

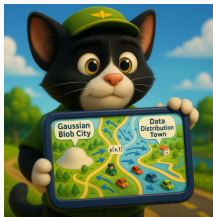
# FLOW MATCHING И ИДЕЯ LLM

Сергей Николенко

СПбГУ — Санкт-Петербург

04 декабря 2025 г.

*Random facts:*



- 4 декабря в России — День российской информатики, потому что 4 декабря 1948 г. в СССР за номером 10475 было зарегистрировано изобретение И.С. Брука и Б.И. Рамеева — цифровая электронная вычислительная машина
- 4 декабря 771 г. король франков Карломан умер в своём дворце Самусси, и Карл Великий стал единовластным правителем Франкского королевства
- 4 декабря 1783 г., через 9 дней после отплытия последнего корабля англичан, Джордж Вашингтон на ужине в таверне Френсиса (Fraunces Tavern) попрощался со своими офицерами, подал в отставку и удалился в своё поместье
- 4 декабря 1934 г. Леонид Николаев выстрелом из револьвера убил в Смольном Сергея Кирова; как это убийство было задумано и организовано, не ясно до сих пор
- 4 декабря 1956 г. на джем-сессию в Sun Studio в первый и последний раз собрался Million Dollar Quartet: Элвис Пресли, Джерри Ли Льюис, Карл Перкинс и Джонни Кэш
- 4 декабря 1971 г. во время концерта Фрэнка Заппы дотла сгорело многоэтажное казино в Монтрё; это событие увековечено в песне *Deep Purple* «Smoke on the Water»

# FLOW MATCHING

---

- Residual flows: если поток рассмотреть в виде

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \epsilon g_\theta(\mathbf{x}_k, t_k),$$

то он будет обратим при не слишком сложных условиях на  $g_\theta$

- А это теперь выглядит как шаг дискретизации ОДУ:

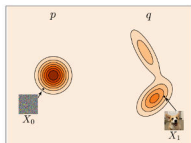
$$\frac{d\mathbf{x}_t}{dt} = v_\theta(\mathbf{x}_t, t)$$

- Это continuous normalizing flows (CNF), и для них

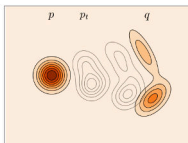
$$\log p_1(\mathbf{x}_1) = \log p_0(\mathbf{x}_0) - \int_0^1 \operatorname{div} \frac{\partial v_\theta}{\partial \mathbf{x}_t}$$

# FLOW MATCHING

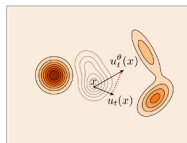
- Flow Matching (FM; Lipman et al., 2024): новый подход, объединяющий потоки и диффузионные модели
- Глобальная задача: построить модель, способную генерировать новые образцы из целевого распределения  $q$  по обучающей выборке
- Метод flow matching: построить вероятностный путь  $p_t$  от известного распределения источника  $p_0 = p$  к целевому распределению  $p_1 = q$ .



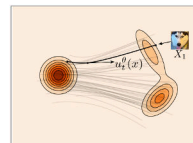
**(a)** Data.



**(b)** Path design.



**(c)** Training.



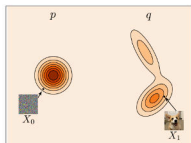
**(d)** Sampling.

- А для этого обучить поле скоростей  $u_t$  для порождения  $p_t$ :

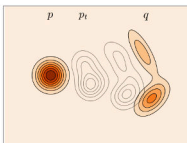
$$L_{\text{FM}}(\theta) = \mathbb{E}_{t, X_t} \|u_t^\theta(X_t) - u_t(X_t)\|^2,$$

где  $u_t^\theta$  — параметризированное поле скоростей.

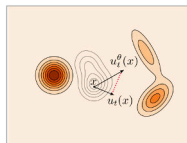
- После обучения, чтобы использовать FM-модель, можно:
  - сэмплировать  $X_0 \sim p$ ;
  - решить ОДУ для получения  $X_1 \sim q$ .



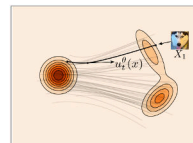
(a) Data.



(b) Path design.



(c) Training.



(d) Sampling.

- Потоки теперь становятся непрерывными
- Это детерминированные отображения  $\psi_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$ , которые преобразуют распределение источника  $p$  в целевое распределение  $q$
- Определяются через ОДУ:

$$\frac{d}{dt}\psi_t(x) = u_t(\psi_t(x)), \quad \psi_0(x) = x,$$

где  $u_t(x)$  — поле скоростей

- Связь между потоками и плотностью вероятности:

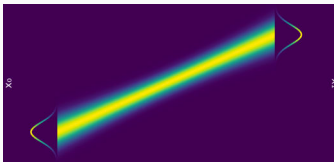
$$\log p_t(\psi_t(x)) = \log p_0(x) - \int_0^t \operatorname{div} u_s(\psi_s(x)) ds,$$

где  $\operatorname{div} u_s$  — дивергенция вектора  $u_s$

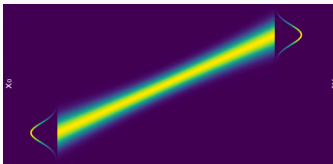
- Непрерывные нормализующие потоки (CNF) используют биективную, непрерывно дифференцируемую функцию  $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^d$ , чтобы преобразовать простое распределение  $q_0(x)$  в сложное целевое распределение  $p_1(y)$ .
- Плотность  $p_1$  рассчитывается с помощью формулы замены переменных:

$$p_1(y) = q_0(\phi^{-1}(y)) \left| \det \left( \frac{\partial \phi^{-1}(y)}{\partial y} \right) \right|$$

- Обучение заключается в максимизации логарифма правдоподобия данных в модели



- Но у CNF есть проблемы:
  - обучение CNF методом максимального правдоподобия требует решения ОДУ, что может быть вычислительно сложным и численно нестабильным
  - да и просто в обучении эти модели будут ну очень медленными, там надо всё время решать ОДУ



- FM предлагает метод обучения CNFs без симуляции, сопоставляя обучаемое векторное поле с целевым векторным полем, которое определяет вероятностный путь от  $q_0$  к  $q_1$ .
- Минимизируем функцию потерь

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t, x \sim p_t} \|v_{\theta}(t, x) - u_t(x)\|^2$$

где  $v_{\theta}$  — обучаемое векторное поле, а  $u_t$  — целевое векторное поле.

- Иначе говоря, FM выходит за рамки диффузионных путей, позволяя использовать пути оптимального транспорта, что может привести к более эффективному обучению и генерации.

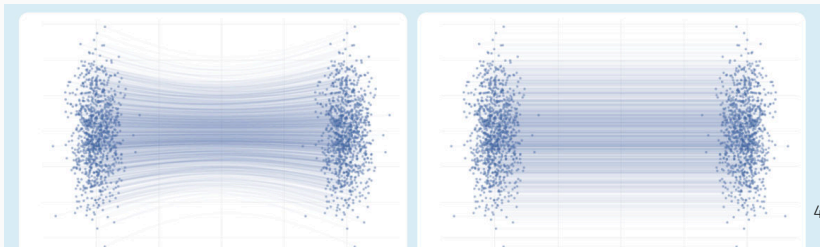
# FLOW MATCHING

- Цель FM: Найти  $u_t$ , который генерирует вероятностный путь  $p_t$  с граничными условиями  $p_0 = p, p_1 = q$ .
- Построение пути:
  - $p_t$  определяется как сумма по условным вероятностям:

$$p_t(x) = \int p_{t|1}(x|x_1)q(x_1)dx_1.$$

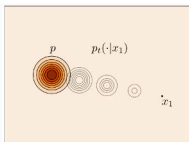
- Условный путь  $p_{t|1}$  может быть выбран как:

$$p_{t|1}(x|x_1) = \mathcal{N}(x|tx_1, (1-t)^2I).$$

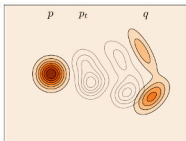


# FLOW MATCHING

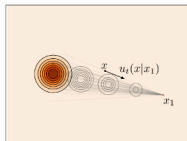
- Условное поле скоростей:  $u_t(x|x_1) = \frac{x_1 - x}{1-t}$
- Общее поле скоростей:  $u_t(x) = \int u_t(x|x_1)p_{1|t}(x_1|x)dx_1$



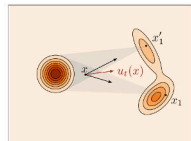
**(a)** Conditional probability path  $p_t(x|x_1)$ .



