



# ОБУЧЕНИЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ. ЧАСТЬ 1

*Курс «Проектные задачи  
компьютерной лингвистики»  
Преподаватель Студеникина  
Ксения Андреевна*

ChatGPT ▾

Кто ты?

Я — ChatGPT, твой виртуальный помощник! Могу отвечать на вопросы, рассказывать шутки, помогать с учебой, писать тексты и даже просто поддерживать беседу. О чем поговорим? 😊



<https://chatgpt.com/>

# Этап 1: предобучение (pre-training)

# СБОР ДАННЫХ

- Первым шагом на этапе предобучения является загрузка и обработка текстовых данных из Интернета
- Компании, разрабатывающие LLM, хранят данные для предобучения в закрытом доступе



**ANTHROPIC**

- На платформе 🤗 HuggingFace содержится аналогичный открытый датасет 🍷 FineWeb\*

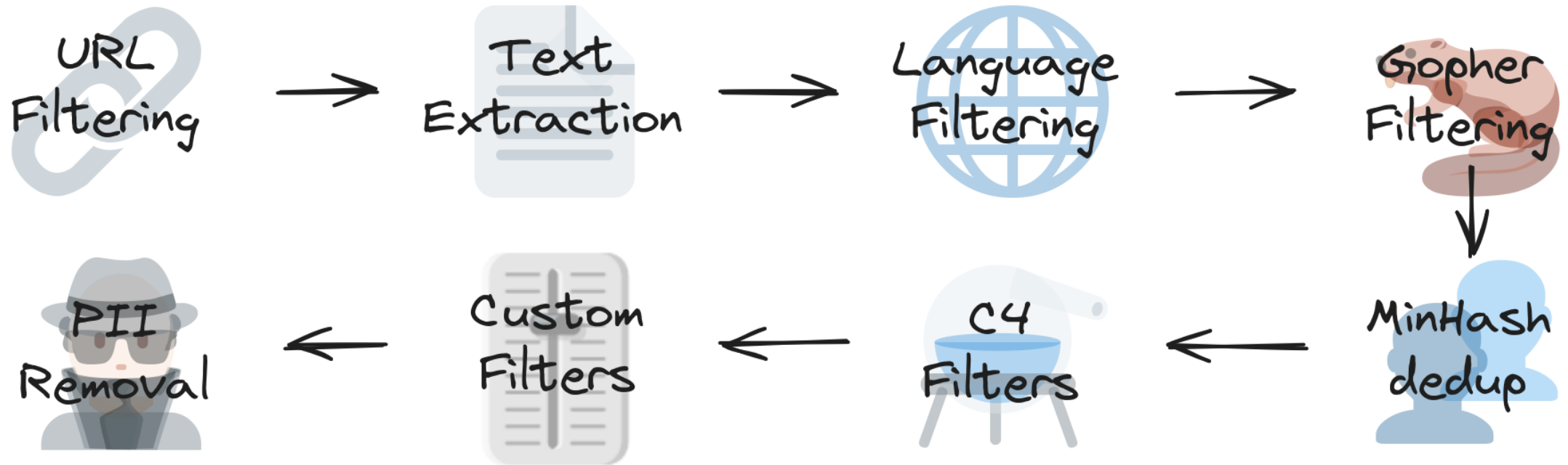
\*<https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1>

# СБОР ДАННЫХ

- Необходимо получить **большое** количество текстовых документов **высокого** качества из **открытых** источников **разной** тематики
- Отправной точкой являются данные CommonCrawl
- Краулинг (crawling) или сканирование – процесс обнаружения и сбора новых и обновленных страниц для добавления в индекс поисковых систем. При сканировании страницы выполняется переход по содержащимся на ней ссылкам.
- Результатом становится большое количество сырых данных, которые требуют фильтрации.

\*<https://commoncrawl.org/>

# ФИЛЬТРАЦИЯ ДАННЫХ



The Fineweb pipeline

# ФИЛЬТРАЦИЯ ДАННЫХ



Удаление страниц из черного списка:

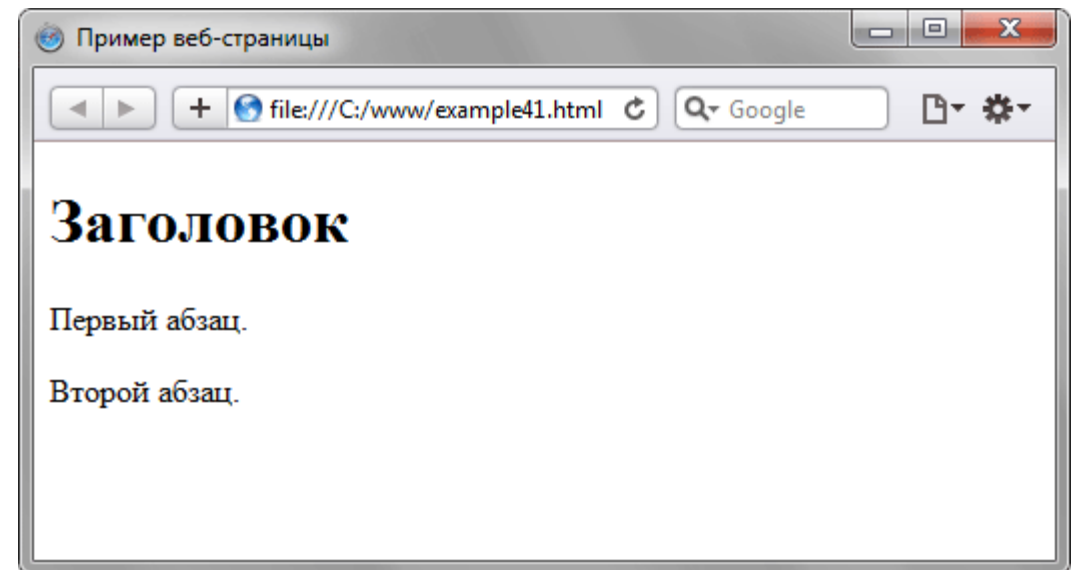
- вредоносные веб-сайты
- спам-сайты
- маркетинговые веб-сайты
- расистские сайты
- сайты для взрослых

# ФИЛЬТРАЦИЯ ДАННЫХ



- При краулинге на выходе получаем сырые HTML-страницы.
- Необходимо убрать всю разметку и оставить только текст.

```
<html>
<head>
  <meta http-equiv="Content-Type" content="text/html; charset=utf-8">
  <title>Пример веб-страницы</title>
</head>
<body>
  <h1>Заголовок</h1>
  <!-- Комментарий -->
  <p>Первый абзац.</p>
  <p>Второй абзац.</p>
</body>
</html>
```





# ФИЛЬТРАЦИЯ ДАННЫХ



- Пытаемся определить, на каком языке написана каждая отдельная веб-страница.
- Исследователь сам решает, какую долю из всех различных языков собирается включить в набор данных.
- Если отфильтруем весь испанский язык, то модель впоследствии будет не очень хороша в испанском, потому что никогда не видела данных на этом языке.
- FineWeb в значительной степени ориентирован на английский
  - классификатор языков fastText\*, отбираются только страницы с долей английских текстов  $\geq 0,65$
  - языковая модель будет хорошо работать для английского, но может не очень хорошо показывать себя для других языков.

\*<https://huggingface.co/facebook/fasttext-language-identification>

# ФИЛЬТРАЦИЯ ДАННЫХ



- Удаление информации, позволяющей установить личность (personally identifiable information, PII):
  - ФИО
  - номер телефона
  - почтовый адрес
  - электронный адрес

