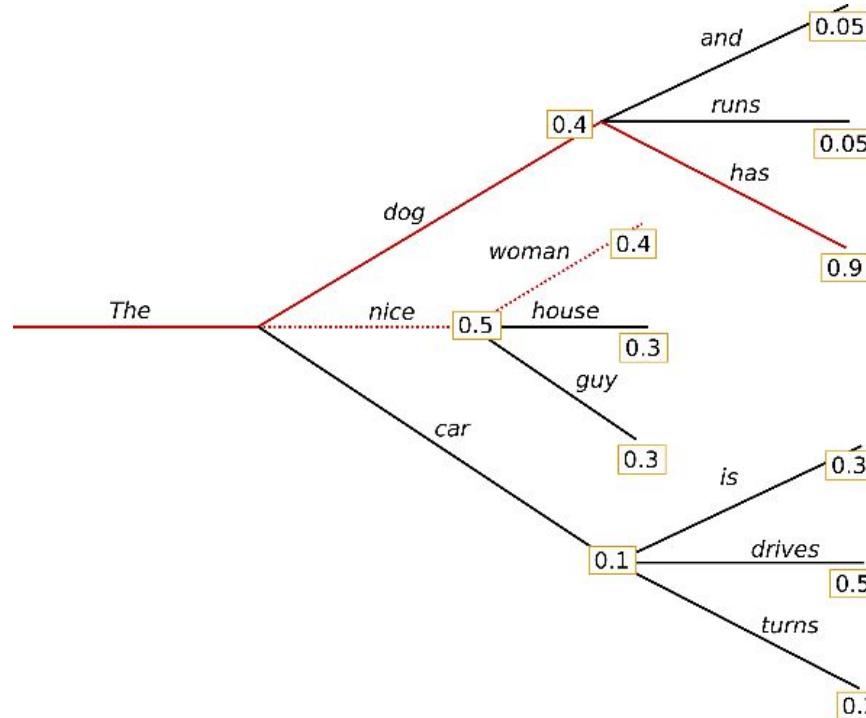
Как сэмплируются токены?

Random Sampling



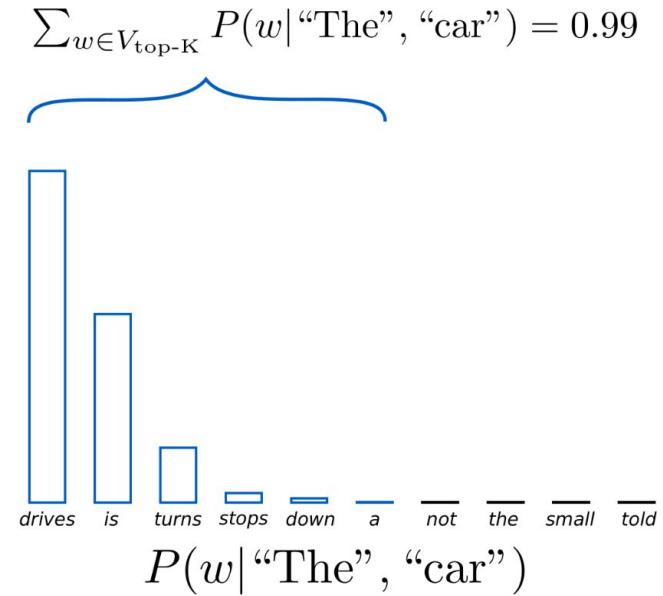
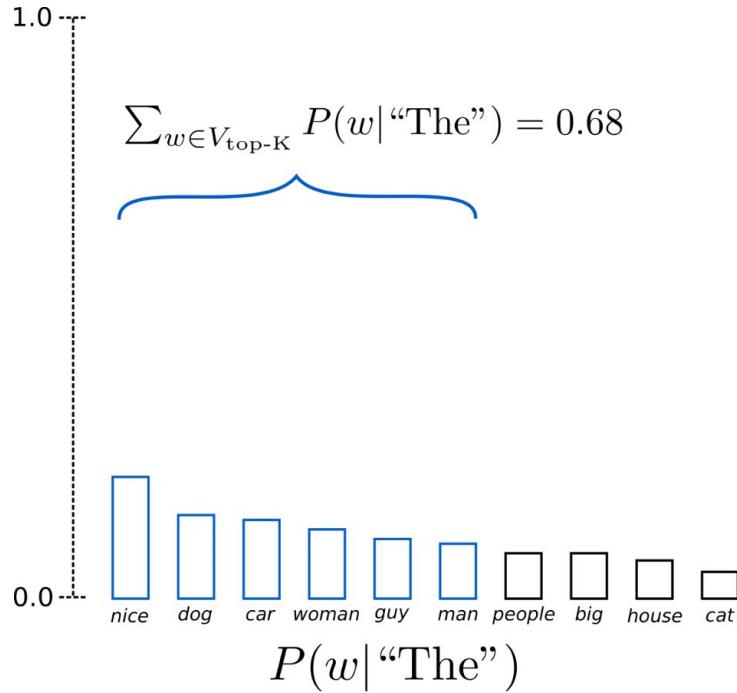
Как сэмплируются токены?

Beam-Search



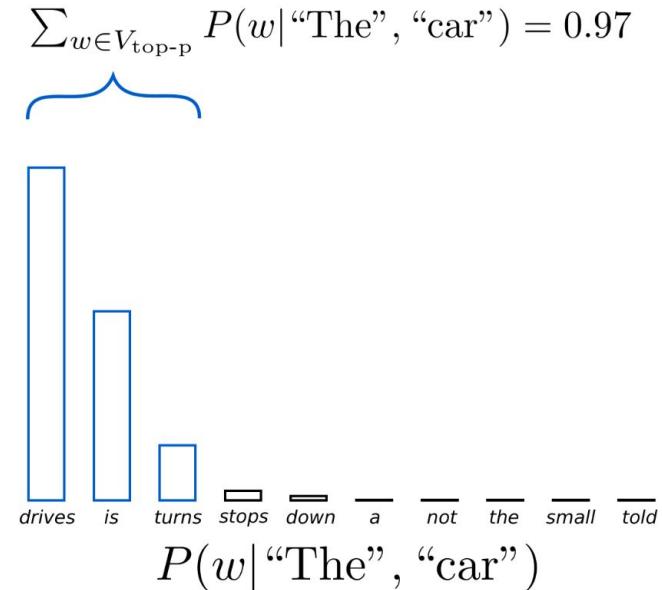
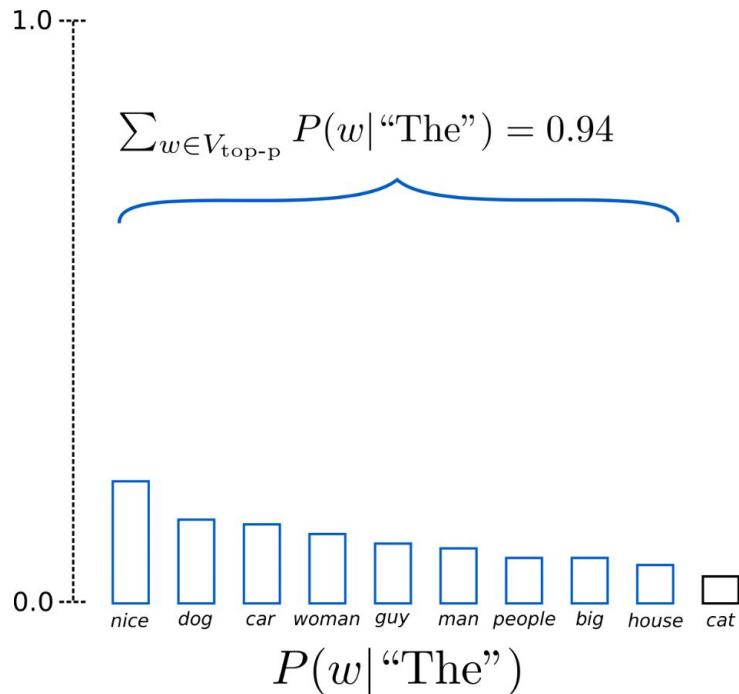
Как сэмплируются токены?

Top-K Sampling



Как сэмплируются токены?

Top-P Sampling



Как сэмплируются токены?