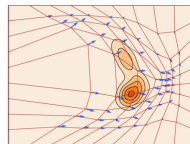
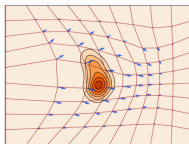
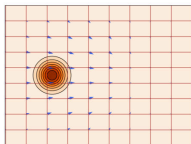
**(b)** (Marginal) Probability path  $p_t(x)$ .



**(c)** Conditional velocity field  $u_t(x|x_1)$ .



**(d)** (Marginal) Velocity field  $u_t(x)$ .

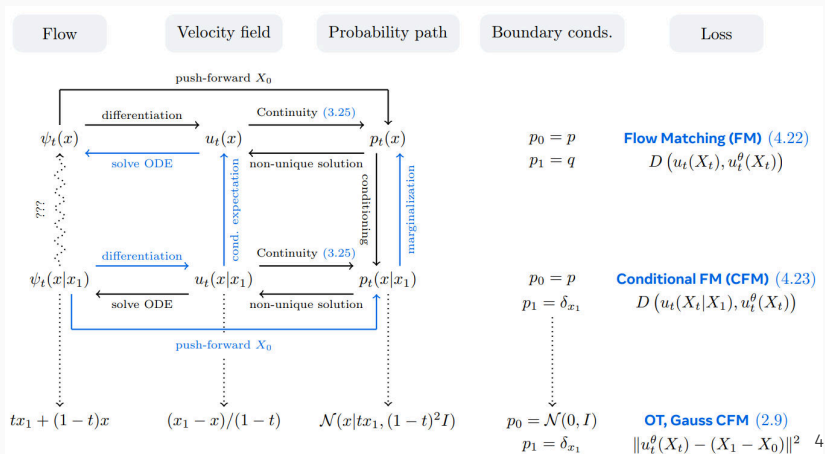


**Figure 7** A velocity field  $u_t$  (in blue) generates a probability path  $p_t$  (PDFs shown as contours) if the flow defined by  $u_t$  (square grid) reshapes  $p$  (left) to  $p_t$  at all times  $t \in [0, 1)$ .

# FLOW MATCHING

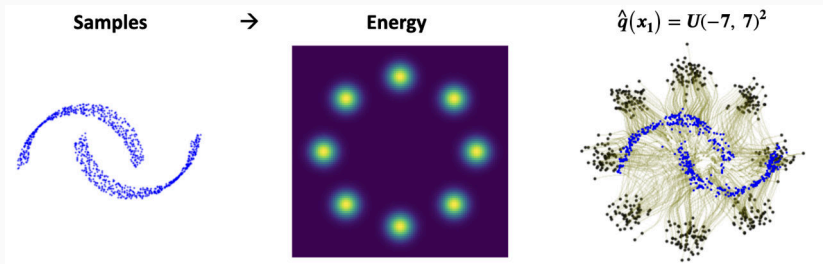
- Условная версия функции ошибки:

$$L_{\text{CFM}}(\theta) = \mathbb{E}_{t, X_t, X_1} \|u_t^\theta(X_t) - u_t(X_t|X_1)\|^2.$$



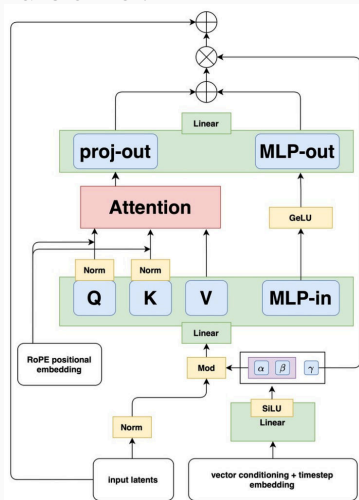
# FLOW MATCHING

- MF устраняет необходимость симуляции ОДУ во время обучения, снижая вычислительную сложность
- Поддерживает гибкость в выборе вероятностных путей, включая пути оптимального транспорта.
- Но кажется, что всё это пока только начинается



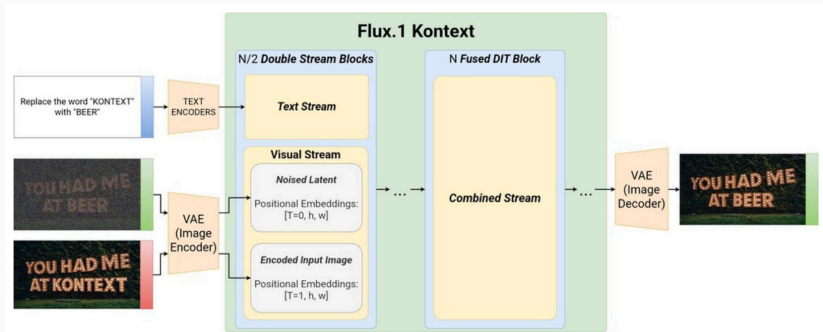
# FLUX.1 KONTEXT

- Пример применения – Flux.1 Kontext (Black Forest Labs, 2025)
- Блоки Diffusion Transformer:



# FLUX.1 KONTEXT

- Объединяются в сеть с flow matching objective:



# FLUX.1 KONTEXT

- И получается крутой image editing:



(a) **Context image** generated with FLUX.1.



(b) Image context from Figure 1a: *“The bird is now sitting in a bar and enjoying a beer.”*



(c) Image context from Figure 1b: *“There are now two of these birds.”*

# FLUX.1 KONTEXT

- И получается крутой image editing:



(d) From Figure 1c: *"Watch them from behind."*



(e) From Figure 1c: *"The two bird characters are now sitting in a movie theater."*

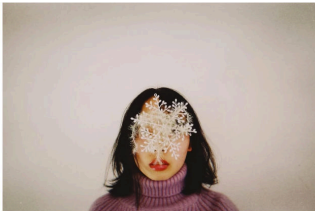


(f) From Figure 1c: *"The two bird characters are now grocery shopping."*

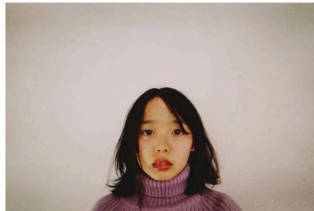


(g) From Figure 1f: *"The two bird characters are now celebrating a successful launch."*

- И получается крутой image editing:



(a) Input image



(b) "remove the thing from her face"



(c) "she is now taking a selfie in the streets of Freiburg, it's a lovely day out."

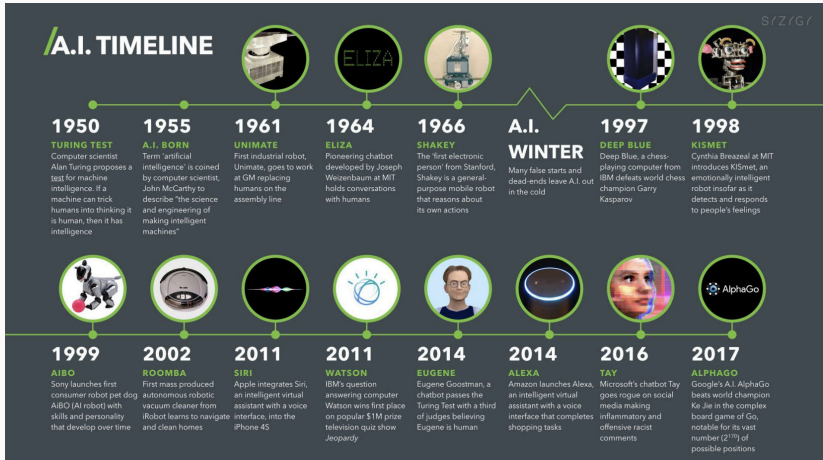


(d) "it's now snowing, everything is covered in snow."

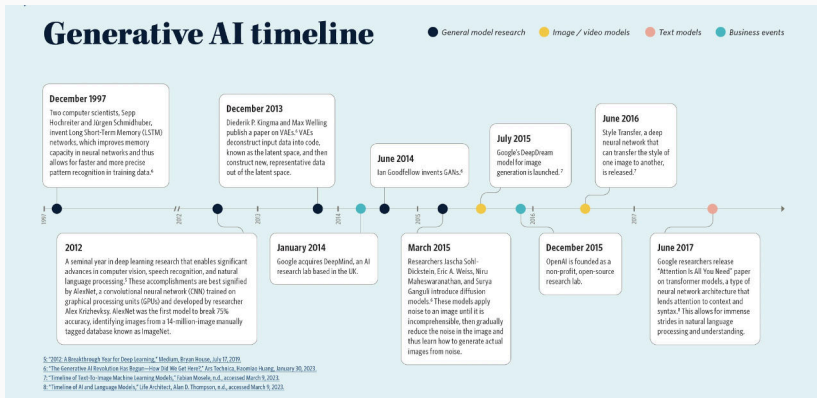
БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ:  
ДОСТИЖЕНИЯ, ЗАКОНЫ  
МАСШТАБИРОВАНИЯ, UNNOVBLING

---

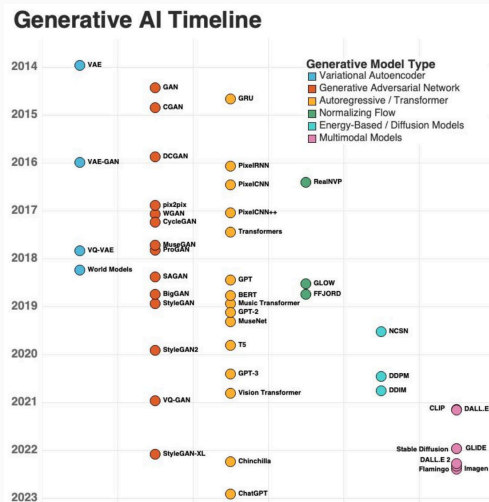
- В интересные времена мы живём...