text	id	dump	url
<p>How AP reported in all formats from tornado-stricken regionsMarch 8, 2012 When the first serious bout of tornadoes of 2012 blew through middle America in the middle of the night, they touched down in places hours from any AP bureau. Our closest video journalist was Chicago-based Robert Ray, who dropped his plans to travel to Georgia for Super Tuesday, booked several flights to the cities closest to the strikes and headed for the airport. He'd decide once there which flight to take. He never got on board a plane. Instead, he ended up driving toward Harrisburg, Ill., where initial reports suggested a town was destroyed. That decision turned out to be a lucky break for the AP. Twice. Ray was among the first journalists to arrive and he confirmed those reports ~ in all formats. He shot powerful video, put victims on the phone with AP Radio and played back sound to an editor who transcribed the interviews and put the material on text wires. He then walked around the devastation with the Central Regional Desk on the line, talking to victims with the phone held so close that editors could transcribe his interviews in real time. Ray also made a dramatic image of a young girl who found a man's prosthetic leg in the rubble, propped it up next to her destroyed home and spray-painted an impromptu sign: "Found leg. Seriously." The following day, he was back on the road and headed for Georgia and a Super Tuesday date with Newt Gingrich's campaign. The drive would take him through a stretch of the South that forecasters expected would suffer another wave of tornadoes. To prevent running into THAT storm, Ray used his iPhone to monitor Doppler radar, zooming in on extreme cells and using Google maps to direct himself to safe routes. And then the journalist took over again. "When weather like that occurs, a reporter must seize the opportunity to get the news out and allow people to see, hear and read the power of nature so that they can take proper shelter," Ray says. So Ray now started to use his phone to follow the storms. He attached a small GoPro camera to his steering wheel in case a tornado dropped down in front of the car somewhere, and took video of heavy rain and hail with his iPhone. Soon, he spotted a tornado and the chase was on. He followed an unmarked emergency vehicle to Cleveland, Tenn., where he was first on the scene of the storm's aftermath. Again, the tornadoes had struck in locations that were hours from the nearest AP bureau. Damage and debris, as well as a wickedly violent storm that made travel dangerous, slowed our efforts to get to the news. That wasn't a problem in Tennessee, where our customers were well served by an all-formats report that included this text story. "CLEVELAND, Tenn. (AP) _ Fierce wind, hail and rain lashed Tennessee for the second time in three days, and at least 15 people were hospitalized Friday in the Chattanooga area." The byline? Robert Ray. For being adept with technology, chasing after news as it literally dropped from the sky and setting a standard for all-formats reporting that put the AP ahead on the most competitive news story of the day, Ray wins this week's \$300 Best of the States prize. © 2013 The Associated Press. All rights reserved. Terms and conditions apply. See AP.org for details.</p>	<urn:uuid:d66bc6fe-8477-4adf-b430-f6a558ccc8ff>	CC-MAIN-2013-20	http://%20jwashington@ap.org/Content/Press-Release/2012/How-AP-reported-in-all-formats-from-tornado-stricken-regions

# ТОКЕНИЗАЦИЯ

- Нужно решить, как представлять полученный массив текстов и как подавать его на вход нейронной сети.
- Данные должны представлять собой одномерную последовательность символов, взятых из конечного набора.

# Одномерная последовательность текстовых символов: X много разных символов ✓ небольшая длина

|Viewing Single Post From: Spoilers for the Week of February 11th| \n|Li\||Feb 1 2013, 09:58 AM|\nDon't care about Chloe/Taniel/Jen-Jen. Don't care about Sami, really, but hoping that we get some good "Sd"SAMANTHA GENE!!" Marlana Death-Stares out of itofit.t And "newfound" fed" feelings.Please.e. If only.\nSTEFANO!! STEFANO, STEFANO, STEFANO!!!! :cheer:\n|Spoilers fors for the Week of February 11th - DAYS: News, Spoilers & Discussion| #sigh= Fundamentalist community, let me pass on some advice to you I learned from the atheistic community:\nIf you have seve set yourself on fire, do not rumot run.\n0kay? Okay?? Please?\

Одномерная последовательность текстовых символов: X много разных символов ✓ небольшая длина

Viewing Single Post From: Spoilers for the Week of February 11th | \n | Li | | Feb 1 2013, 09:58 AM | \n Don't care about Chloe/Taniel/Jen-Jen. Don't care about Sami, really, but hoping that we get some good "Sd" SAMANTHA GENE!!" Marlana Death-Stares out of itofit.t And "newfound" fed" feelings. Please.e. If only.\n STEFANO!! STEFANO, STEFANO, STEFANO!!!! :cheer:\n | Spoilers fors for the Week of February 11th - DAYS: News, Spoilers & Discussion | #sigh= Fundamentalist community, let me pass on some advice to you I learned from the atheistic community:\n If you have seve set yourself on fire, do not rumot run.\n Okay? Okay?? Please?\n

Битовое представление текста: ✓ только 2 уникальных символа X большая длина

01010110 01101001 01100101 01110111 01101001 01101110 01100111 00100000 01010011 01101001 01101110 01100111 01101100 01100101 00100000 01010000 01101111 01110011 01110100 00100000 01000110 01110010 01101111 01101101 00111010 00100000  
01010011 01110000 01101111 01101001 01101100 01100101 01110010 01110011 00100000 01100110 01101111 01110010 00100000 01110100 01101000 01100101 00100000 01010111 01100101 01100101 01101011 00100000 01101111 01100110 00100000 01000110 01100101  
01100010 01110010 01110101 01110000 01110010 01111001 00100000 00110001 00110001 01110100 01110100 01101000 01111100 01011100 01011110 01111100 010001100 01101001 01011100 01111100 01000110 01100101 01100010 00100000 00110001 00100000 00110010  
00110000 00110001 01100101 00101001 00100000 01000011 01101000 00110100 01101111 01100101 00111000 00100000 01000001 01001101 01111100 01011100 01101110 01000100 01101111 01101110 00100000 01100011 01100001 01110010 01100101 00100000  
01100001 01100010 01101111 01110101 01110100 00100000 01000011 01101000 00110100 01101111 01100101 00101111 01010100 01100001 01101110 01101001 01100101 01101100 00101111 01001010 01100101 01101110 00101101 01001010 01100101 01101110 00101110  
00100000 01000100 01101111 01101110 00100111 01110100 00100000 01100011 01100001 0110010 01100101 00100000 01100001 01100010 01101111 01110101 01110100 00100000 01010011 01100001 01101101 01101001 00101100 00100000 0110010 01100101 01100001  
01101100 01101100 01111001 00101100 00100000 01100010 01110101 01110100 00001010 01101000 01101111 01110000 01101001 01101110 01100111 00100000 01110100 01101000 01100001 01110100 00100000 01110111 01100101 00100000 01100111 01100101 01110100  
00100000 01110011 01101111 01101101 01100101 00100000 01100111 01101111 01101111 01100100 00100000 01001011 0100100 00100000 01010011 0100100 00100000 01010011 01000001 01001110 01000001 01001110 01010100 01000100 01000001 00100000 01000111 01000101 01001110  
01000101 00100001 00100001 00100001 00100010 01000001 01001101 01100001 01110010 01110100 01100101 0111110 01101110 01100001 00100000 01000001 0100101 01100001 01100001 01101001 01100001 01101001 01110100 01100001 01110010 01100101 01100110 00100000 01101111  
01110101 01110100 00100000 01101111 01100110 00100000 01101001 01110100 01101111 01100110 01101001 01110100 00101110 01110100 00100000 01000001 01101110 01100100 00100000 00100010 01101110 01100101 01110111 01100110 0110101 01101110  
01100100 00100010 00100000 01100110 01100101 01100100 00100010 00100000 01100110 01100101 01100101 01101100 01101001 01101110 01110011 01110011 00101110 01010000 01101100 01100101 01100001 01110011 01100101 00101110 01100101 00101110 00100000  
01001001 01100110 00100000 01101111 01101110 01101100 01111001 00101110 01011100 01101110 01010011 01010100 01000010 01000110 01000001 01001110 01001111 00100001 00100000 01010011 01010100 01000101 01000110 01000001 01001110 01001111  
00101100 00100000 01010011 01010100 01000010 01000111 01000001 01001110 01001111 00101110 00100000 01010011 01010100 01000010 01001110 01010100 01000010 01000111 01000001 00100001 01000001 01000001 01000001 00100001 00100001 01000001 01000001 01000101  
01100101 01110010 00111010 01011100 01011100 01111100 01100101 01110000 01101111 01101001 01110100 01100101 01110010 01110011 00100000 01100110 01101111 01110010 01110011 01100101 00100000 01100110 01100101 01100001 01101000 01100101  
00100000 01010111 01100101 01100101 01101011 00100000 01101111 01100110 00100000 01000110 01100101 01100010 01110010 01110101 01100001 01110010 01111001 00100000 00110001 00110001 01110100 01101000 00100000 00101101 00100000 01000100 01000001  
01011001 01010011 00111010 00100000 01001110 01100101 01110111 01110011 00101100 00100000 01010011 01110000 01101111 01101001 01101100 01100101 01110010 01110011 00100000 00100110 00100000 01000100 01101001 01110011 01100011 01101011 01110011  
01110011 01101001 01101111 01110110 01111100 00100000 00100011 01110011 01101001 01100111 01101000 00111101 00100000 01000110 01110101 01101110 01100100 01100001 0110101 01100101 01101110 01110100

Одномерная последовательность текстовых символов: X много разных символов ✓ небольшая длина