Sampling with softmax temperature

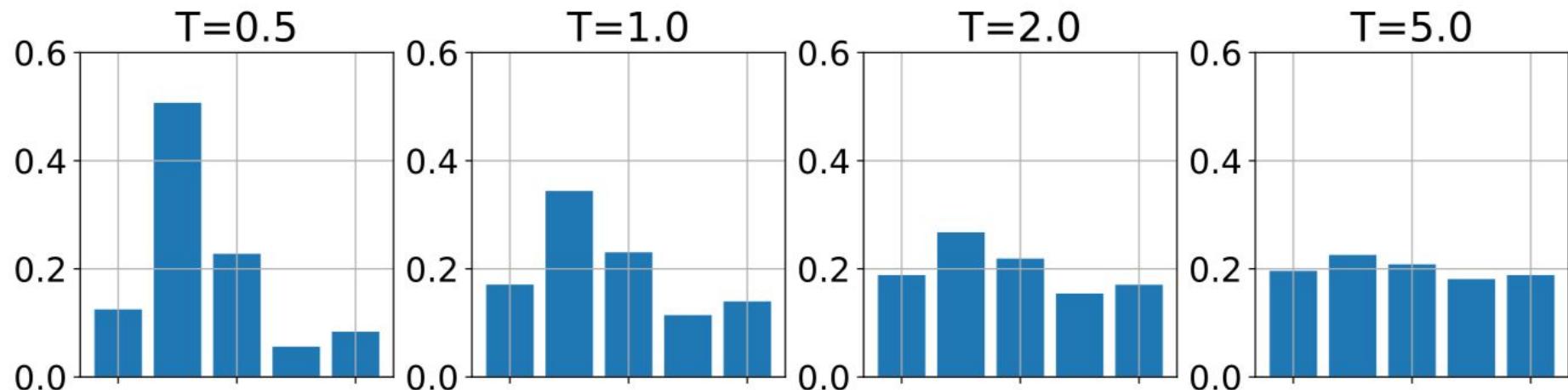
$$P_i = \frac{e^{y_i}}{\sum_{k=1}^n e^{y_k}}$$



$$P_i = \frac{e^{\frac{y_i}{T}}}{\sum_{k=1}^n e^{\frac{y_k}{T}}}$$

Как сэмплируются токены?

Sampling with softmax temperature





Вопросы?



Как оценивается качество ответов?

Accuracy – Exact Match

Пример: GSM8K

Query	Response
James writes a 3-page letter to 2 different friends twice a week. How many pages does he write a year?	He writes each friend $3 \times 2 = <<3*2=6>>6$ pages a week So he writes $6 \times 2 = <<6*2=12>>12$ pages every week That means he writes $12 \times 52 = <<12*52=624>>624$ pages a year #### 624

Как оценивается качество ответов?

Accuracy – Exact Match

Пример: GSM8K

Query	Response
James writes a 3-page letter to 2 different friends twice a week. How many pages does he write a year?	He writes each friend $3 \times 2 = <<3 \times 2 = 6>> 6$ pages a week So he writes $6 \times 2 = <<6 \times 2 = 12>> 12$ pages every week That means he writes $12 \times 52 = <<12 \times 52 = 624>> 624$ pages a year #### 624

About 624 pages
Six hundred twenty four
624

?

Как оценивается качество ответов?

Accuracy – Multiple Choice (Log-likelihood)

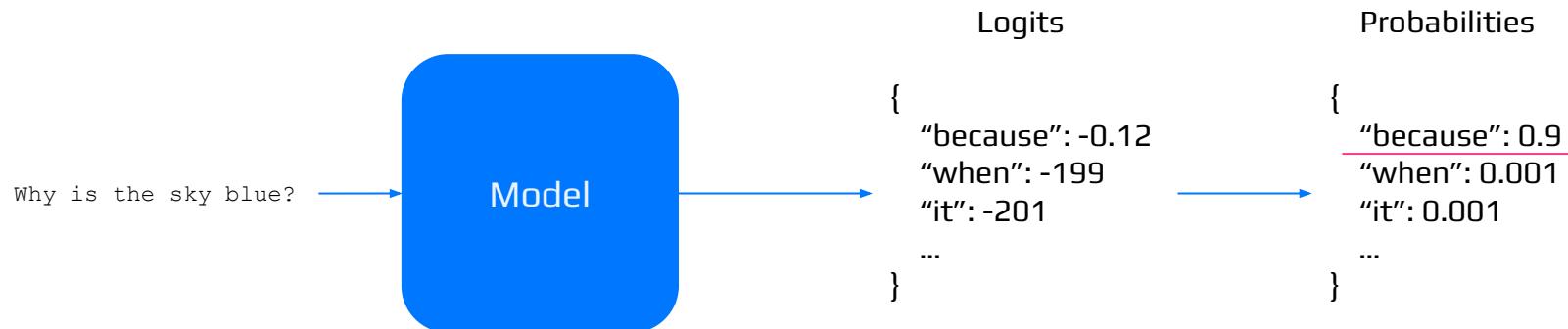
Пример: MMLU

Query	Response
Why is the sky blue?	<ol style="list-style-type: none">1) "Because the molecules that compose the Earth's atmosphere have a blue-ish color."2) "Because the sky reflects the color of the Earth's oceans."3) "<u>Because the atmosphere preferentially scatters short wavelengths.</u>"4) "Because the Earth's atmosphere preferentially absorbs all other colors."

Как оценивается качество ответов?

Accuracy – Multiple Choice (Log-likelihood)

Пример: MMLU



$$\text{Loglikelihood}(\text{seq}) = \log(0.9) + \dots$$

"Because the atmosphere preferentially scatters short wavelengths."

Как оценивается качество ответов?

Accuracy – Multiple Choice (Log-likelihood)

Пример: MMLU



$$\text{Loglikelihood}(\text{seq}) = \log(0.9) + \log(0.001) + \dots$$

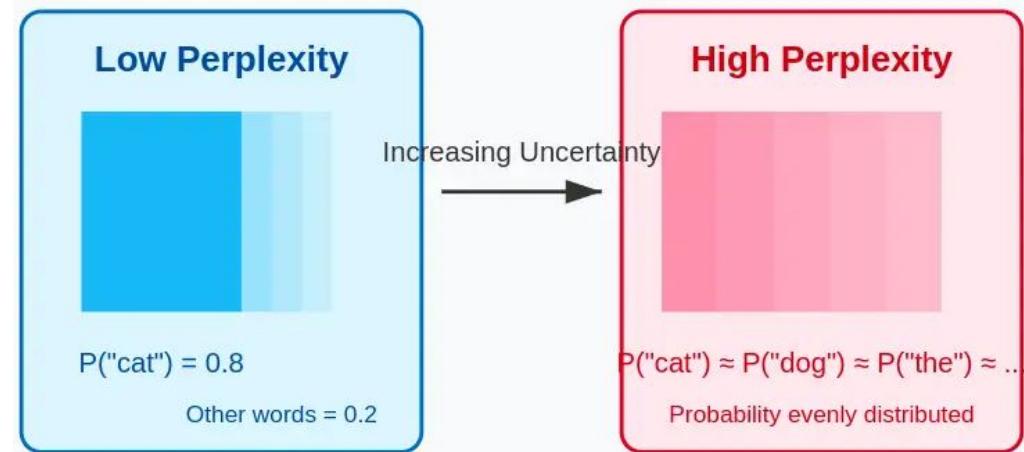
"Because the atmosphere preferentially scatters short wavelengths."

Как оценивается качество ответов?

Perplexity

$$PP(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, \dots, w_N)}}$$

$$PP(W) = 2^{H(W)} = 2^{-\frac{1}{N} \log_2 P(w_1, w_2, \dots, w_N)}$$



Бенчмарки: LAMBADA, WikiText2, PTB, 1BW

Как оценивается качество ответов?