- В интересные времена мы живём...

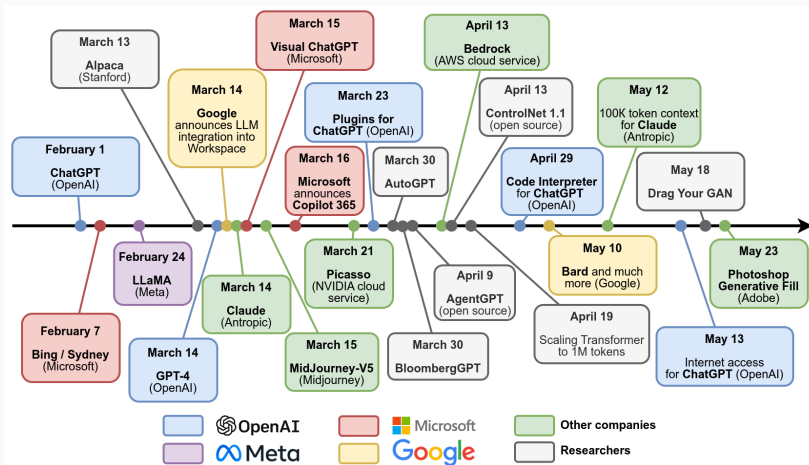


- В интересные времена мы живём...



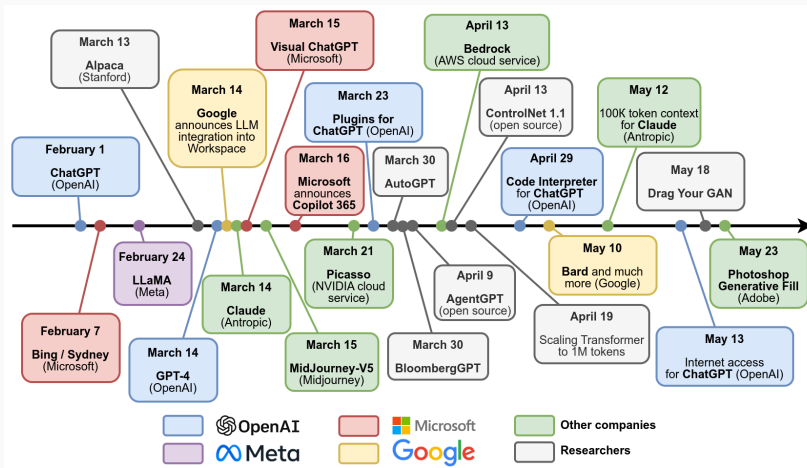
# AI TIMELINES

- В интересные времена мы живём...



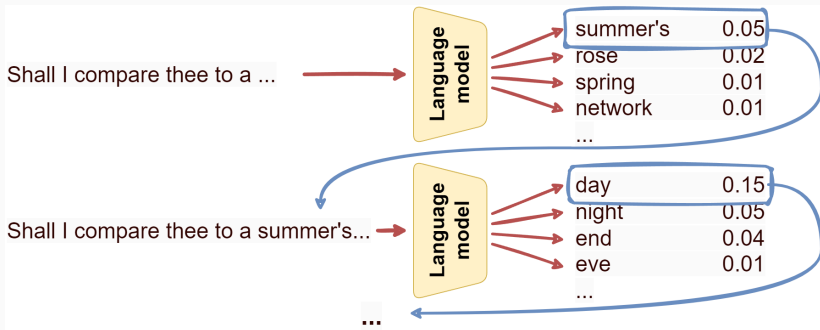
# AI TIMELINES

- В интересные времена мы живём...

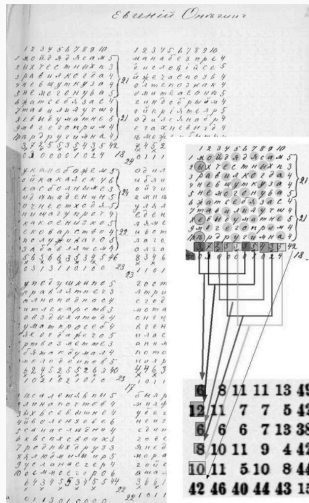


- В основном благодаря LLM

- Языковая модель – это вид постановки задачи:  
авторегрессивно порождать следующий токен на основе уже порождённых



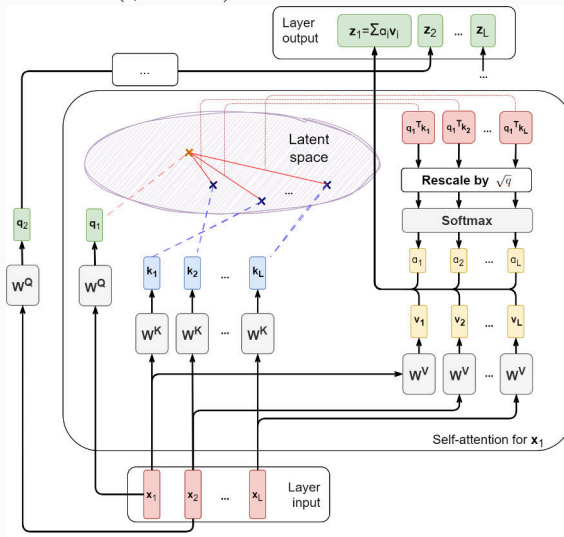
- С языковых моделей когда-то начинались марковские цепи...



- Было много классических подходов:
  - считаем  $n$ -граммы слов;
  - сглаживаем их, fallback на  $n$ -граммы поменьше;
  - например, Kneser-Ney smoothing;
  - вычислительные проблемы и улучшенные алгоритмы ( $n$ -грамм получается много).
- Но это всё давало только, например, помощь при распознавании речи
- А сейчас LLM – это самый вероятный путь к AGI

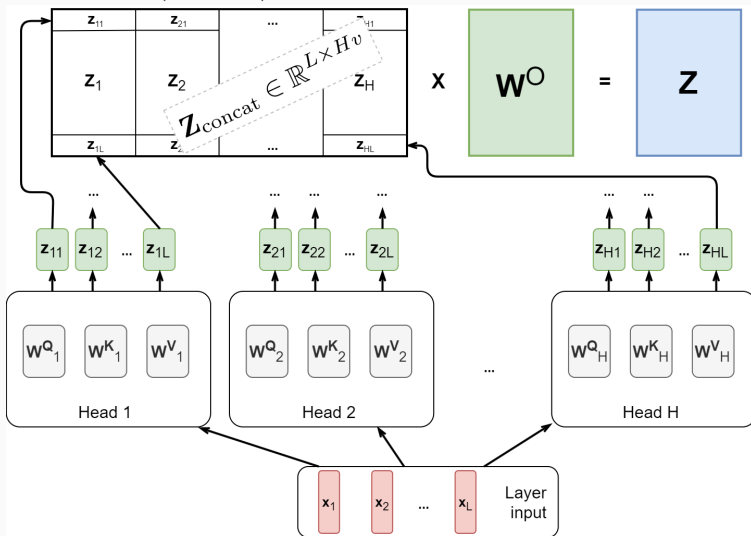
# ПРОБЛЕМЫ ТРАНСФОРМЕРОВ

- Мы уже говорили о трансформерах: self-attention работает как  $\mathbf{z}_i = \text{softmax} \left( \frac{1}{\sqrt{q}} \mathbf{q}_i \mathbf{K}^\top \right) \mathbf{V}$



# ПРОБЛЕМЫ ТРАНСФОРМЕРОВ

- В матричной форме одна "голова" выглядит как  $Z = \text{softmax} \left( \frac{1}{\sqrt{q}} QK^T \right) V$



- И здесь первая проблема:

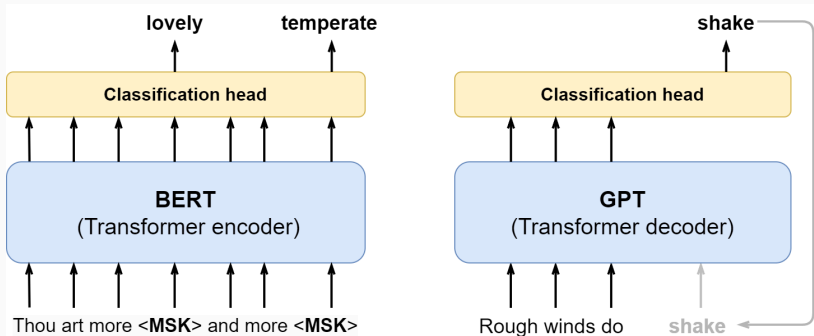
$$Z = \text{softmax} \left( \frac{1}{\sqrt{q}} QK^{\top} \right) V$$

означает вычисление  $L \times L$  матрицы  $QK^{\top}$

- Квадратичная сложность значит, что контекст расширять сложно, а хочется!