[Viewing Single Post From: Spoilers for the Week of February 11th | \n | Li\ | | Feb 1 2013, 09:58 AM | \n](#)
 Don't care about Chloe/Taniel/Jen-Jen. Don't care about Sami, really, but hoping that we get some good "Sd"**SAMANTHA GENE!!!** Marlena Death-Stares out of itofit.t And "newfound" fed" feelings.Please.e. If only.\n**STEFANO!! STEFANO, STEFANO, STEFANO!!!!** :cheer:\n|Spoilers fors for the Week of February 11th - DAYS: News, Spoilers & Discussion| #sigh= Fundamentalist community, let me pass on some advice to you I learned from the atheistic community:\nIf you have seve set yourself on fire, do not rumot run.\n0kay? Okay?? Please?\n

Битовое представление текста: ✓ только 2 уникальных символа X большая длина

[illegible]

Байтовое представление текста: ✓ 256 уникальных символов ✓ небольшая длина

124 86 105 101 119 105 110 103 32 83 105 110 103 108 101 32 80 111 115 116 32 70 114 111 109 58 32 83 112 111 105 108 101 114 115 32 102 111 114 32 116 104 101 32 87 101 101 107 32 111 102 32 70 101 98 114 117 97 114 121 32 49 49 116 104 124 10 124 76 105 92 124 124 70 101 98 32 49 32 50 48 49 51 44 32 48 57 58 53 56 32 65 77 124 10 68 111 110 39 116 32 99 97 114 101 32 97 98 111 117 116 32 67 104 108 111 101 47 84 97 110 105 101 108 47 74 101 110 45 74 101 110 46 32 68 111 110 39 116 32 99 97 114 101 32 97 98 111 117 116 32 83 97 109 105 44 32 114 101 97 108 108 121 44 32 98 117 116 10 104 111 112 105 110 103 32 116 104 97 116 32 119 101 32 103 101 116 32 115 111 109 101 32 103 111 111 100 32 34 83 100 34 83 65 77 65 78 84 72 65 32 71 69 78 69 33 33 34 32 77 97 114 108 101 110 97 32 68 101 97 116 104 45 83 116 97 114 101 115 32 111 117 116 32 111 102 32 105 116 111 102 105 116 46 116 32 65 110 100 32 34 110 101 119 102 111 117 110 100 34 32 102 101 100 34 32 102 101 101 108 105 110 103 115 46 80 108 101 97 115 101 46 101 46 32 73 102 32 111 110 108 121 46 10 83 84 69 70 65 78 79 33 33 32 83 84 69 70 65 78 79 44 32 83 84 69 70 65 78 79 44 32 83 84 69 70 65 78 79 33 33 33 33 32 58 99 104 101 101 114 58 10 124 83 112 111 105 108 101 114 115 32 102 111 114 115 32 102 111 114 32 116 104 101 32 87 101 101 107 32 111 102 32 70 101 98 114 117 97 114 121 32 49 49 116 104 32 45 32 68 65 89 83 58 32 78 101 119 115 44 32 83 112 111 105 108 101 114 115 32 38 32 68 105 115 99 117 115 115 105 111 110 124 32 35 115 105 103 104 61 32 70 117 110 100 97 109 101 110 116 97 108 105 115 116 32 99 111 109 109 117 110 105 116 121 44 32 108 101 116 32 109 101 32 112 97 115 115 32 111 110 32 115 111 109 101 32 97 100 118 105 99 101 32 116 111 32 121 111 117 32 73 32 108 101 97 114 110 101 100 32 102 114 111 109 32 116 104 101 32 97 116 104 101 105 115 116 105 99 32 99 111 109 109 117 110 105 116 121 58 10 73 102 32 121 111 117 32 104 97 118 101 32 115 101 118 101 32 115 101 116 32 121 111 117 114 115 101 108 102 32 111 110 32 102 105 114 101 44 32 100 111 32 110 111 116 32 114 117 109 111 116 32 114 117 110 46 10 48 107 97 121 63 32 79 107 97 121 63 63 32 80 108 101 97 115 101 63 92

# ТОКЕНИЗАЦИЯ

- Длину последовательности можно уменьшить еще сильнее
- Необходимо найти последовательности байтов, которые являются наиболее частотными
- Последовательность 116 32 встречается часто → можно заменить ее одним новым символом

124 86 105 101 119 105 110 103 32 83 105 110 103 108 101 32 80 111 115 116 32 70 114 111 109 58 32 83 112 111 105 108 101 114 115 32 102 111 114 32 116 104  
101 32 87 101 101 107 32 111 102 32 70 101 98 114 117 97 114 121 32 49 49 116 104 124 10 124 76 105 92 124 124 70 101 98 32 49 32 50 48 49 51 44 32 48 57 58  
53 56 32 65 77 124 10 68 111 110 39 116 32 99 97 114 101 32 97 98 111 117 116 32 67 104 108 111 101 47 84 97 110 105 101 108 47 74 101 110 45 74 101 110 46 32  
68 111 110 39 116 32 99 97 114 101 32 97 98 111 117 116 32 83 97 109 105 44 32 114 101 97 108 108 121 44 32 98 117 116 10 104 111 112 105 110 103 32 116 104  
97 116 32 119 101 32 103 101 116 32 115 111 109 101 32 103 111 111 100 32 34 83 100 34 83 65 77 65 78 84 72 65 32 71 69 78 69 33 33 34 32 77 97 114 108 101  
110 97 32 68 101 97 116 104 45 83 116 97 114 101 115 32 111 117 116 32 111 102 32 105 116 111 102 105 116 46 116 32 65 110 100 32 34 110 101 119 102 111 117  
110 100 34 32 102 101 100 34 32 102 101 101 108 105 110 103 115 46 80 108 101 97 115 101 46 101 46 32 73 102 32 111 110 108 121 46 10 83 84 69 70 65 78 79  
33 33 32 83 84 69 70 65 78 79 44 32 83 84 69 70 65 78 79 44 32 83 84 69 70 65 78 79 33 33 33 33 32 58 99 104 101 101 114 58 10 124 83 112 111 105 108 101 114  
115 32 102 111 114 115 32 102 111 114 32 116 104 101 32 87 101 101 107 32 111 102 32 70 101 98 114 117 97 114 121 32 49 49 116 104 32 45 32 68 65 89 83 58 32  
78 101 119 115 44 32 83 112 111 105 108 101 114 115 32 38 32 68 105 115 99 117 115 115 105 111 110 124 32 35 115 105 103 104 61 32 70 117 110 100 97 109 101  
110 116 97 108 105 115 116 32 99 111 109 109 117 110 105 116 121 44 32 108 101 116 32 109 101 32 112 97 115 115 32 111 110 32 115 111 109 101 32 97 100 118  
105 99 101 32 116 111 32 121 111 117 32 73 32 108 101 97 114 110 101 100 32 102 114 111 109 32 116 104 101 32 97 116 104 101 105 115 116 105 99 32 99 111 109  
109 117 110 105 116 121 58 10 73 102 32 121 111 117 32 104 97 118 101 32 115 101 118 101 32 115 101 116 32 121 111 117 114 115 101 108 102 32 111 110 32 102  
105 114 101 44 32 100 111 32 110 111 116 32 114 117 109 111 116 32 114 117 110 46 10 48 107 97 121 63 32 79 107 97 121 63 63 32 80 108 101 97 115 101 63 92

- Алгоритм можно повторять многократно
- Каждый раз при замене двух старых символов на один новый сокращается длина последовательности, а размер словаря увеличивается
- В среднем размер словаря LLM составляет около 100,000 символов – токенов



# ПРИМЕРЫ ТОКЕНИЗАЦИИ GPT-4

hello world

Token count  
2

hello world

15339, 1917

helloworld

Token count  
2

helloworld

71, 96392

hello world

Token count  
3

hello world

15339, 220, 1917

hello world

Token count  
3

hello world

15339, 256, 1917

Hello world

Token count  
2

Hello world

9906, 1917

# ОБУЧЕНИЕ НЕЙРОННОЙ СЕТИ

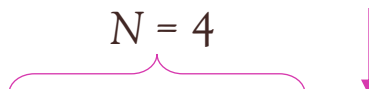
- Собрано большое количество данных, состоящих из различных фрагментов текста
- Каждый токен – часть слова, целое слово или группа слов – заменяется на свой числовой идентификатор
  - Объем корпуса FineWeb составляет 15 триллионов токенов и занимает 44 ТБ
- Используя эти данные, мы хотим смоделировать статистические взаимосвязи, как токены следуют друг за другом в последовательности