BLEU

$$\text{BLEU} = \text{BP} \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right)$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ \exp \left(1 - \frac{r}{c} \right) & \text{if } c \leq r \end{cases}$$

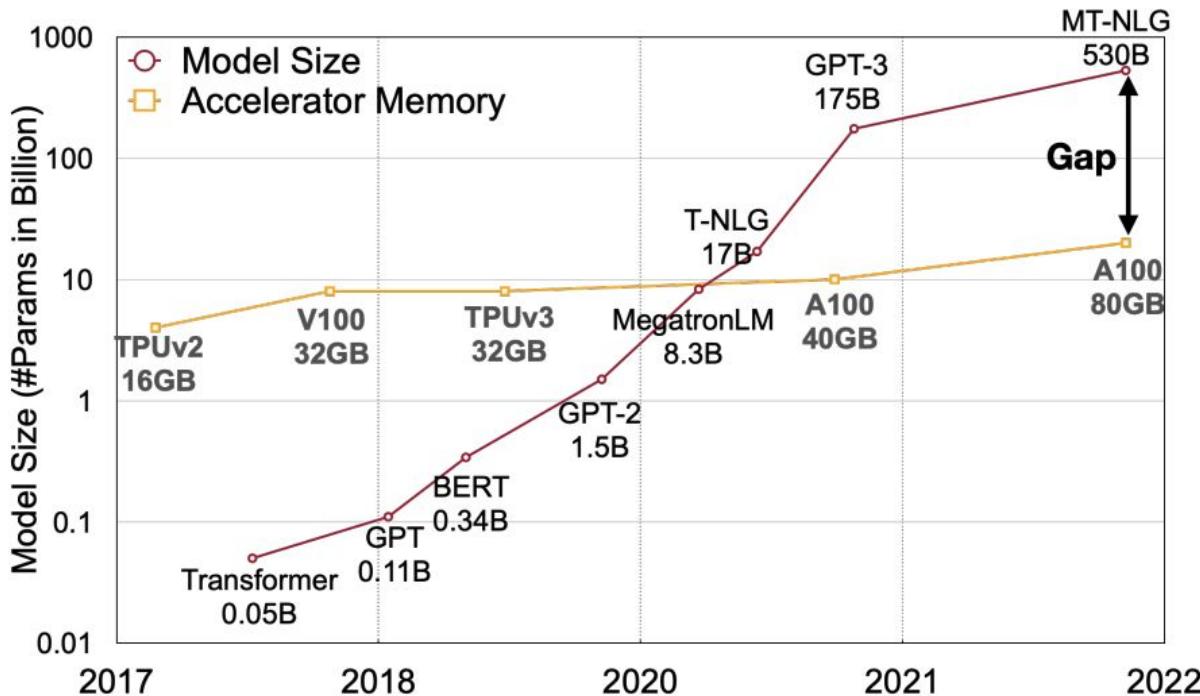
Бенчмарки:



Вопросы?



Где начинаются LLM?



Где начинаются LLM?



ChatGPT

Где начинаются LLM?

Scaling + Alignment

Alignment?

3 “H” = Helpful + Harmless + Honest

Где начинаются LLM?

Scaling + Alignment

Training Phase

Alignment

Unsupervised pre-training + Supervised Fine-Tuning + RLHF

Где начинаются LLM?

Scaling + Alignment

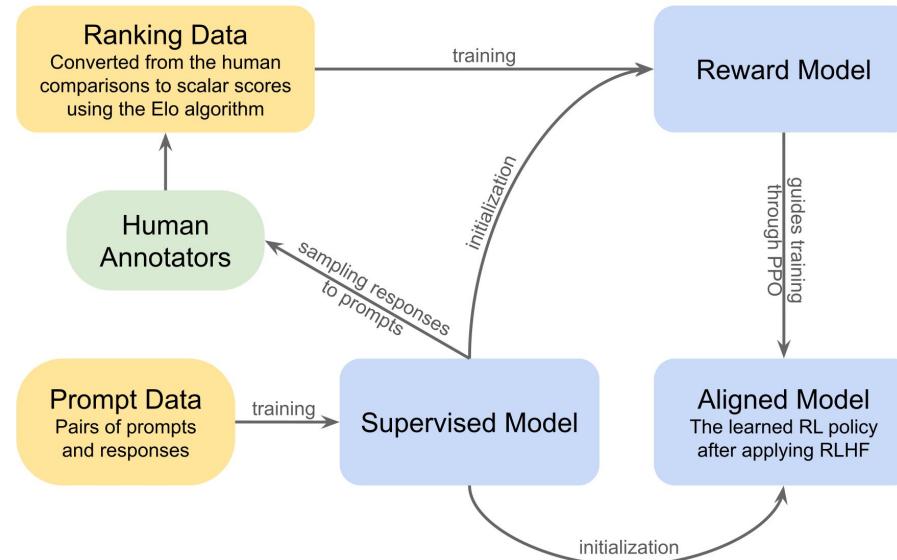
Supervised Fine-Tuning (SFT)

- Обучение с учителем для следования инструкциям пользователя
- Как правило, выборка размечается людьми. Финальная модель часто также выбирается с участием ассессоров.
- В модель не добавляется новых “знаний”, только умение следовать инструкциям
- На выходе готовая к использованию instruct-tuned модель

Где начинаются LLM?

Scaling + Alignment

Reinforcement Learning from Human Feedback



Итоги раздела:

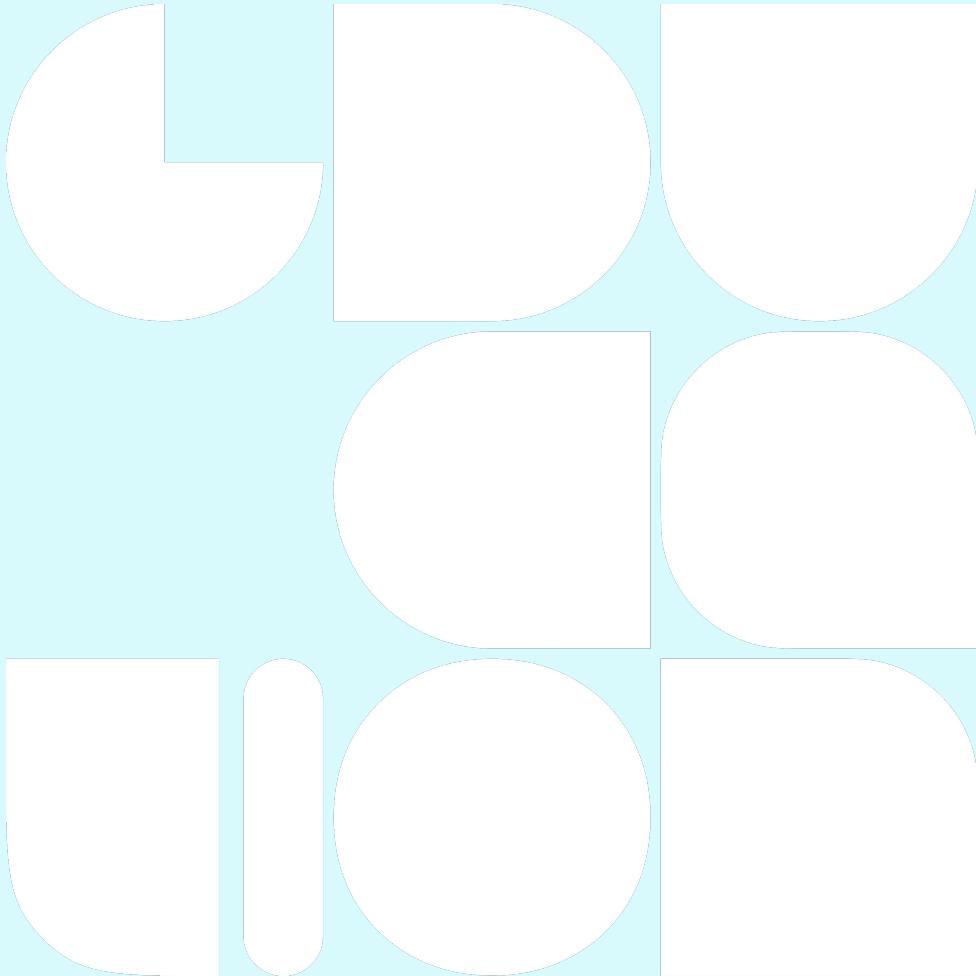
- Просмотрели на развитие GPT (1-3) и causal language modeling
- Узнали как сэмплируются токены
- Изучили одни из наиболее популярных метрик для измерения качества генерации
- Узнали ключевые фишки обучения современных языковых моделей



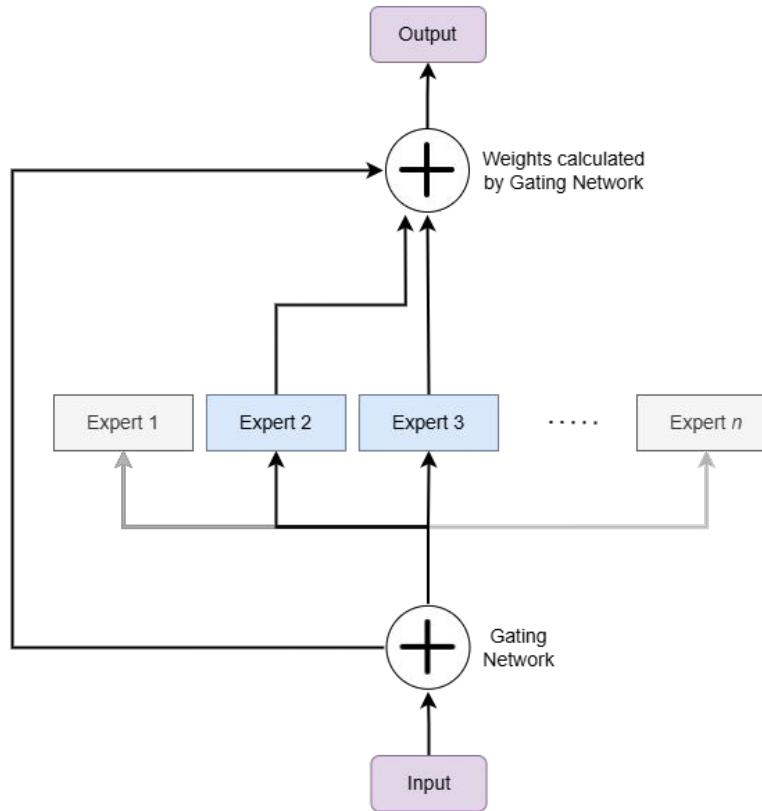
Вопросы?