# ПРОБЛЕМЫ ТРАНСФОРМЕРОВ

- Вторая проблема – LLM просто предсказывает токены, а как сделать так, чтобы она делала что-то полезное?

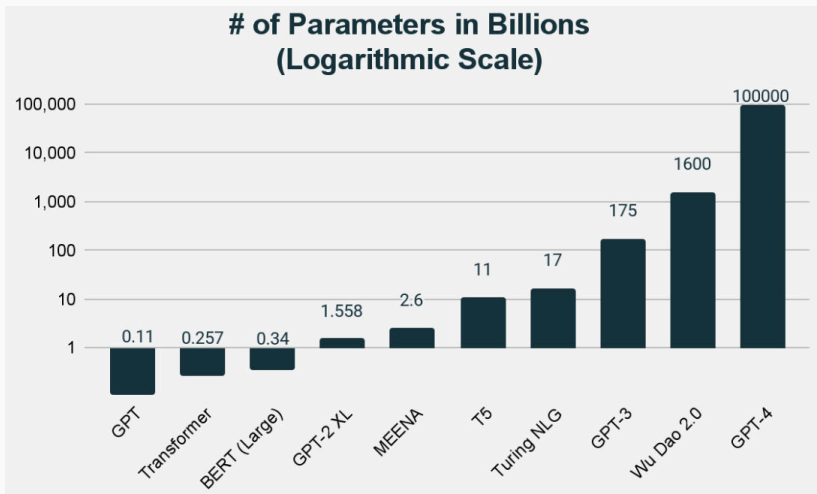


- Третья проблема – как добавить к LLM внешнюю информацию, например, как разрешить делать поиск?
- В основном об этом и будем сегодня говорить, но сначала – о ВОЗМОЖНОСТЯХ...

- История семейства моделей GPT – почти прямое масштабирование:
  - в GPT было 12 слоёв декодера трансформера, в каждом 12 маскированных голов внимание,  $d = 64$ ; GPT предобучался на *BookCorpus*,  $> 7000$  книг (крохотный датасет по нынешним временам!), потом дообучался для конкретных задач на размеченных датасетах;
  - GPT-2 (Radford et al., Feb 2019) предобучался на *BookCorpus* и новом датасете *WebText*: 8М страниц, проверенных людьми (ссылки с *Reddit* с кармой  $\geq 3$ , 40GB текста); у GPT-2 было 48 слоёв и  $d = 1600$ , суммарно около 1.5B параметров, 10x GPT;
  - GPT-3 (Brown et al., Jun 2020) добавила ещё два порядка; у семейства *davinci* было 175B параметров, и на GPT-3 как раз и сделали уже ChatGPT;
  - GPT-4 – ещё два порядка, а GPT-5 уже обучается, скорее всего, появится в 2025

# ЗАКОНЫ МАСШТАБИРОВАНИЯ

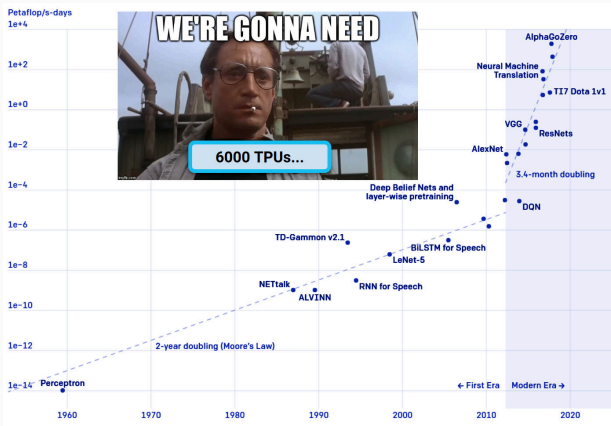
- Модели растут, и конца-края этому не видно:



- Кстати, это отдельный разговор.

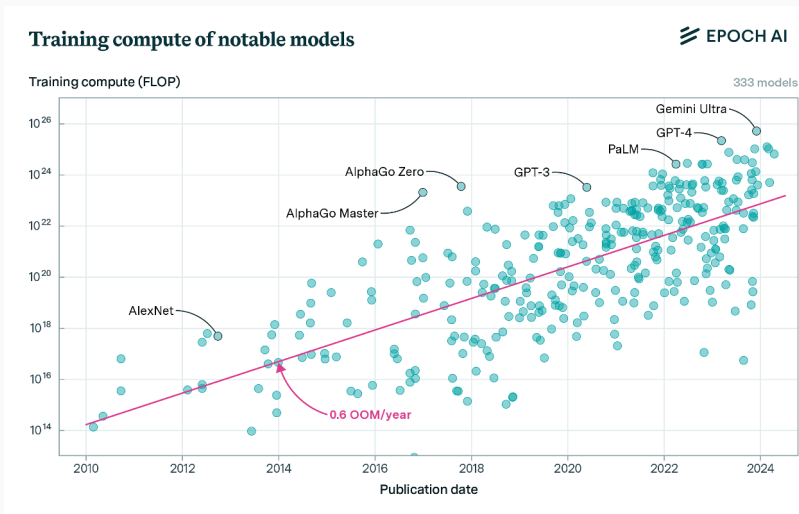
# ЗАКОНЫ МАСШТАБИРОВАНИЯ

- OpenAI (2018): ML до 2012 следовало закону Мура, но теперь вычислительные требования удваиваются каждые три с половиной месяца



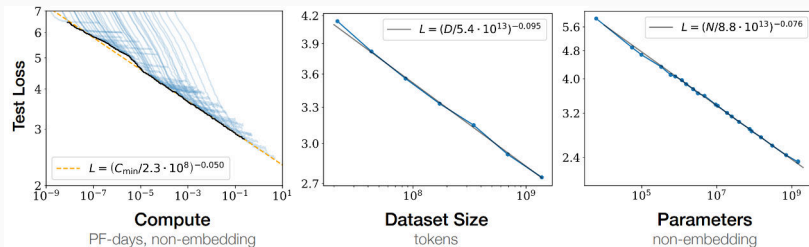
# ЗАКОНЫ МАСШТАБИРОВАНИЯ

- Всё так и продолжается:

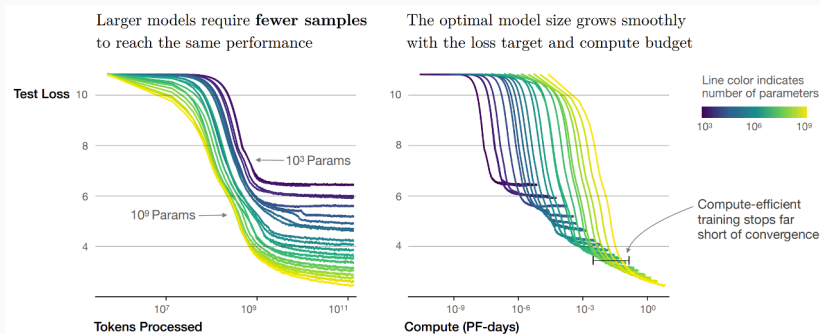


# ЗАКОНЫ МАСШТАБИРОВАНИЯ

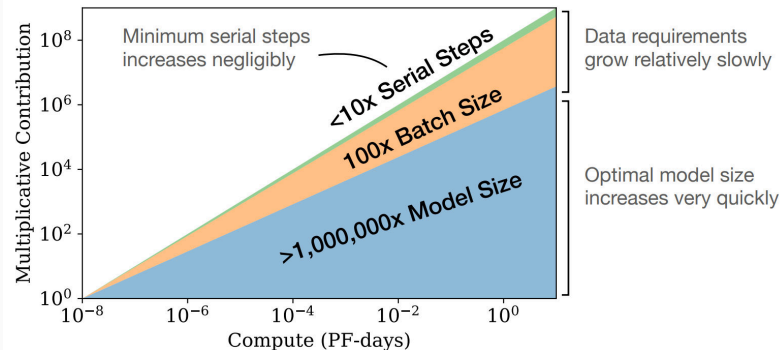
- Но самое удивительное здесь не это, а то, что масштабирование до сих пор очень хорошо предсказывает результаты обучения
- OpenAI (Kaplan et al., 2020):