91 860 287 11579 3962 105 110 103 32 83 105 110 103 108 101 32 80 111 115 116 32 70 114 111 109 58 32 83 112  
111 105 108 101 114 115 32 102 111 114 32 116 104 101 32 87 101 101 107 32 111 102 32 70 101 98 114 117 97 114  
121 32 49 49 116 104 124 10 124 76 105 92 124 124 70 101 98 32 49 32 50 48 49 51 44 32 48 57 58 53 56 32 65 77  
124 10 68 111 110 39 116 32 99 97 114 101 32 97 98 111 117 116 32 67 104 108 111 101 47 84 97 110 105 101 108 47  
74 101 110 45 74 101 110 46 32 68 111 110 39 116 32 99 97 114 101 32 97 98 111 117 116 32 83 97 109 105 44 32 114  
101 97 108 108 121 44 32 98 117 116 10 104 111 112 105 110 103 32 116 104 97 116 32 119 101 32 103 101 116 32  
115 111 109 101 32 103 111 111 100 32 34 83 100 34 83 65 77 65 78 84 72 65 32 71 69 78 69 33 33 34 32 77 97 114  
108 101 110 97 32 68 101 97 116 104 45 83 116 97 114 101 115 32 111 117 116 32 111 102 32 105 116 111 102 105  
116 46 116 32 65 110 100 32 34 110 101 119 102 111 117 110 100 34 32 102 101 100 34 32 102 101 101 108 105 110  
103 115 46 80 108 101 97 115 101 46 101 46 32 73 102 32 111 110 108 121 46 10 83 84 69 70 65 78 79

# ОБУЧЕНИЕ НЕЙРОННОЙ СЕТИ

- Из всей последовательности индексов выбирается контекст – некоторый фрагмент длины  $N$
- $N$  может быть от 0 до некоторого максимального значения, которое мы устанавливаем сами: например, до 4000 или до 16000
- Однако нужно помнить, что обработка очень длинных фрагментов потребует больших вычислительных затрат
- Мы хотим предсказывать, какой token будет следующим в последовательности

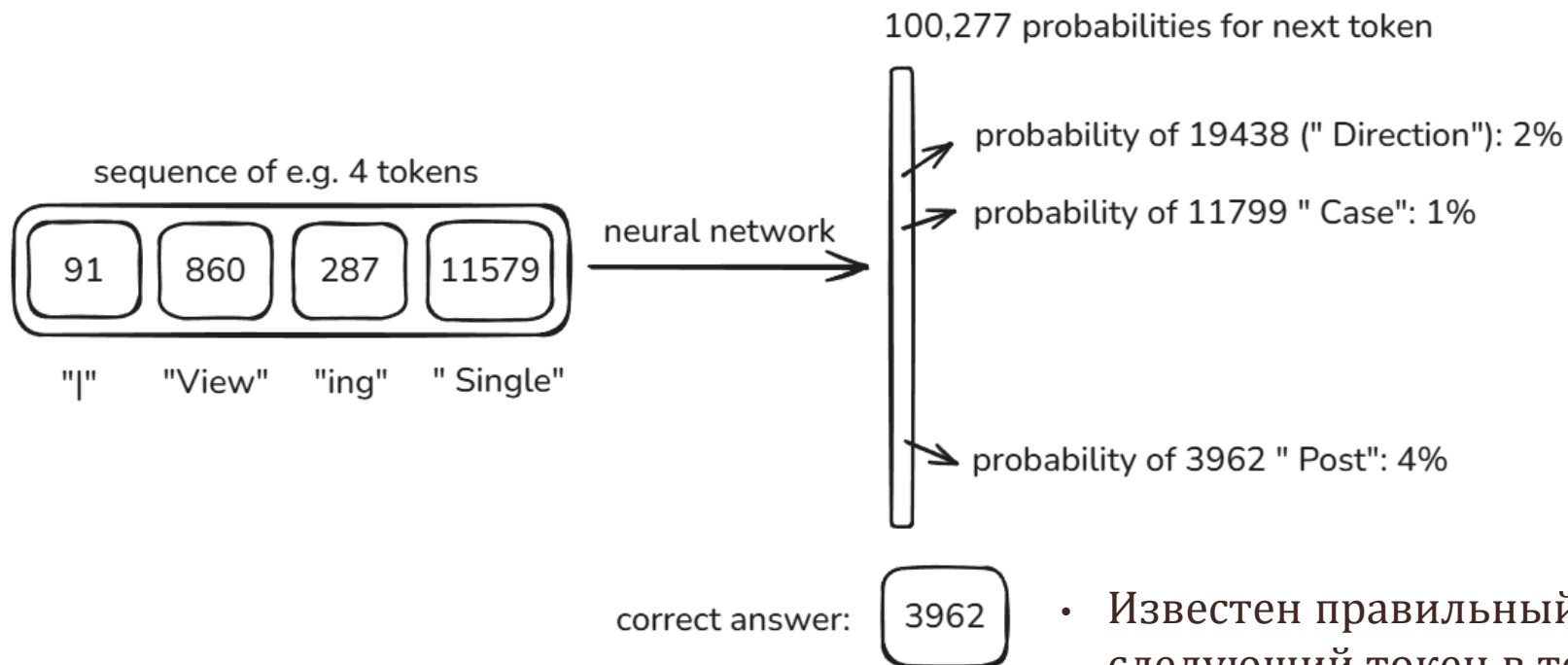
$N = 4$



91 860 287 11579 3962 105 110 103 32 83 105 110 103 108 101 32 80 111 115 116 32 70 114 111 109 58 32 83 112  
111 105 108 101 114 115 32 102 111 114 32 116 104 101 32 87 101 101 107 32 111 102 32 70 101 98 114 117 97 114  
121 32 49 49 116 104 124 10 124 76 105 92 124 124 70 101 98 32 49 32 50 48 49 51 44 32 48 57 58 53 56 32 65 77  
124 10 68 111 110 39 116 32 99 97 114 101 32 97 98 111 117 116 32 67 104 108 111 101 47 84 97 110 105 101 108 47  
74 101 110 45 74 101 110 46 32 68 111 110 39 116 32 99 97 114 101 32 97 98 111 117 116 32 83 97 109 105 44 32 114  
101 97 108 108 121 44 32 98 117 116 10 104 111 112 105 110 103 32 116 104 97 116 32 119 101 32 103 101 116 32  
115 111 109 101 32 103 111 111 100 32 34 83 100 34 83 65 77 65 78 84 72 65 32 71 69 78 69 33 33 34 32 77 97 114  
108 101 110 97 32 68 101 97 116 104 45 83 116 97 114 101 115 32 111 117 116 32 111 102 32 105 116 111 102 105  
116 46 116 32 65 110 100 32 34 110 101 119 102 111 117 110 100 34 32 102 101 100 34 32 102 101 101 108 105 110  
103 115 46 80 108 101 97 115 101 46 101 46 32 73 102 32 111 110 108 121 46 10 83 84 69 70 65 78 79

# ОБУЧЕНИЕ НЕЙРОННОЙ СЕТИ

- На входе: контекст – последовательность из  $N$  токенов
- На выходе:  $M$  значений вероятностей для следующего токена

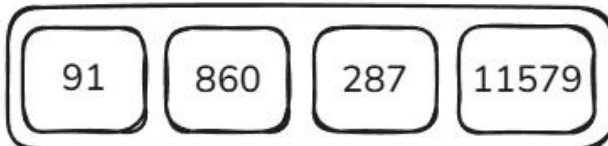


- Вероятность корректного токена должна увеличиваться, а вероятность остальных токенов – уменьшаться

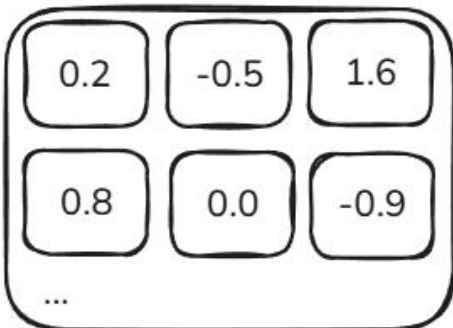
- Известен правильный ответ – следующий токен в тексте
- Этот процесс происходит параллельно для всех фрагментов длины  $N$

# ОБУЧЕНИЕ НЕЙРОННОЙ СЕТИ

input sequence tokens  $x$   
anywhere from 1 to e.g. 8,000 tokens



parameters (/ "weights")  $w$   
usually billions of these



- Входные данные  $x$  объединяются в гигантское математическое выражение вместе с параметрами или весами нейронной сети  $w$