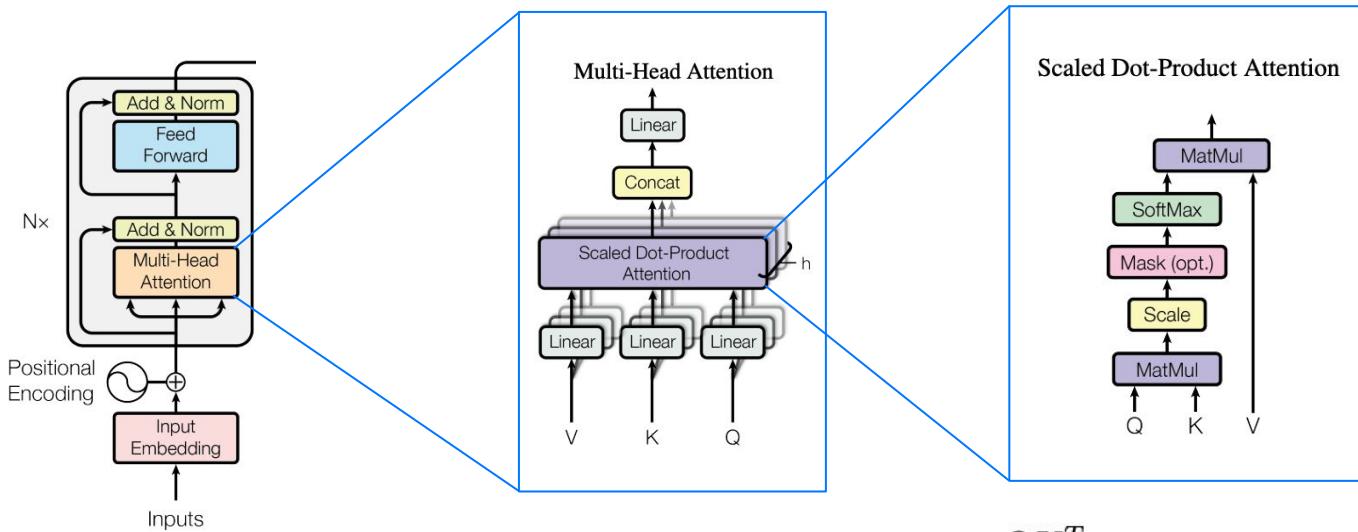
Что позволило масштабировать модели?