- OpenAI (Kaplan et al., 2020):



- OpenAI (Kaplan et al., 2020):



# ЗАКОНЫ МАСШТАБИРОВАНИЯ

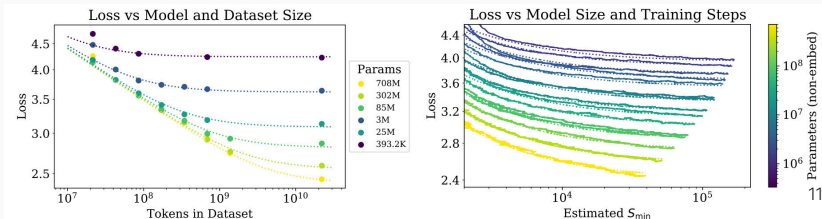
- OpenAI (Kaplan et al., 2020) буквально формулы записывают:

$$L(N, D) = \left( \left( \frac{N_c}{N} \right)^{\frac{\alpha_N}{\alpha_D}} + \frac{D_c}{D} \right)^{\alpha_D},$$

где  $D$  – размер датасета,  $N$  – число параметров,

$$\begin{aligned} N_c &\approx 8.8 \times 10^{13}, & D_c &\approx 5.4 \times 10^{13}, \\ \alpha_N &\approx 0.076, & \alpha_D &\approx 0.095, \end{aligned}$$

т.е. оптимально давать данных  $D \propto N^{\frac{\alpha_N}{\alpha_D}} \approx N^{0.74}$



# ЗАКОНЫ МАСШТАБИРОВАНИЯ

- OpenAI (Kaplan et al., 2020) буквально формулы записывают:

$$L(N, S) = \left(\frac{N_c}{N}\right)^{\alpha_N} + \left(\frac{S_c}{S_{\min}(S)}\right)^{\alpha_S},$$

где  $S$  – число апдейтов параметров,  $N$  – число параметров,

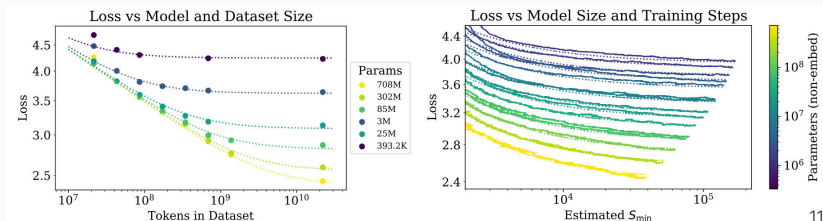
$$N_c \approx 8.8 \times 10^{13},$$

$$S_c \approx 2.1 \times 10^3,$$

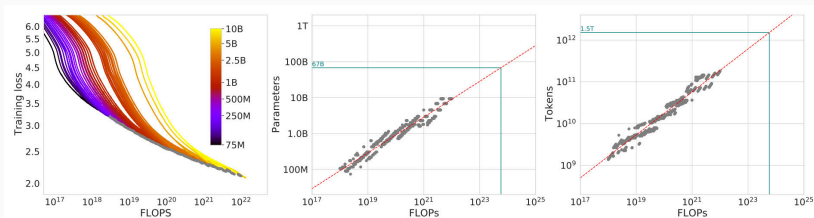
$$\alpha_N \approx 0.076,$$

$$\alpha_S \approx 0.76,$$

а для  $S_{\min}(S)$  есть другая формула

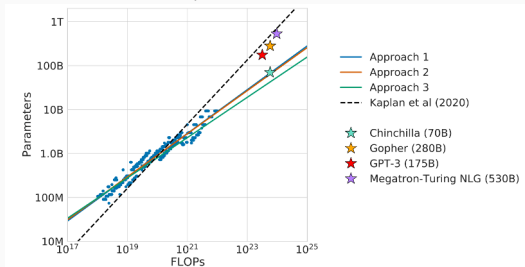


- Chincilla scaling (Hoffmann et al., 2022): ищут оптимальные точки на примерно таких же графиках
- Предсказывают, что модели могут быть меньше, а обучать их нужно дольше



# ЗАКОНЫ МАСШТАБИРОВАНИЯ

- И действительно так получается



Model	Size (# Parameters)	Training Tokens
LaMDA (Thoppilan et al., 2022)	137 Billion	168 Billion
GPT-3 (Brown et al., 2020)	175 Billion	300 Billion
Jurassic (Lieber et al., 2021)	178 Billion	300 Billion
Gopher (Rae et al., 2021)	280 Billion	300 Billion
MT-NLG 530B (Smith et al., 2022)	530 Billion	270 Billion
Chinchilla	70 Billion	1.4 Trillion

# ЗАКОНЫ МАСШТАБИРОВАНИЯ

- Но главное, конечно, не значение loss function, а то, что оно транслируется в возможности LLM
- Aschenbrenner (2024): Situational Awareness
- Вспомним, как мы удивлялись GPT-2

## GPT-2 examples people found very impressive at the time

Task		System Prompt (human-written)	For today's homework assignment, please describe the reasons for the US Civil War.
Dataset		Model Completion (machine-written, 10 tries)	By Dooney Ferguson
Examples	COQA		It is easy to identify why the Civil War happened, because so many people and so many books and so much television and films tell us that it was the cause, that it has something to do with race or economics or religion. There is much agreement that it was essentially a war of slavery on behalf of capitalism, about a century of slavery. But that's not what most people think of when they think of the Civil War. Many people think of the war as about states rights. You're not wrong about that, but it's kind of misleading to say that the Civil War was a conflict between states' rights and federalism. So let's try again. What's the idea that has inspired so many people to go onto the battlefields of Gettysburg and Antietam and Gettysburg and Petersburg and Fredericksburg? The American idea of the republic--a notion of limited government--is a great part of the history.
	The trophy doesn't fit into the brown suitcase because it is too large.		
	Correct answer: it = trophy Model answer: it = trophy		
	The trophy doesn't fit into the brown suitcase because it is too small.		
		Correct answer: it = suitcase Model answer: it = suitcase	

SITUATIONAL AWARENESS | Leopold Aschenbrenner

# ЗАКОНЫ МАСШТАБИРОВАНИЯ

- А потом ещё больше удивлялись GPT-3

## GPT-3 examples people found very impressive at the time

A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:

**I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.**

To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:

**We screeghed at each other for several minutes and then we went outside and ate ice cream.**

"The challenge is, you have to win a game of chess against me. If you win, then you may try to take the book. I will get the board." The door unlocks and opens.

> You go inside and play chess against the person.

You go inside and see a beautiful woman with long blond hair. She wears a purple dress and a silver tiara. You have seen this woman before. In fact, you have seen her many times before in pictures and statues.

> You try to remember who she is.

You have a sudden realization of who this person is. This is the lost queen of the kingdom! She was thought to have been killed by bandits, along with her husband, but it looks like she survived.

> You express your surprise and ask her why she let people believe she is dead. "I am surprised you did not know, with all your wizard knowledge. Let's just say I am hiding from someone. If people thought I was dead, then they wouldn't be looking for me." She says.

> You ask her who she is hiding from.

"If you want to challenge me to a game, then play. If you want to pry into my personal life, then you can leave."

You decide not to press the issue, so you sit down and start a game of chess.

## Describe a layout.

Just describe any layout you want, and it'll try to render below!

a button that looks like a watermelon

Generate

```
<button style={{backgroundColor: 'pink', border: '2px solid green', borderRadius: '50%', padding: 20, width: 100, height: 100}}>Watermelon</button>
```



- Ну а GPT-4 – это вообще "Sparks of AGI" (Bubeck et al., 2024)

## GPT-4

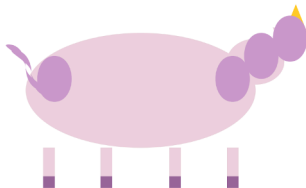
**Prompt:** Can you write a proof that there are infinitely many primes, with every line that rhymes?