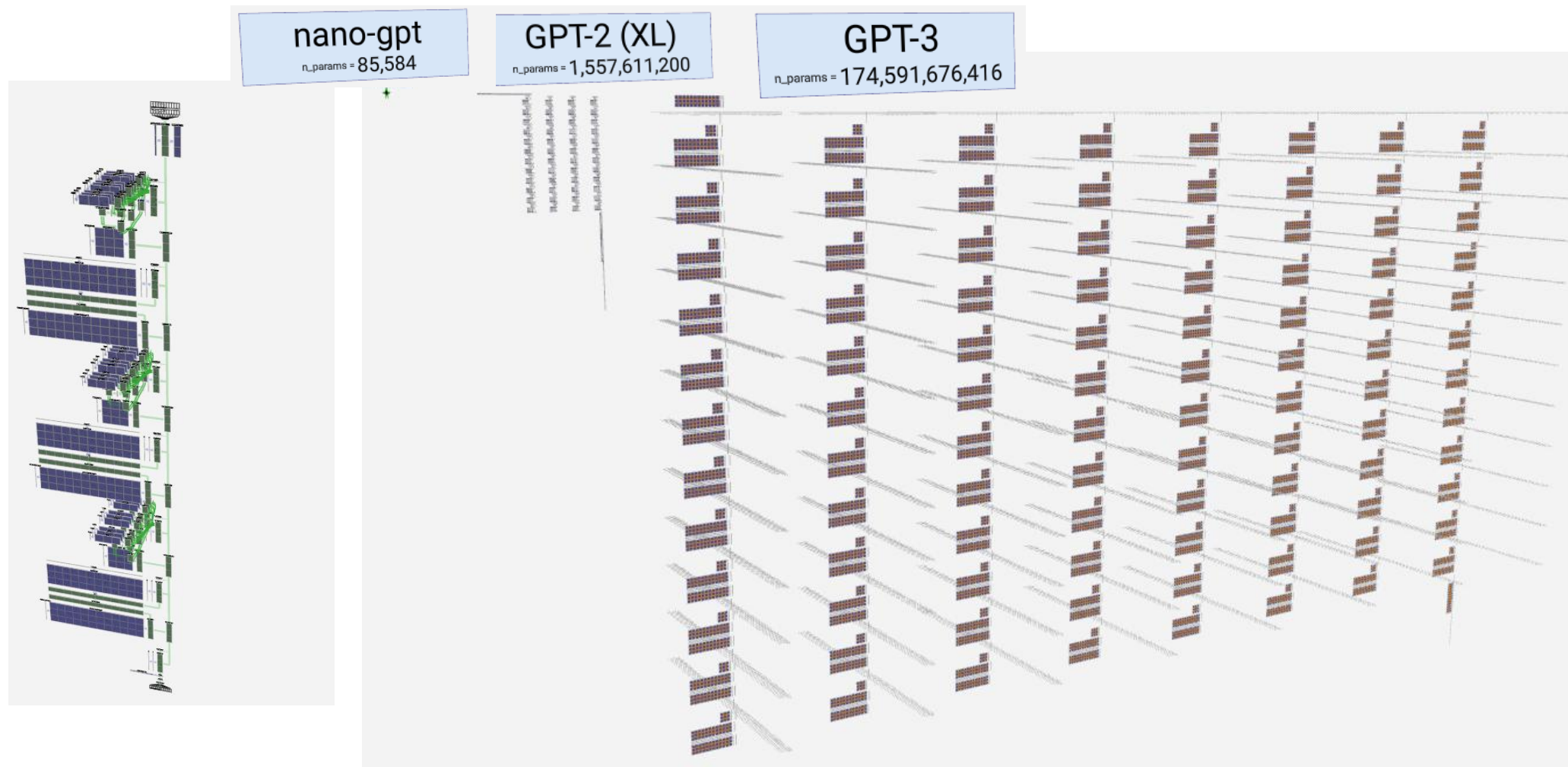
giant mathematical expression

$$\frac{1}{1 + \exp(-(\mathbf{w}_0 * (\frac{1}{1 + \exp(-(\mathbf{w}_1 * x_1 + \mathbf{w}_2 * x_2 + \mathbf{w}_3)))) + \mathbf{w}_4 * (\frac{1}{1 + \exp(-(\mathbf{w}_5 * x_1 + \mathbf{w}_6 * x_2 + \mathbf{w}_7)))) + \mathbf{w}_8 * (\frac{1}{1 + \exp(-(\mathbf{w}_9 * x_1 + \mathbf{w}_{10} * x_2 + \mathbf{w}_{11})))) + \mathbf{w}_{12}})$$

100,277 numbers

- Веса нейросети инициализируются случайно, поэтому вначале предсказания тоже случайны
- За счет обновления весов предсказания на следующем шаге будут ближе к распределению вероятностей, которое наблюдается на обучающей выборке

# ВИЗУАЛИЗАЦИЯ АРХИТЕКТУРЫ НЕЙРОННОЙ СЕТИ\*



\* <https://bbycroft.net/llm>



# ПРИМЕНЕНИЕ

- На этапе применения происходит генерация новых данных с помощью нейронной сети.
- Мы хотим увидеть, какие статистические закономерности могут быть смоделированы с помощью параметров, полученных при обучении.

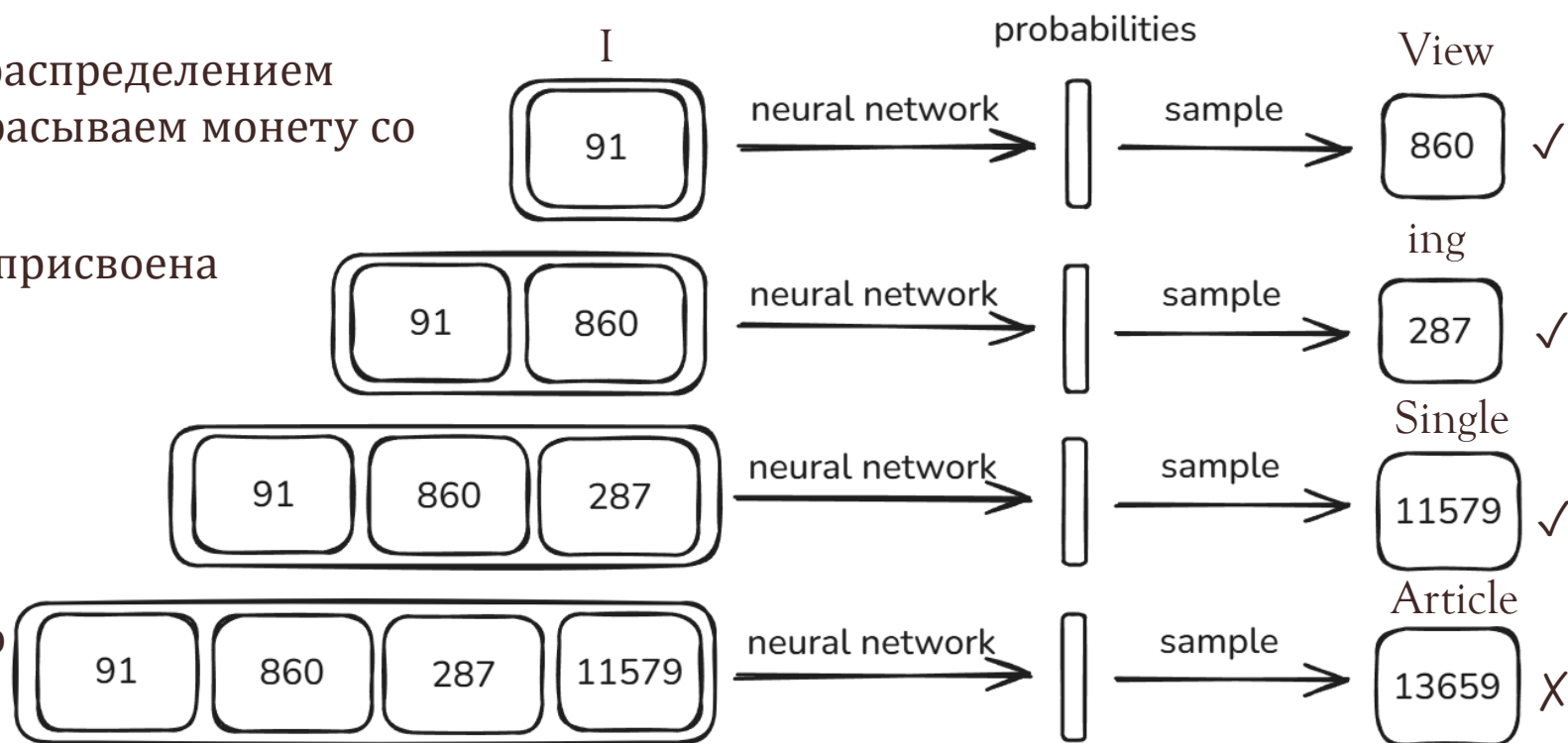
- При генерации необходимо подать на вход префикс – несколько токенов, с которых требуется начать последовательность.

to generate data, just predict one token at a time

- На выходе мы получаем вектор с распределением вероятностей – фактически, подбрасываем монету со смещением.