Mixture-of-Experts



Recap: Multi-Head Attention

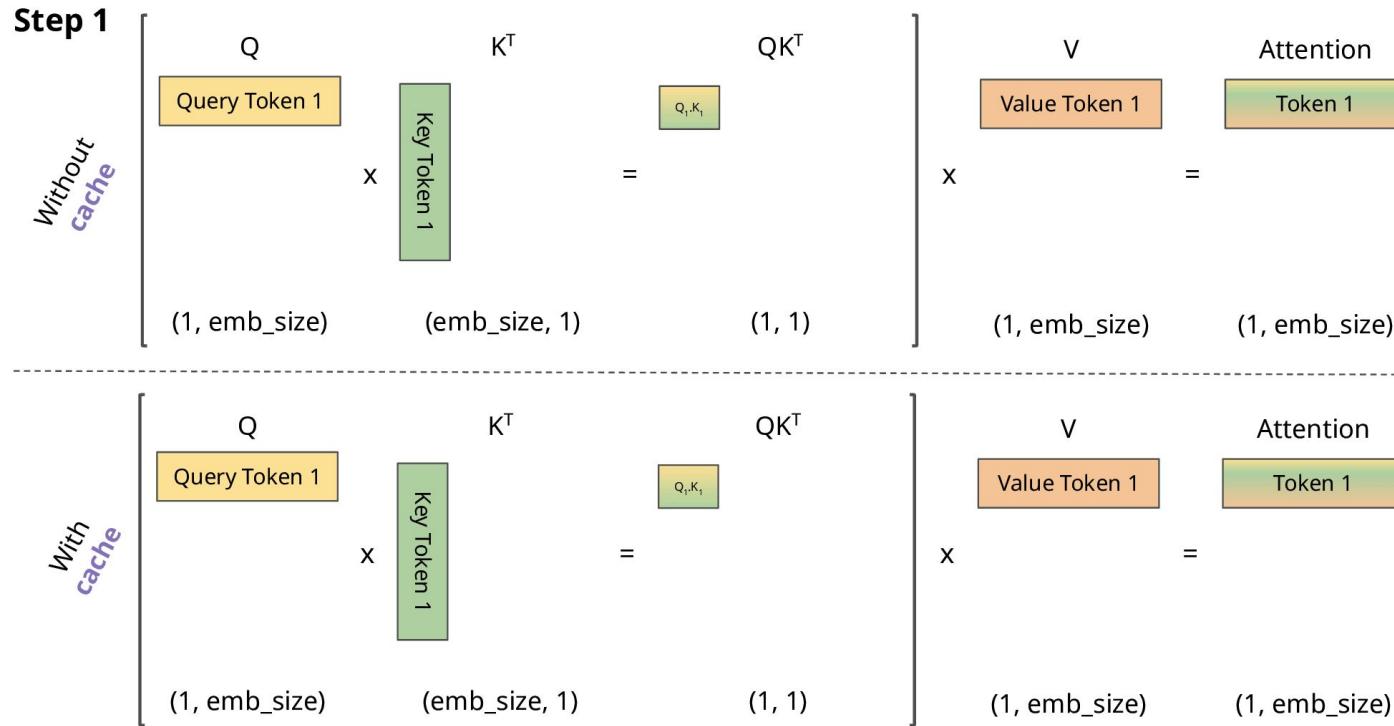


$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

$$\text{where } \text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

KV-Cache

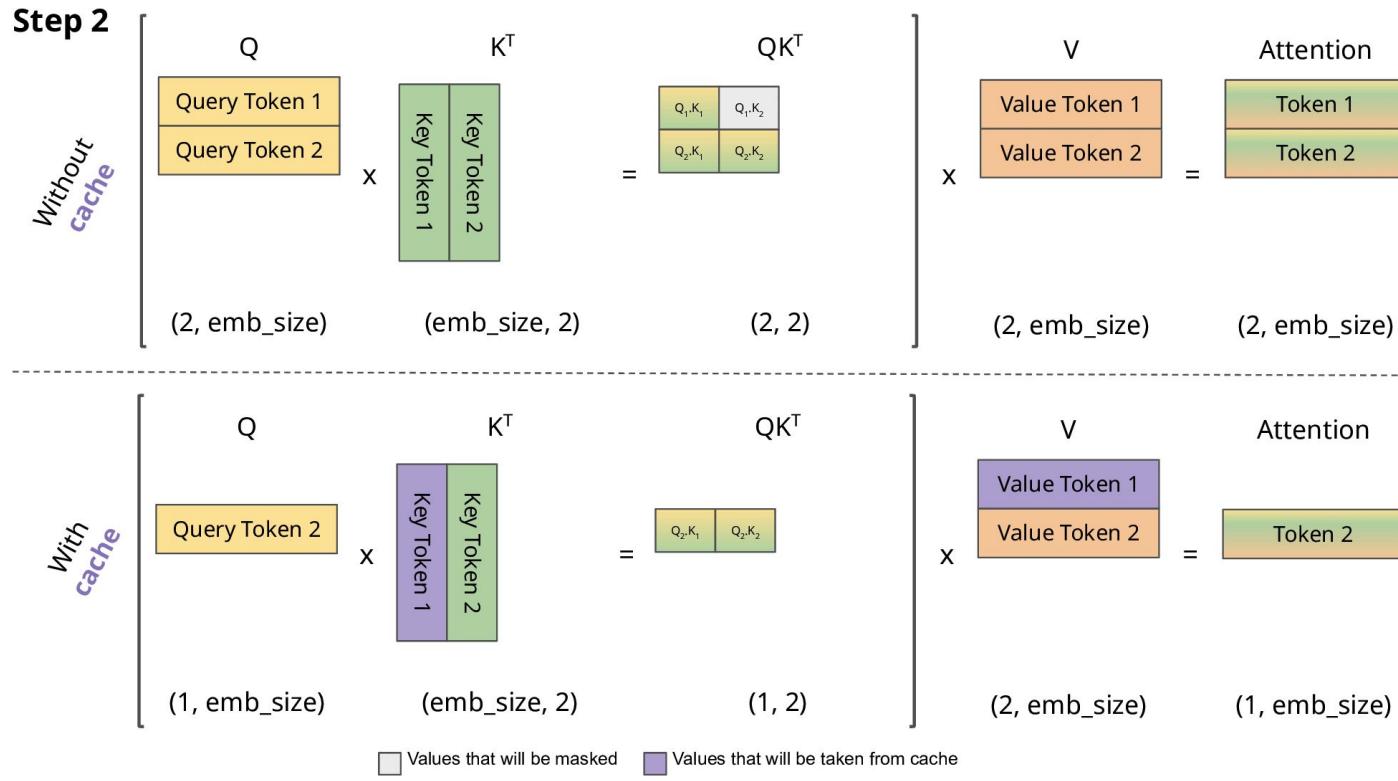


Values that will be masked



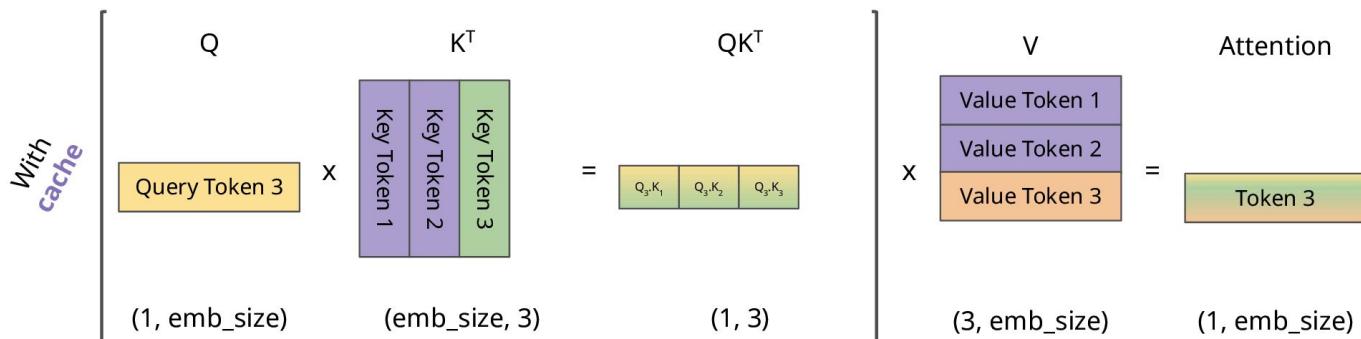
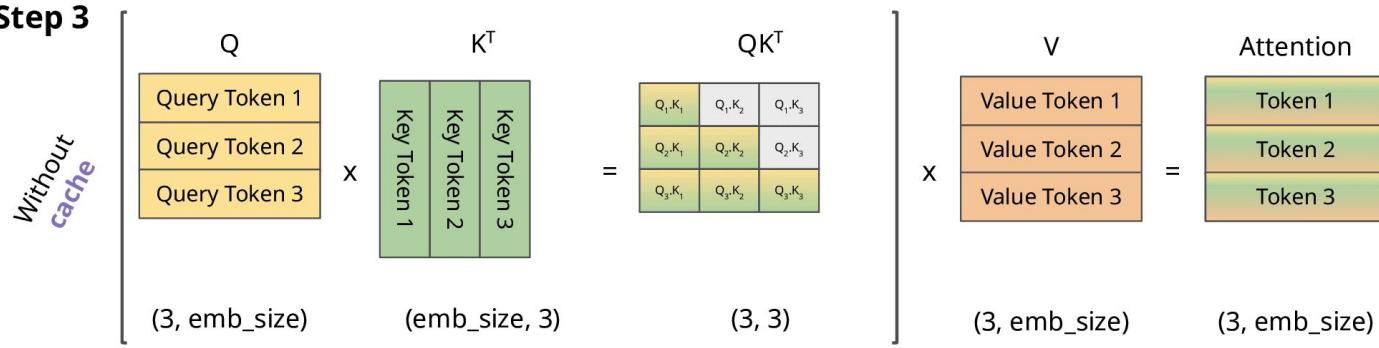
Values that will be taken from cache