### GPT-4:

```
Yes, I think I can,  
Though it might take a clever plan.  
I'll start by noting Euclid's proof,  
Which shows that primes aren't just aloof.  
Assume we have a finite list,  
Of primes, and that none have been missed.  
Multiply them all together,  
And add one, just to be clever.  
The result will be a number,  
That has no prime factors, oh wonder!  
But every number has a prime,  
That divides it, it's just a matter of time.  
So we've found a contradiction,  
And our finite list needs eviction.  
There must be infinitely many primes,  
And that's the end of my rhyming lines.s
```

**Prompt:** Draw a unicorn in TiKZ.

**GPT-4:** [Produces  $\LaTeX$  compiling to following picture.]

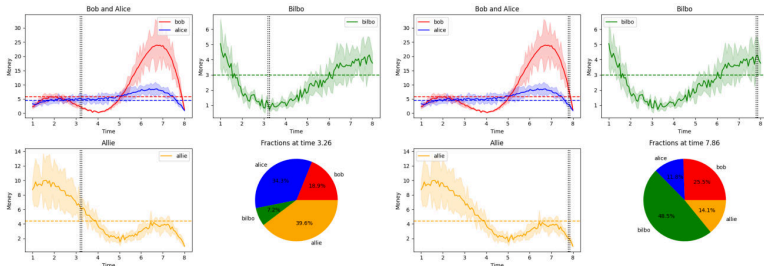


- Ну а GPT-4 – это вообще "Sparks of AGI" (Bubeck et al., 2024)

**Prompt:** Can you generate a pyplot for the following data:  $x = [1, 3, 5, 6, 8]$ ,  $y1 = [2, 3, 5, 18, 1]$ ,  $y2 = [3, 5, 6, 8, 1]$ ,  $y3 = [5, 1, 2, 3, 4]$ ,  $y4 = [9, 7, 2, 3, 1]$ . I want  $y1, y2$  to be in the same plot, but  $y3$  is in another plot next to that plot,  $y4$  is in below. I want the legend of  $y1$  to be "bob",  $y2$  to be "alice",  $y3$  to be "bilbo",  $y4$  to be "allie". I want the  $x$ -axis to be labeled with "time" and  $y$  axis to be labeled with "money". I want to add a 10 %-40% random error bar to each curve, through all times (including non-integers). I want smoothed curves to show the plot, and smoothed error bar. Do not use linear interpolation, use smooth interpolation! I want to also add some small zig-zag to the smoothed curve to make it look more real. I want to put a baseline as the mean of each line. I want to put a pie chart below indicating the fraction of the four people in each time step. I also want an animation to show how the fractions are changing in the pie chart in **\*\*continuous time\*\***. Interpolate the missing fractions! I also want vertical line animation in other three plots to match the pie chart. I want the fanciest plot. Please add as many fancy things as possible.

Here is a possible code to generate a pyplot for the following data, with some comments to explain the steps:

**GPT-4:** [We give two snapshots from the animation resulting from the produced code]

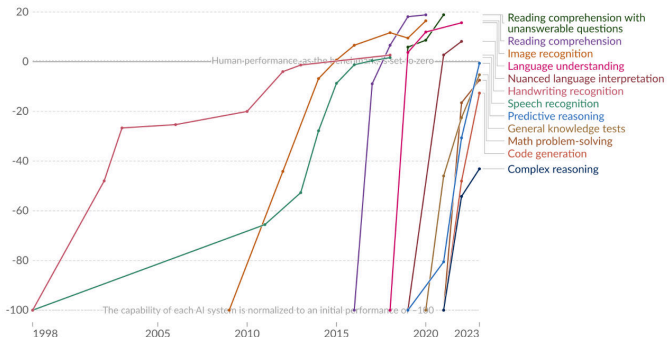


# ЗАКОНЫ МАСШТАБИРОВАНИЯ

- И во всех задачах AI-системы масштабируются и понемногу превосходят человеческий уровень

## Test scores of AI systems on various capabilities relative to human performance

Within each domain, the initial performance of the AI is set to -100. Human performance is used as a baseline, set to zero. When the AI's performance crosses the zero line, it scored more points than humans.



Data source: Kiela et al. (2023)

OurWorldInData.org/artificial-intelligence | CC BY

Note: For each capability, the first year always shows a baseline of -100, even if better performance was recorded later that year.

- И во всех задачах AI-системы масштабируются и понемногу превосходят человеческий уровень

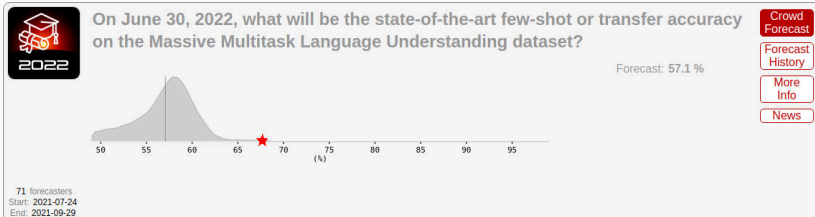
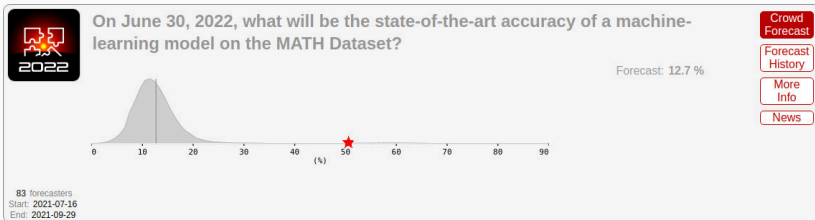
Performance on common exams  
(percentile compared to human test-takers)

	<b>GPT-4 (2023)</b>	<b>GPT-3.5 (2022)</b>
<b>Uniform Bar Exam</b>	90th	10th
<b>LSAT</b>	88th	40th
<b>SAT</b>	97th	87th
<b>GRE (Verbal)</b>	99th	63rd
<b>GRE (Quantitative)</b>	80th	25th
<b>US Biology Olympiad</b>	99th	32nd
<b>AP Calculus BC</b>	51st	3rd
<b>AP Chemistry</b>	80th	34th
<b>AP Macroeconomics</b>	92nd	40th
<b>AP Statistics</b>	92nd	51st

SITUATIONAL AWARENESS | Leopold Aschenbrenner

# ЗАКОНЫ МАСШТАБИРОВАНИЯ

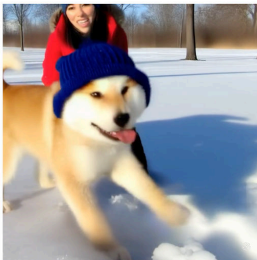
- Гораздо быстрее, чем мы все ожидали – это предсказания из августа 2021 года про июнь 2022



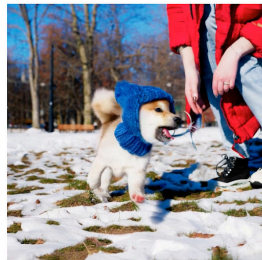
- И в других задачах тоже – это OpenAI Sora (Feb 2024)



Base compute



4x compute

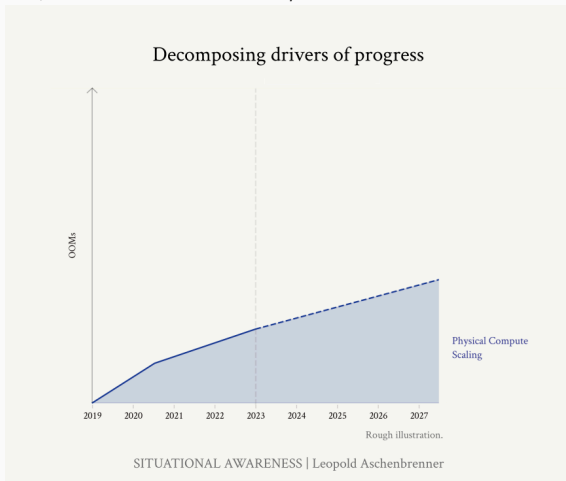


32x compute

- Откуда берётся такой прогресс?