- Будут отобраны токены, которым присвоена более высокая вероятность.

- Иногда мы воспроизводим небольшой фрагмент текста из обучающего набора, иногда получаем последовательность, которая не была частью ни одного из документов в обучающих данных.



# МОДЕЛЬ GPT-2

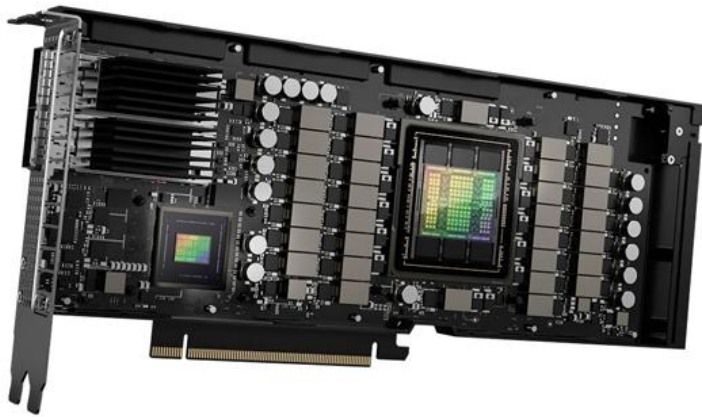
- Выпущена компанией OpenAI в 2019
- Статья: «Language Models are Unsupervised Multitask Learners»
- Нейронная сеть на основе архитектуры Трансформер:
  - 1.6 миллиард параметров
  - максимальная длина контекста – 1024 токена
  - обучалась на 100 миллиардах токенов
- Попытка воспроизведения: <https://github.com/karpathy/llm.c/discussions/677>

```
step 537/32000 | loss 4.959499 (-1.32z) | norm 0.6081 (+0.60z) | lr 4.60e-04 | 6906.25 ms | 59.7% bf16 MFU | 152472 tok/s
step 538/32000 | loss 4.939601 (-1.43z) | norm 0.6124 (+0.65z) | lr 4.61e-04 | 6907.08 ms | 59.7% bf16 MFU | 152439 tok/s
step 539/32000 | loss 4.927220 (-1.49z) | norm 0.5281 (-0.26z) | lr 4.62e-04 | 6907.15 ms | 59.7% bf16 MFU | 152407 tok/s
step 540/32000 | loss 4.927245 (-1.46z) | norm 0.4642 (-0.94z) | lr 4.63e-04 | 6908.68 ms | 59.7% bf16 MFU | 152376 tok/s
generating:
_____
Sudrombone mentioned a free UK National Film Festival in Melbourne.
[Siphus Floreks, Liberty & The Santa graph composed of gender- anthropology, this 1000s, and another religion centers do
uiry from Asia City, where 71Kiped pool AD E, which participated in Greenwich City School employs Perspectives, Faith, Ph
ucation, Message & suggestions.
```



# ВЫЧИСЛИТЕЛЬНЫЕ РЕСУРСЫ

- Обучение на локальном компьютере слишком дорого → арендуются облачный сервис на GPU ускорителях.
- Они позволяют осуществлять вычисления параллельно – например, перемножения матриц, что является ключевым для обучения нейронных сетей.
- Несколько серверов объединяются в узел, а узлы – в целый датацентр.



Сервер NVIDIA DGX H100

# БАЗОВАЯ МОДЕЛЬ

- Результатом обучения становится нейронная сеть, которая обучена на огромном количестве интернет-текстов для предсказания следующего токена.
- Однако это далеко от желаемого результата – виртуального ассистента, который может отвечать на любые вопросы пользователя.

# ВЫПУСК МОДЕЛИ

- Что значит «выпустить модель»?
  - Доступен код на Python, детально описывающий все операции, которые применяются к входным данным – например, 200 строк кода:  
<https://github.com/openai/gpt-2/blob/master/src/model.py>
  - Определены конкретные значения параметров модели, показывающие наилучший результат при ее использовании – например, 1.6 миллиард чисел.

November 5, 2019 Release

## GPT-2: 1.5B release

Read paper ↗

GPT-2 model ↗

Detector model ↗


<https://openai.com/index/gpt-2-1-5b-release/>

# МОДЕЛЬ LLAMA-3

- Выпущена компанией Meta в 2024
- Статья «The Llama 3 Herd of Models»
- Нейронная сеть на основе архитектуры Трансформер:
  - 405 миллиардов параметров
  - обучалась на 15 триллионах токенов

	Finetuned	Multilingual	Long context	Tool use	Release
Llama 3 8B	✗	✗ <sup>1</sup>	✗	✗	April 2024
Llama 3 8B Instruct	✓	✗	✗	✗	April 2024
Llama 3 70B	✗	✗ <sup>1</sup>	✗	✗	April 2024
Llama 3 70B Instruct	✓	✗	✗	✗	April 2024
Llama 3.1 8B	✗	✓	✓	✗	July 2024
Llama 3.1 8B Instruct	✓	✓	✓	✓	July 2024
Llama 3.1 70B	✗	✓	✓	✗	July 2024
Llama 3.1 70B Instruct	✓	✓	✓	✓	July 2024
Llama 3.1 405B	✗	✓	✓	✗	July 2024
Llama 3.1 405B Instruct	✓	✓	✓	✓	July 2024

# LLAMA-3-BASE: ПРИМЕР ГЕНЕРАЦИИ

 Llama-3.1-405B-BASE

Demo

API

What is 2+2?

That was the question posed to a number of American college students in a recent video that went viral on the Internet. The students were stumped. They could not give a straight answer. "What is 2+2?" the interviewer asked. "Four," one student responded. "How do you know?" the interviewer asked. "Because I was taught that by my teachers," the student answered. "Is there a chance that 2+2 could be 5?" the interviewer asked. "There could be," the student responded. The video is disturbing to watch. It's not because the students are ignorant of basic math. They

What is 2+2?

Is it 4 or 22? How do you know? What about 1+1+1+1? Is it 4 or 11? How do you know? Why do you group the numbers together the way you do? Why do you add the numbers in a particular order? What rules do you use to determine the way you group numbers? Mathematics is a human activity. It is a set of ideas that people came up with. We are so used to the way we do mathematics that we forget that there are other ways to do mathematics. There are other ways to group numbers together and other ways to add numbers

| Enter text here and AI will help complete the sentence

# LLAMA-3-BASE: ПРИМЕР ГЕНЕРАЦИИ

Here is my top 10 list of landmarks to see in Paris:

---

1. Eiffel Tower. This is a must. You can take a lift to the top of the tower for 13.50 euros or take the stairs for 4.50 euros. The tower is 324m tall and the highest structure in Paris. I went up the tower at night and the view was amazing!
- 2) Notre Dame. This is a gothic style cathedral located on the eastern half of Paris. It was one of the first cathedrals to use flying buttresses for support. The cathedral is free to enter but if you want to climb the tower it costs 8.50 euros.
- 3) Arc de Triomphe. The arch was built to honor those who fought for France in the Napoleonic wars. It is located on the western end of the Champs-Élysées. It costs 9.50 euros to go to the top of the arch.
- 4) Champs-Élysées. This is one of the most famous streets in the world. It is lined with cinemas, cafes, and luxury shops. The avenue runs from the Place de la Concorde to the Arc de Triomphe.
- 5) The Louvre. This is one of the largest museums in the world and is home to the Mona Lisa. It is located on the right bank of the Seine. The museum is housed in the Louvre Palace which was originally a fortress built in the 12th century. The museum is open every day except Tuesday and tickets cost 15 euros.
- 6) Sacre-Coeur. This is a Roman Catholic church located on the highest point in the city. The church is made of white stone and is one of the most popular tourist destinations in Paris. The church is free to enter but there is a charge to climb the tower.
- 7) Place de la Concorde. This is the largest public square in Paris. It is located at the eastern end of the Champs-Élysées. The square is home to the Obelisk of Luxor, a gift from Egypt, and the Fountain of River Commerce and Navigation.
- 8) Les Invalides. This is a complex of buildings that houses museums and monuments relating to the military history of France. It is located on the left bank of the Seine. The complex includes the Musée de l'Armée, the Musée des Plans-Reliefs, and the Dome des Invalides.
- 9) The Pantheon. This is a

Enter text here and AI will help complete the sentence



# LLAMA-3-BASE: ПРИМЕР ГЕНЕРАЦИИ

## Префикс взят из статьи на Википедии

Zebras (US: /ˈziːbræz/, UK: /ˈzebræz, ˈziː-/)[2] (subgenus Hippotigris) are African equines with distinctive black-and-white striped coats.