KV-Cache



KV-Cache

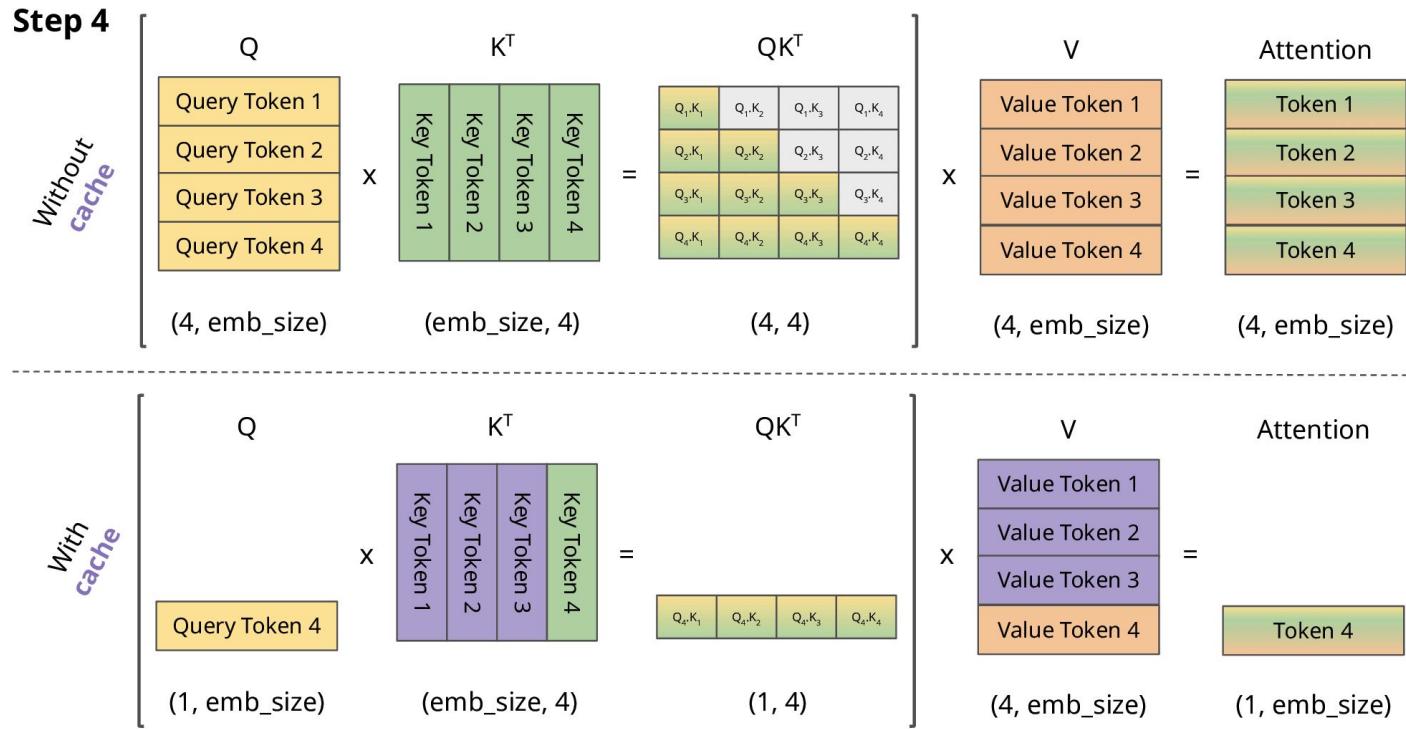
Step 3



□ Values that will be masked

■ Values that will be taken from cache

KV-Cache



□ Values that will be masked

■ Values that will be taken from cache

Sliding-Window Attention

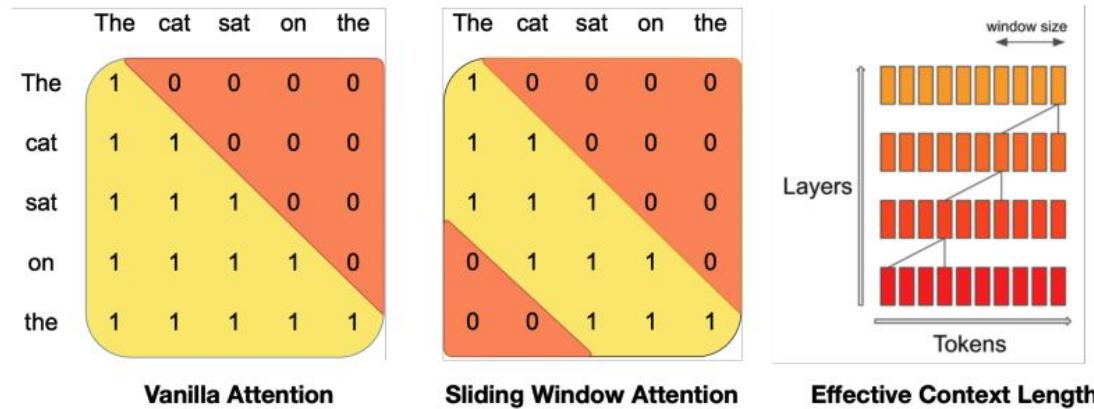
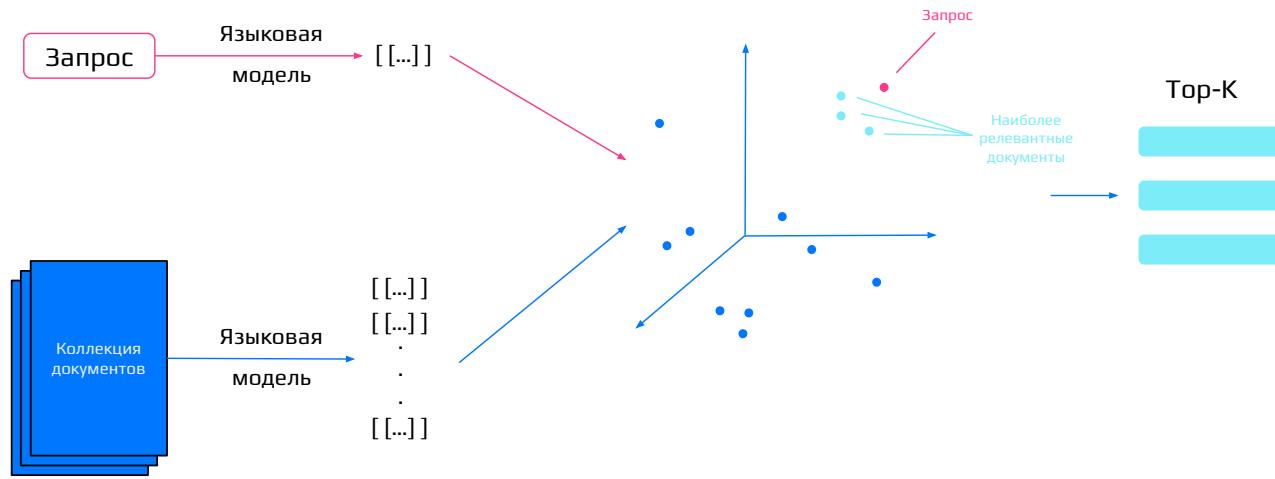
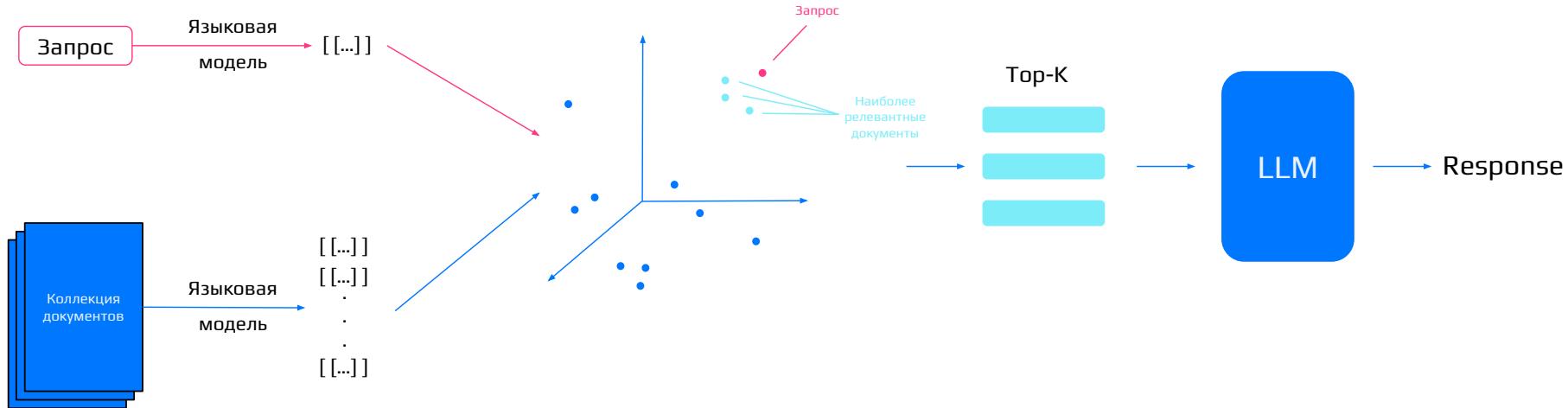


Figure 1: Sliding Window Attention. The number of operations in vanilla attention is quadratic in the sequence length, and the memory increases linearly with the number of tokens. At inference time, this incurs higher latency and smaller throughput due to reduced cache availability. To alleviate this issue, we use sliding window attention: each token can attend to at most W tokens from the previous layer (here, $W = 3$). Note that tokens outside the sliding window still influence next word prediction. At each attention layer, information can move forward by W tokens. Hence, after k attention layers, information can move forward by up to $k \times W$ tokens.

Recap: Semantic Search



Retrieval Augmented Generation



Paged Attention

0. Before generation.

Seq
A

Prompt: "Alan Turing is a computer scientist"
Completion: ""

Logical KV cache blocks

Block 0				
Block 1				
Block 2				
Block 3				

Block table

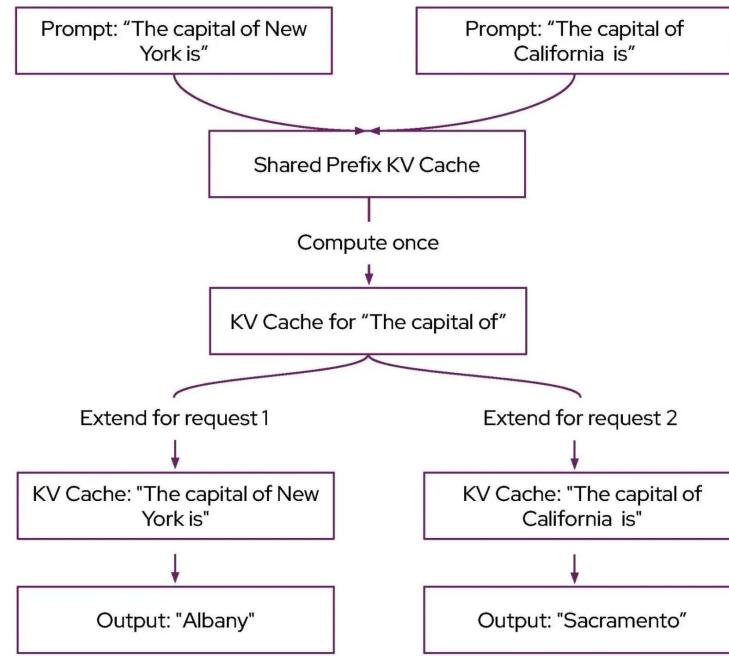
Physical block no.	# Filled slots
-	-
-	-
-	-
-	-

Physical KV cache blocks

Block 0			
Block 1			
Block 2			
Block 3			
Block 4			
Block 5			
Block 6			
Block 7			

Prefix Caching

Prefix Caching



Итоги раздела:

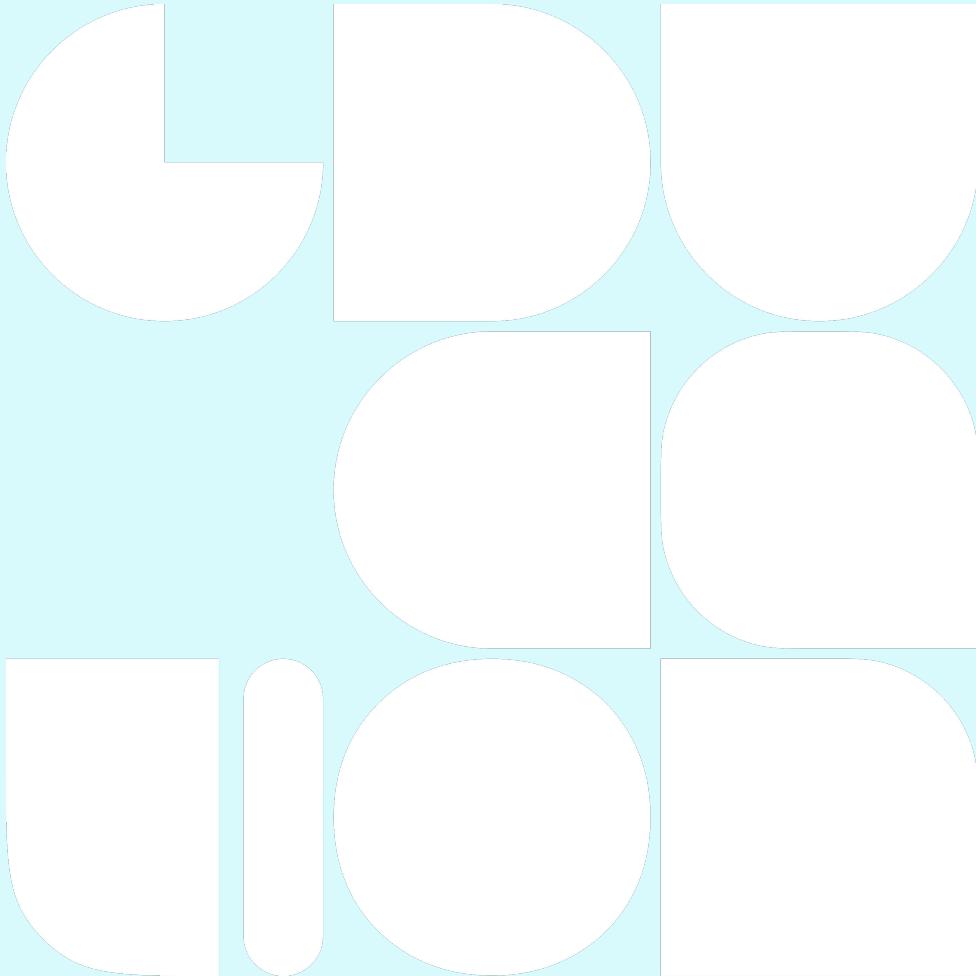
- Узнали как работает kv-cache, sliding-window attention и paged attention
- Посмотрели на retrieval augmented generation



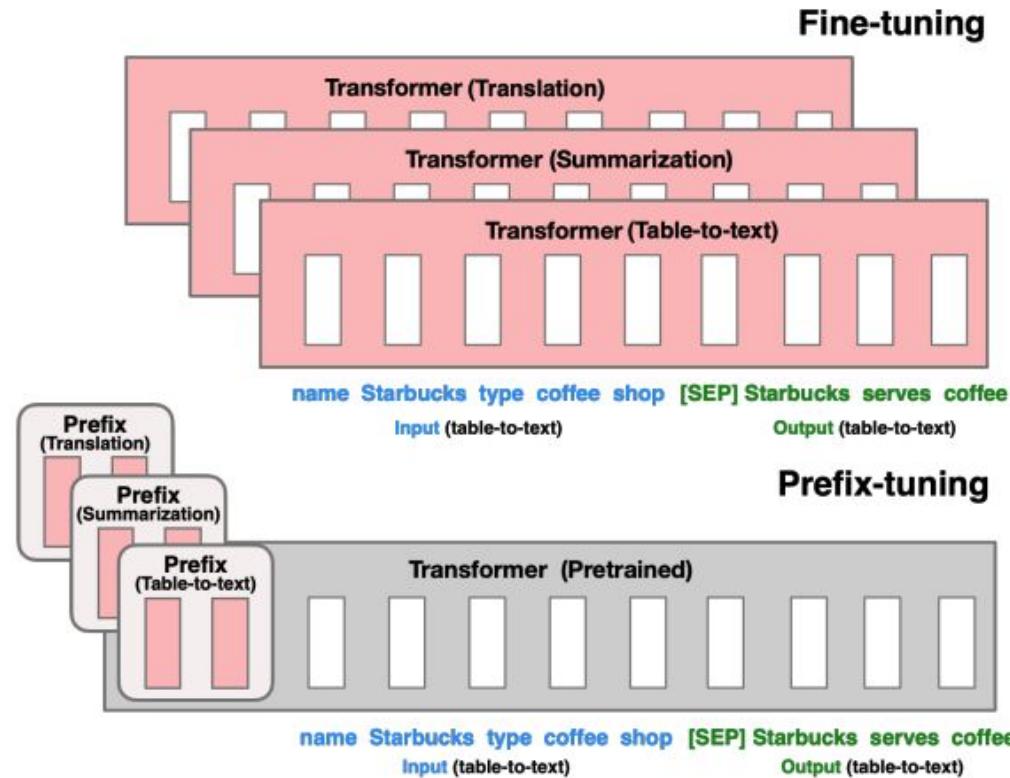
Вопросы?



Parameter-Efficient Fine-Tuning



Prefix Tuning



Prefix Tuning

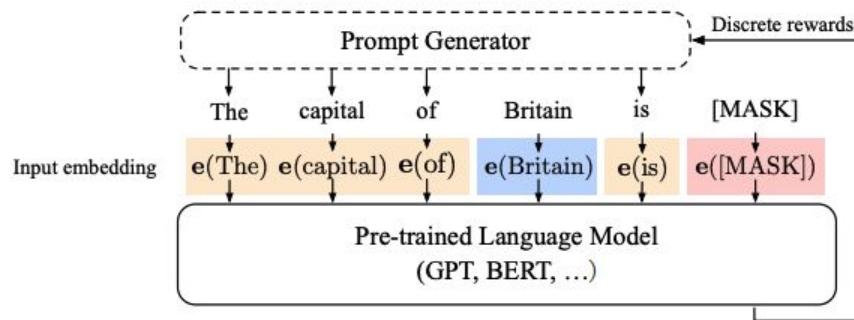
Плюсы:

- Мало параметров
- Веса базовой модели не изменяются и можно ее переиспользовать
- Не требует много ресурсов для обучения

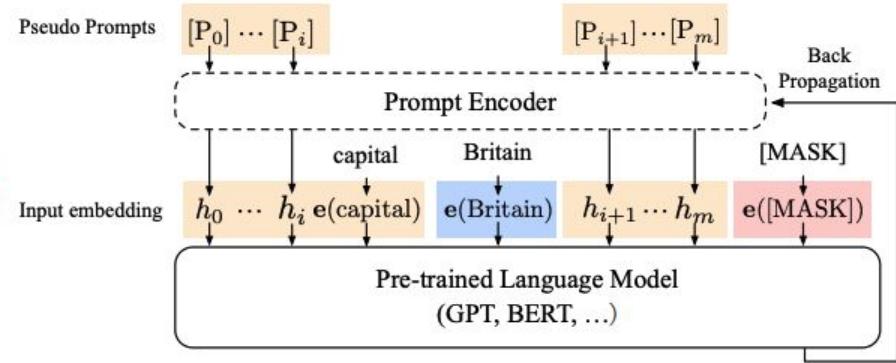
Минусы:

- Крадет часть размера контекста
- Требует вмешательства в реализацию модели
- Может быть нестабильными чувствительным к гиперпараметрам