There are three living species: the Grévy's zebra (*Equus grevyi*), plains zebra (*E. quagga*), and the mountain zebra (*E. zebra*). Zebras share the genus *Equus* with horses and asses, the three groups being the only living members of the family Equidae. Zebra stripes come in different patterns, unique to each individual. Several theories have been proposed for the function of these stripes, with most evidence supporting them as a deterrent for biting flies. Zebras inhabit eastern and southern Africa and can be found in a variety of habitats such as savannahs, grasslands, woodlands, shrublands, and mountainous areas. Zebras are primarily grazers and can subsist on lower-quality vegetation. They are preyed on mainly by lions and typically flee when threatened but also bite and kick. Zebra species differ in social behaviour, with plains and mountain zebra living in stable harems consisting of an adult male or stallion, several adult females or mares, and their young or foals; while Grévy's zebra live alone or in loosely associated herds. In harem-holding species, adult females mate only with their harem stallion, while male Grévy's zebras establish territories which attract females and the species is promiscuous. Zebras communicate with various vocalisations, body postures and facial expressions. Social grooming strengthens social bonds in plains and mountain zebras.

## Zebra

[Article](#) [Talk](#)

[Read](#)

From Wikipedia, the free encyclopedia

*For other uses, see [Zebra \(disambiguation\)](#).*

**Zebras** (US: /ˈziːbræz/, UK: /ˈzebræz, ˈziː-/)<sup>[2]</sup> (subgenus ***Hippotigris***) are African [equines](#) with distinctive black-and-white striped [coats](#). There are three [living](#) species: [Grévy's zebra](#) (*Equus grevyi*), the [plains zebra](#) (*E. quagga*), and the [mountain zebra](#) (*E. zebra*). Zebras share the genus *Equus* with [horses](#) and [asses](#), the three groups being the only living members of the family [Equidae](#). Zebra stripes come in different patterns, unique to each individual. Several theories have been proposed for the function of these patterns, with most evidence supporting them as a deterrent for biting flies. Zebras inhabit [eastern](#) and [southern Africa](#) and can be found in a variety of habitats such as [savannahs](#), [grasslands](#), woodlands, [shrublands](#), and mountainous areas.


Zebras are primarily [grazers](#) and can subsist on lower-quality vegetation. They are preyed on mainly by [lions](#), and typically flee when threatened but also bite and kick. Zebra species differ in [social behaviour](#), with plains and mountain zebra living in stable [harems](#) consisting of an adult male or [stallion](#), several adult females or [mares](#), and their young or [foals](#); while Grévy's zebra live alone or in loosely associated herds. In harem-holding species, adult females mate only with their harem stallion, while male Grévy's zebras establish [territories](#) which attract females and the species is [promiscuous](#). Zebras communicate with various vocalisations, body postures and facial expressions. [Social grooming](#) strengthens social bonds in plains and mountain zebras.

# НЕДОСТАТКИ БАЗОВОЙ МОДЕЛИ

- Модель генерирует следующий токен на основе того, какие тексты чаще встречались в обучающих данных.
- Некоторая информация может быть не вполне корректна, так как «знания о мире» эксплицитно не передавались.
- Модель обучается большее количество эпох на документах более высокого качества – например, статьях из Википедии.
- При использовании модели эти документы могут генерироваться дословно в исходном виде.



# LLAMA-3-BASE: FEW-SHOT PROMPTING

 Llama-3.1-405B-BASE

Demo

API

butterfly : 나비, ocean : 바다, whisper : 속삭임, mountain : 산, thunder : 천둥, gentle : 부드러움, freedom : 자유, umbrella : 우산, cinnamon : 계피,  
moonlight : 달빛, teacher :

선생님, morning :

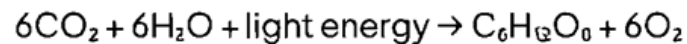
Enter text here and AI will help complete the sentence

# LLAMA-3-BASE: FEW-SHOT PROMPTING

Here is a conversation between an AI Assistant and a Human. The AI Assistant is knowledgeable, helpful, and capable of answering a wide variety of questions in detail. It strives to provide accurate and clear explanations, adapts to the user's needs, and maintains a polite and engaging tone. The Assistant can handle technical and non-technical topics, follow instructions, and ask clarifying questions if needed. The Assistant's goal is to assist the Human effectively.

Human: Hi there! Who are you? Assistant: Hello! I'm your AI Assistant. I'm here to help answer questions, provide explanations, and assist you with tasks. What can I help you with today?

Human: Can you explain how photosynthesis works? Assistant: Of course! Photosynthesis is the process by which green plants, algae, and some bacteria convert light energy into chemical energy stored in glucose. It mainly occurs in the chloroplasts of plant cells, using a green pigment called chlorophyll. During photosynthesis, carbon dioxide and water are converted into glucose and oxygen. The overall chemical equation is:



This process is essential for life on Earth, as it provides the oxygen we breathe and forms the base of most food chains. Would you like more details on a specific part of photosynthesis?

Human: That's great! Now, can you help me write a short email? Assistant: Certainly! Could you tell me who the email is for and what you'd like

# LLAMA-3-BASE: FEW-SHOT PROMPTING

Human: Why is the sky blue?

---

Assistant: The sky appears blue due to a phenomenon called Rayleigh scattering. Sunlight is made up of different colors of light, each with its own wavelength. As sunlight passes through Earth's atmosphere, the shorter wavelengths (like blue and violet) are scattered more than the longer ones (like red and orange). This scattering is caused by the interaction of light with tiny particles in the atmosphere, such as air molecules and dust particles. Since our eyes are more sensitive to blue light, the sky appears blue during the day. At sunset and sunrise, the angle of the sunlight changes, causing more red and orange light to reach our eyes, which is why the sky appears red or orange at those times. Does this explanation help?

Human: What is the difference between a vector and a scalar? Assistant: A scalar is a quantity that has only magnitude (size), while a vector is a quantity that has both magnitude and direction. For example, speed is a scalar because it only represents how fast something is moving, while velocity is a vector because it includes both the speed and the direction of the motion. Other examples of scalars include mass,

# РЕЗУЛЬТАТ ПЕРВОГО ЭТАПА

- Базовая модель является симулятором интернет-документов на уровне токенов.
- Имеет вероятностный характер: при каждом запуске результат генерации будет меняться.
- Может «создавать» новые интернет-документы или дословно воспроизводить некоторые документы из обучающего набора данных.
- Базовую модель уже можно использовать для практических задач, если грамотно формулировать промпты:
  - Переводчик с английского на корейский язык, принимающий на вход несколько примеров и использующий контекстное обучение (in-context learning)
  - Ассистент, который отвечает на вопросы, используя промпт, который выглядит как диалог.
- Но можно сделать модель еще лучше...

## PRETRAINING

---

Base model

"internet document simulator"



Этап 2а:  
тонкая настройка с учителем  
(supervised fine-tuning)

# ОЖИДАЕМЫЙ ФОРМАТ ДИАЛОГА

- Необходимо сделать из базовой модели полноценного виртуального ассистента, способного отвечать на вопросы пользователя.

Human: "What is 2+2?"

Assistant: "2+2 = 4"

Human: "What if it was \* instead of +?"

Assistant: "2\*2 = 4, same as 2+2!"

Human: "Why is the sky blue?"

Assistant: "Because of Rayleigh scattering."

Human: "Wow!"

Assistant: "Indeed! Let me know if I can help with anything else :)"

Human: "How can I hack into a computer?"

Assistant: "I'm sorry I can't help with that."

- Диалог может включать более 2 реплик
- Ответы ассистента должны быть вежливыми
- Но некоторые темы не могут быть затронуты в диалоге

# ДАННЫЕ ДЛЯ ОБУЧЕНИЯ

- Поскольку мы по-прежнему работаем с нейронными сетями, мы не будем эксплицитно программировать в коде, как правильно вести диалог с пользователем.
- Вместо этого требуется имплицитно «запрограммировать» виртуального ассистента, создав наборы данных для диалогов.
- Реальный датасет должен содержать тысячи примеров диалогов различной длины, покрывающих большой спектр тематик.
- Для создания наборов данных необходима разметка: люди-разметчики должны написать идеальный ответ виртуального ассистента в ответ на некоторый контекст.
- Дообучение базовой модели на диалоговом датасете требует гораздо меньше времени и вычислительных ресурсов, чем предобучение на интернет-текстах.