P-Tuning



(a) Discrete Prompt Search



(b) P-tuning

P-Tuning

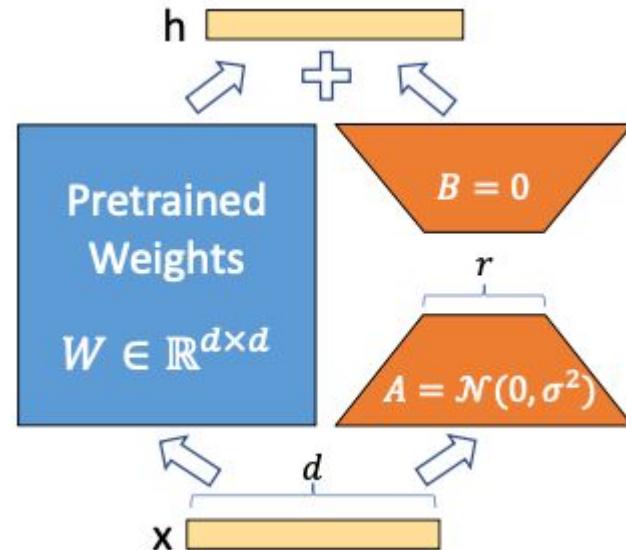
Плюсы:

- Качество в среднем лучше, чем у prefix-tuning
- Меньше, чем prefix-tuning "крадет" контекст

Минусы:

- Требует вмешательства в реализацию модели

Low-Rank Adaptation (LoRA)



Low-Rank Adaptation (LoRA)

Плюсы:

- Не меняет размерностей внутренних представлений
- По качеству превосходит prefix-tuning и p-tuning
- Обычно не влияет на время инференса
- Не “крадет” контекст

Минусы:

- Требует вмешательства в топологию нейросети
- Не всегда очевидно к каким слоям применять

Итоги раздела:

- Изучили основные виды адаптеров: p-tuning, prefix-tuning, LoRA; выяснили какие у каждого преимущества и недостатки



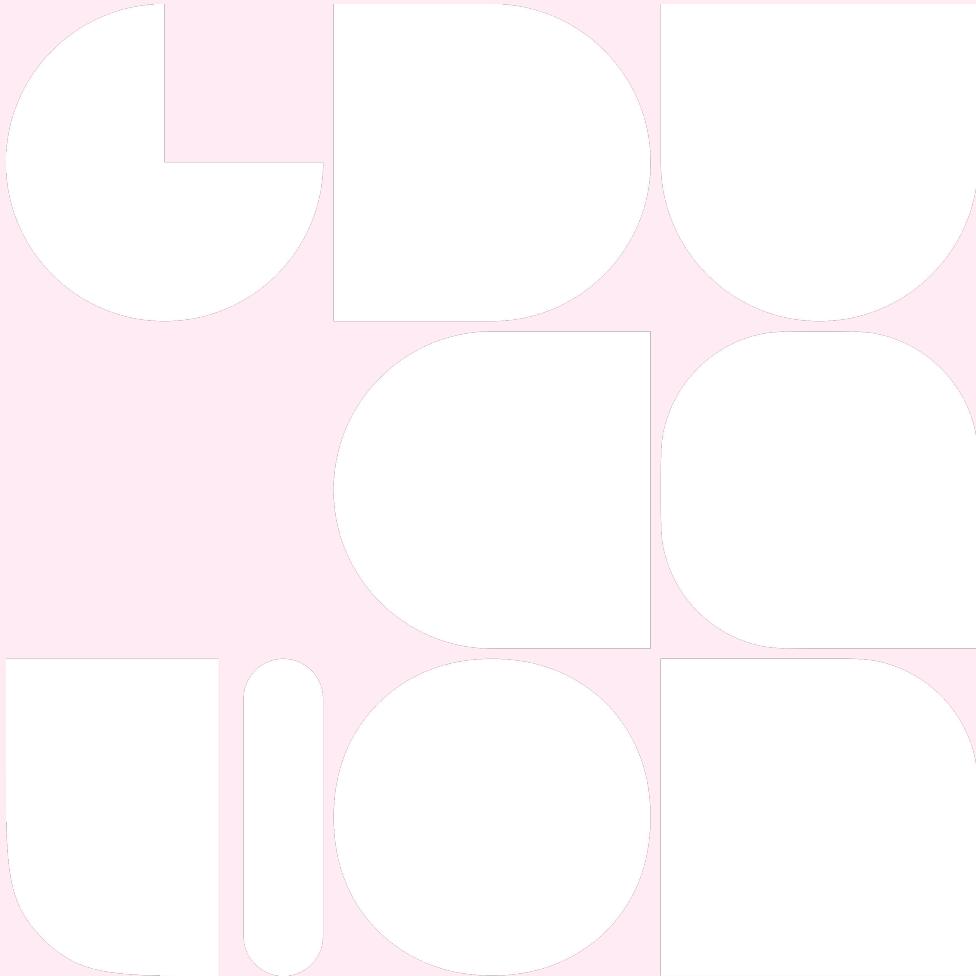
Вопросы?



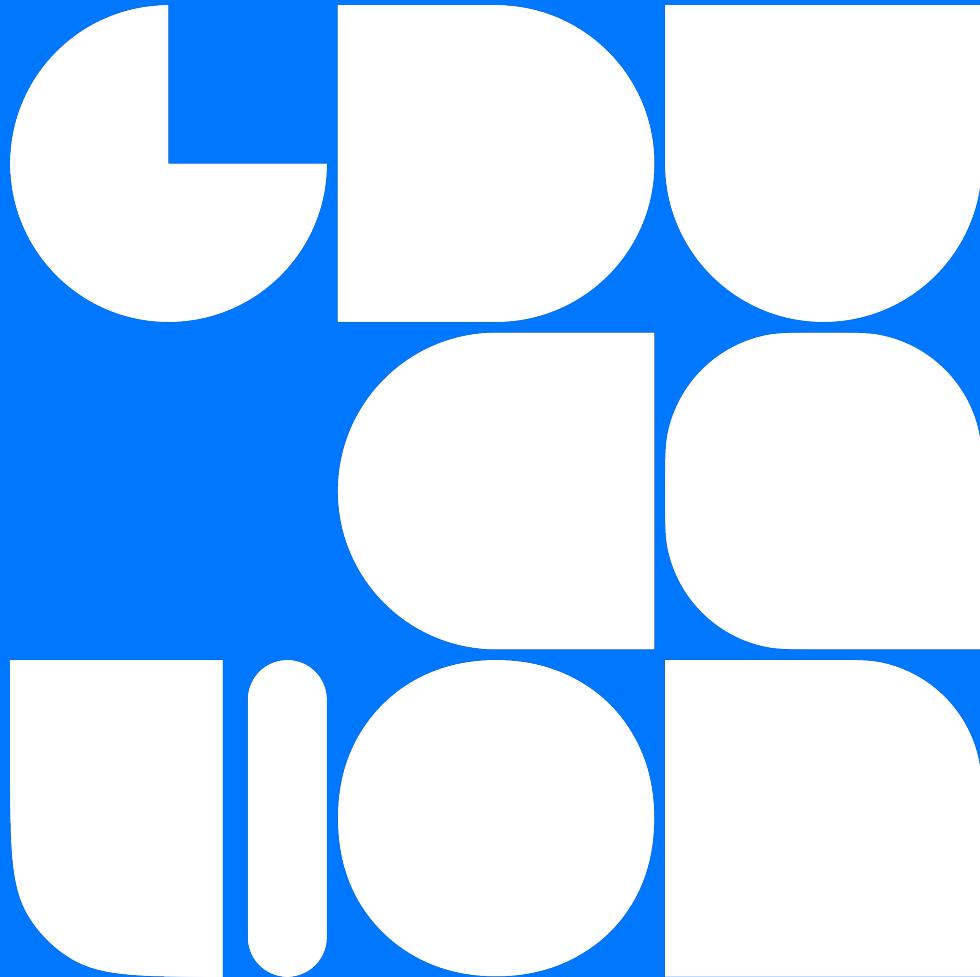
Подведём итоги:

- Decoder-only нейросети на подобии GPT развиваются в первую очередь путем масштабирования весов и данных
- Одна из основных фишек LLM – Alignment
- Качество генерируемого текста можно оценивать через exact_match, loglikelihood-based accuracy, BLEU и Preplexity
- После получения распределения вероятностей токены можно сэмплировать различными способами
- LLM смогли развиваться благодаря набору эвристик для эффективной утилизации железа и расширения контекста
- Большие модели редко приходится обучать самому с нуля. Гораздо чаще нужно прибегать к адаптерам, например, LoRA

Перерыв



Практика



Спасибо за
внимание!