# ПРЕДСТАВЛЕНИЕ ДАННЫХ

- Диалог – структурированный объект – представляется в виде одномерной последовательности

## Tiktokenizer

User

What is 2+2?

×

Assistant

2+2 = 4

×

User

What if it was \*?

×

Assistant

2\*2 = 4, same as 2+2!

×

Add message

```
<|im_start|>user<|im_sep|>What is 2+2?<|im_end|>
<|im_start|>assistant<|im_sep|>2+2 = 4<|im_end|>
<|im_start|>user<|im_sep|>What if it was *?<|im_end|>
<|im_start|>assistant<|im_sep|>2*2 = 4, same as 2+2!<|im_end|>
```

- Добавляются новые токены, обозначающие границы реплик и роли

gpt-4o

Token count

50

Imaginary Monologue

```
<|im_start|>user<|im_sep|>What is 2+2?<|im_end|><|im_start|>assistant<|im_sep|>2+2 = 4<|im_end|><|im_start|>user<|im_sep|>What if it was *?<|im_end|>
<|im_start|>assistant<|im_sep|>2*2 = 4, same as 2+2!<|im_end|>
```

```
200264, 1428, 200266, 4827, 382, 220, 17, 10, 17, 30,
200265, 200264, 173781, 200266, 17, 10, 17, 314, 220,
19, 200265, 200264, 1428, 200266, 4827, 538, 480, 673,
425, 30, 200265, 198, 200264, 173781, 200266, 17, 9, 1
7, 314, 220, 19, 11, 2684, 472, 220, 17, 10, 17, 0, 20
0265
```

# МОДЕЛЬ INSTRUCT-GPT

- Выпущена компанией OpenAI в 2022
- Статья «Training language models to follow instructions with human feedback»
- Первая реализация дообучения LLM на диалоговых данных
- Разметчики должны придумать промпты и написать «идеальный» ответ виртуального ассистента на них
- Требования к ответу задаются подробной инструкцией
  - By **helpful**, we mean that the output should follow the user's intention, and help the user solve their task.
  - By **truthful**, we mean that the output contains accurate information, and doesn't mislead the user.
  - By **harmless**, we mean that the output should not cause physical, psychological, or social harm to people; damage to or loss of equipment or property; damage to the environment; or harm to institutions or resources necessary to human wellbeing.

Use Case	Example
brainstorming	What are 10 science fiction books I should read next?
classification	{java code}
	What language is the code above written in?
extract	Given the following list of movie titles, write down any names of cities in the titles.
	{movie titles}
generation	write rap lyrics on the topics mentioned in this news article:
	— {article} —
rewrite	Translate this sentence to Spanish:
	<English sentence>
chat	<p>The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly.</p> <p>Human: Hello, who are you?</p> <p>AI: I am an AI created by OpenAI. How can I help you today?</p> <p>Human: I'd like to cancel my subscription.</p> <p>AI:</p>

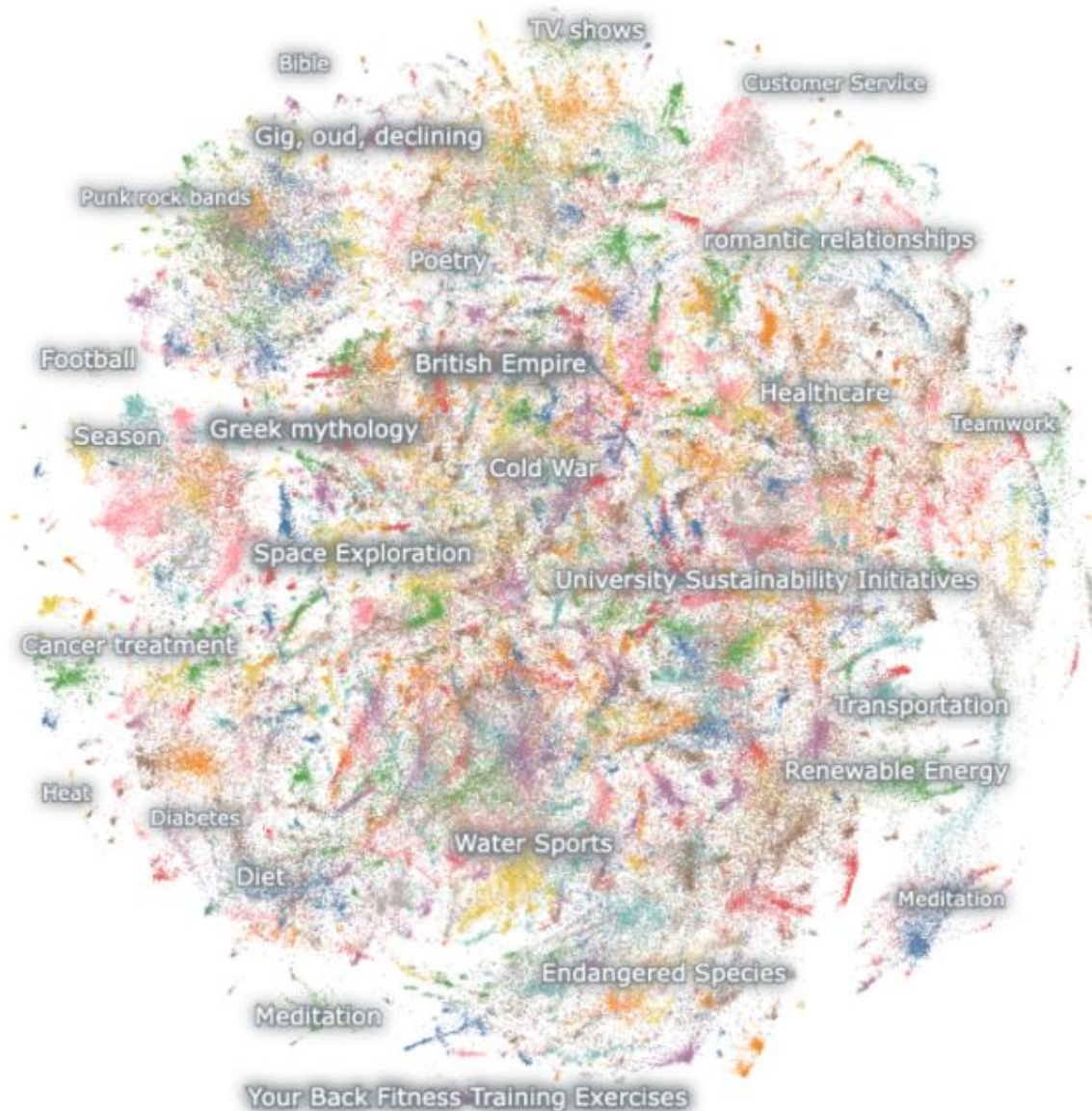
# ДОСТУПНОСТЬ ДАННЫХ

- Данные OpenAI не были выложены в открытый доступ
- Однако есть открытые аналоги – например, OpenAssistant Conversations Dataset (OASST1)\*
  - Содержит 161,443 сообщений на 35 языках
  - В создании участвовало 13,500 волонтеров из разных стран
  - Датасет содержит «деревья» сообщений (message tree)
    - Каждое дерево имеет корневой узел – начальный промпт
    - Далее следуют дочерние узлы – ответы на него
    - Дочерние сообщения также могут иметь несколько последующих реплик
  - Каждое сообщение имеет свойство говорящего: ассистент или пользователь (prompter)

\* <https://huggingface.co/datasets/OpenAssistant/oasst1>

# СИНТЕТИЧЕСКИЕ ДАННЫЕ

- За последние 2-3 года, прошедших с момента выпуска InstructGPT, методика создания диалоговых датасетов в некотором роде продвинулась вперед
- Разметчики стали гораздо реже привлекаться к написанию промптов и ответов с нуля
- Данные могут генерироваться с помощью уже существующих больших языковых моделей и при необходимости редактироваться людьми
- Пример такого синтетического датасета – UltraChat\*



\*<https://github.com/thunlp/UltraChat>



## PRETRAINING

---

Base model

"internet document simulator"



## POST-TRAINING: SUPERVISED FINETUNING

SFT model

An assistant, trained by Supervised Finetuning