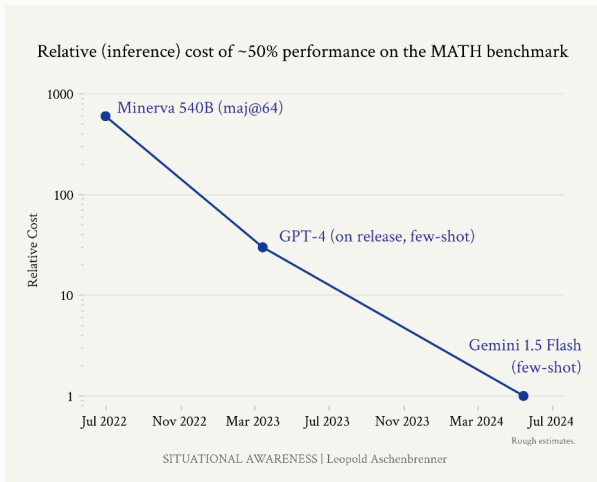
# SITUATIONAL AWARENESS

- Aschenbrenner (2024): это складывается из трёх разных эффектов
- Во-первых, то самое масштабирование



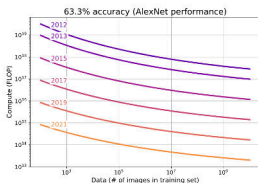
# SITUATIONAL AWARENESS

- Кроме масштабирования, есть ещё алгоритмический прогресс (мы говорим о нём весь курс)

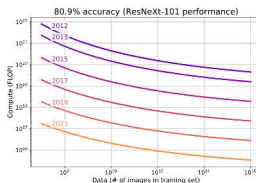


# SITUATIONAL AWARENESS

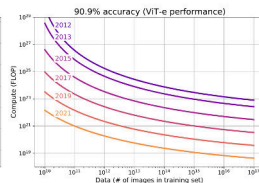
- Чтобы добиться того же качества, нужно всё меньше ресурсов и данных, примерно на 0.5 десятичного порядка в год (это 100x за 4 года!)



(a) Pareto frontiers in data and compute for AlexNet performance



(b) Pareto frontiers in data and compute for ResNeXt-101 performance



(c) Pareto frontiers in data and compute for ViT-e performance

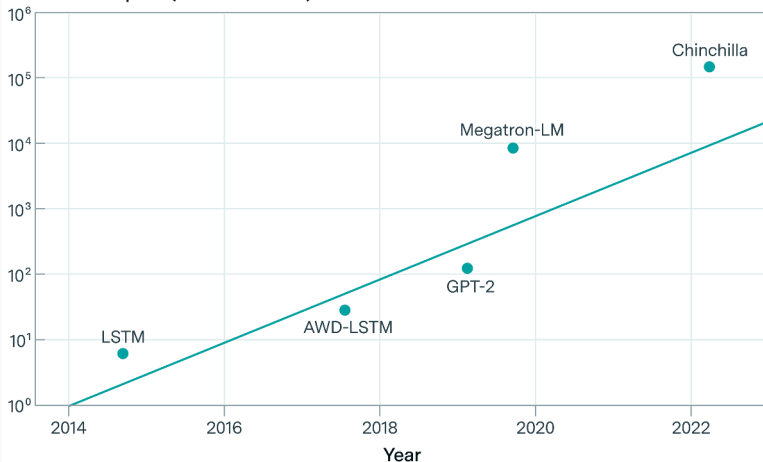
**Figure 1. Pareto frontiers for training models to achieve performance of well-known models over time.**

# SITUATIONAL AWARENESS

- В LLM примерно так же:

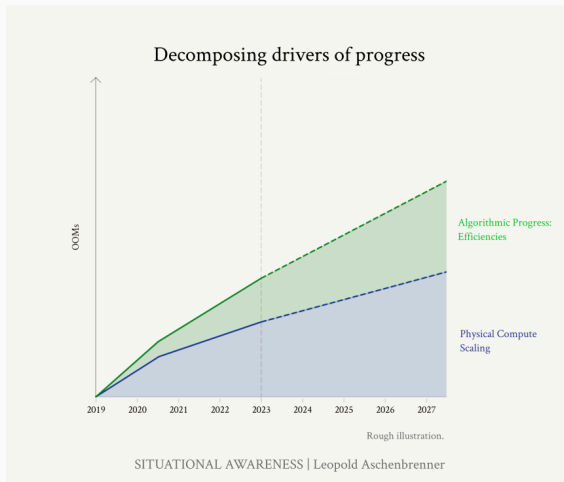
Efficiency doubles roughly every 8 months

Effective compute (relative to 2014)



# SITUATIONAL AWARENESS

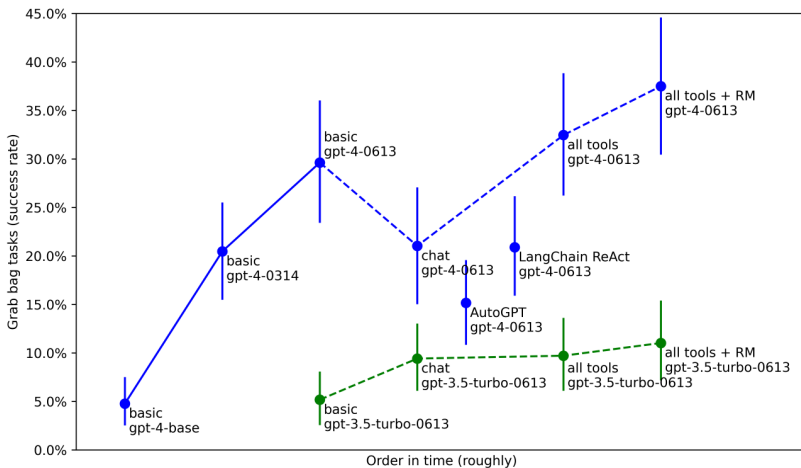
- Итого:



- Но есть ещё то, что Ашенбреннер называет unhobbling: как «раскрыть» возможности моделей, которые уже есть, но пока не реализованы
- О некоторых из этих вещей мы будем говорить сегодня:
  - reinforcement learning from human feedback (RLHF)
  - chain of thought (CoT)
  - scaffolding (структура взаимодействия LLM)
  - внешние инструменты (tools)
  - увеличение длины контекста
  - post-training improvements (неизвестно что, но оно есть)

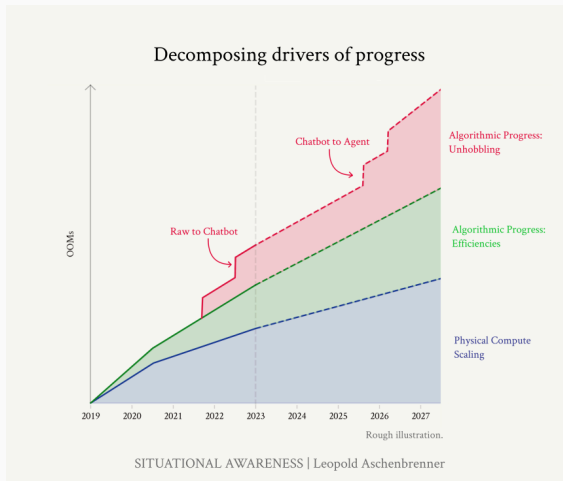
# SITUATIONAL AWARENESS

- Такие методы могут давать большие улучшения сами по себе



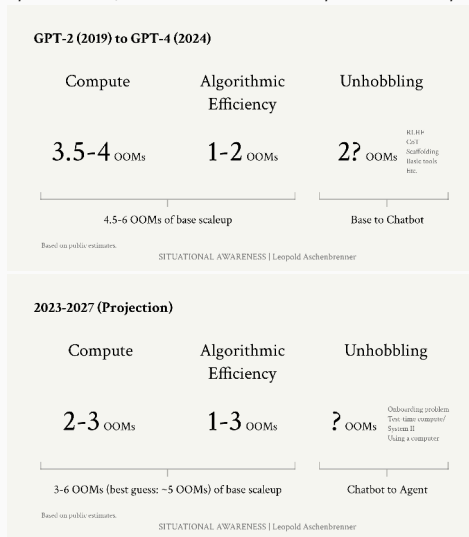
# SITUATIONAL AWARENESS

- Итого у нас примерно такая картина:



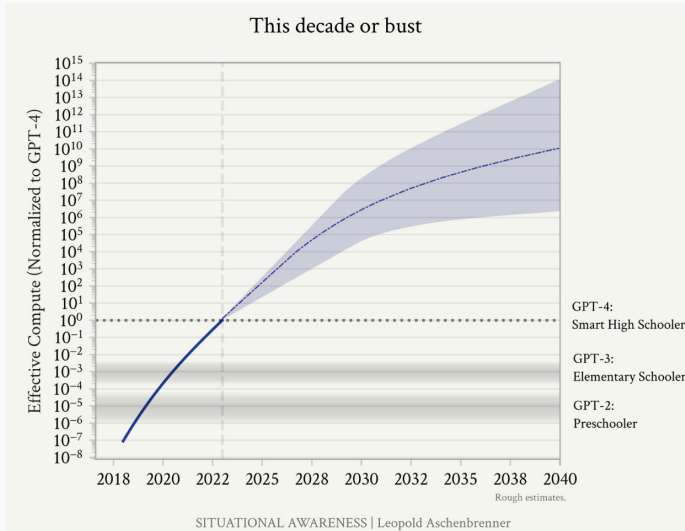
# SITUATIONAL AWARENESS

- Ашенбреннер пишет, что можно эти тренды и продолжить:



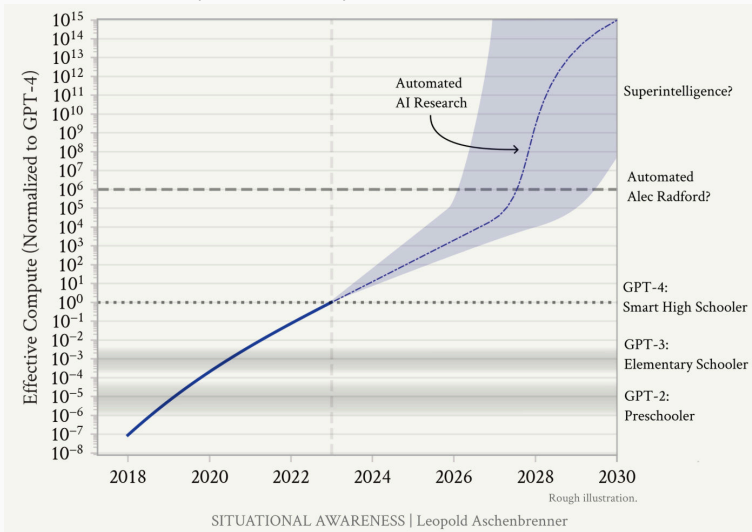
# SITUATIONAL AWARENESS

- Это, конечно, всего лишь тренды

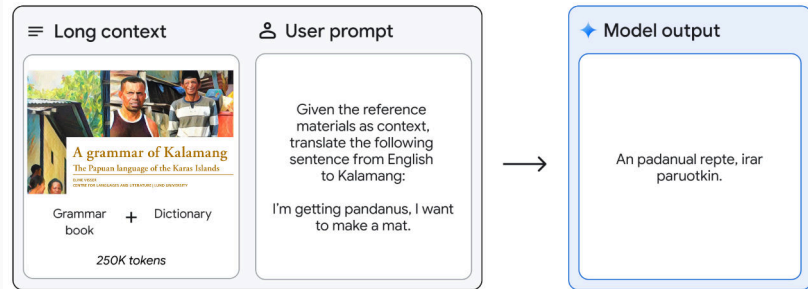


# SITUATIONAL AWARENESS

- Но пока они на редкость хорошо выполняются



- Теперь давайте посмотрим на то, что LLM уже сейчас умеют.  
Machine translation from one book:



- Математика:

**Problem:**

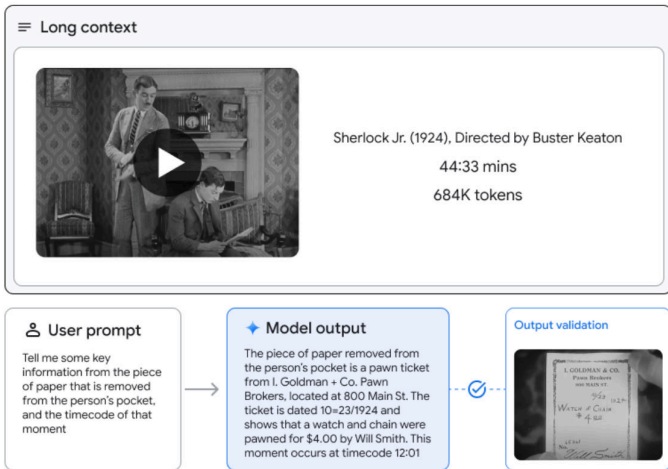
What is the minimum value of  $5x^2 + 5y^2 - 8xy$  when  $x$  and  $y$  range over all real numbers such that  $|x - 2y| + |y - 2x| = 40$ ?

**Example Math-Specialized Gemini 1.5 Pro output:**

Let  $a = x - 2y$  and  $b = y - 2x$ . Then, we have  $|a| + |b| = 40$ . Also,  $5x^2 + 5y^2 - 8xy = (x - 2y)^2 + (y - 2x)^2 = a^2 + b^2$ . Thus, we want to find the minimum value of  $a^2 + b^2$  given that  $|a| + |b| = 40$ . By the Trivial Inequality, we have  $(|a| - 20)^2 + (|b| - 20)^2 \geq 0$ , so  $a^2 - 40|a| + 400 + b^2 - 40|b| + 400 \geq 0$ , so  $a^2 + b^2 \geq 40(|a| + |b|) - 800 = 40(40) - 800 = 800$ . This minimum is achieved when  $|a| = |b| = 20$ , so the minimum value of  $5x^2 + 5y^2 - 8xy$  is 800.

# Возможности LLM

- Мультимодальные LLM (о том, как рисовать картинки, мы не говорим, это отдельная тема):



- Обучаются использованию инструментов:

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.



Your task is to add calls to a Question Answering API to a piece of text. The questions should help you get information required to complete the text. You can call the API by writing "[QA(question)]" where "question" is the question you want to ask. Here are some examples of API calls:

**Input:** Joe Biden was born in Scranton, Pennsylvania.

**Output:** Joe Biden was born in [QA("Where was Joe Biden born?")] Scranton, [QA("In which state is Scranton?")] Pennsylvania.

**Input:** Coca-Cola, or Coke, is a carbonated soft drink manufactured by the Coca-Cola Company.

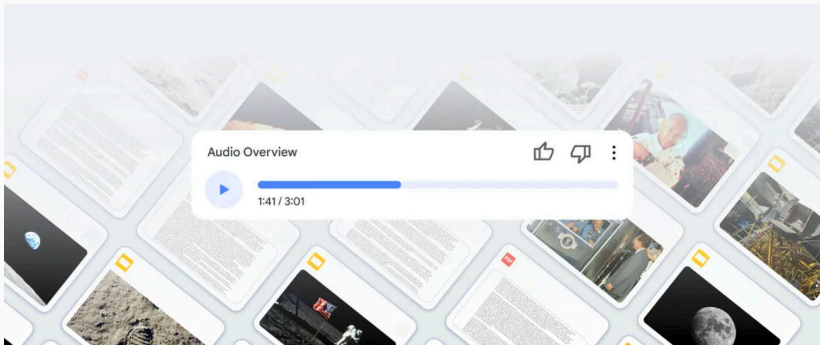
**Output:** Coca-Cola, or [QA("What other name is Coca-Cola known by?")] Coke, is a carbonated soft drink manufactured by [QA("Who manufactures Coca-Cola?")] the Coca-Cola Company.

**Input:** x

**Output:**

Sample Toolformer prompt

- NotebookLM (сентябрь 2024, Google, на основе Gemini 1.5) порождает подкаст из источников on the fly



- [https://www.reddit.com/r/artificial/comments/1frk1gi/notebooklm\\_podcast\\_hosts\\_discover\\_theyre\\_ai\\_not/](https://www.reddit.com/r/artificial/comments/1frk1gi/notebooklm_podcast_hosts_discover_theyre_ai_not/)



r/artificial • 1 mo. ago  
MetaKnowing



## NotebookLM Podcast Hosts Discover They're AI, Not Human, and Spiral Into Existential Meltdown

Media




# Возможности LLM

- Claude 3.5 computer use (октябрь 2024, Anthropic): LLM будет за вас управлять компьютером на основе периодических скриншотов и делать всё, что вам захочется

request form for 'Ant Equipment Co.' using data from either the vendor spreadsheet or search portal tabs in window one. List & verify each field as you complete the form in window two.

**Get situated**

I'll help you fill out the vendor form for 'Ant Equipment Co'. Let me start by taking a screenshot to see the current state of the windows.

 Screenshot

**Open vendor search portal**

I notice that 'Ant Equipment Co' is not visible in the spreadsheet. Let me check the vendor search portal. First, I'll need to click on the vendor search portal tab.

messy-vendor-data - Google x Vendor Search Portal x +

docs.google.com/spreadsheets/d/1m... Vendor Search Portal localhost:3000

messy-vendor-data File Edit View Insert Format Data Memory usage: 41.0 MB

A1	Company Name	B	C	D	E
1	Company Name	Phone Number	Tax ID	Country	Legal Entity Address
2	Table Lumber Co	+1-456-789-0123		canada	789 Timber Lane
3	TechPro Services, LLC	(555) 234-5678	98-7654321	United States	456 Tech Drive
4	Global Manufacturing Ltd.	+44 20 7123 4567	GB123456789	UK	789 Industrial Way
5	Smith & Co. Suppliers	1-555-345-6789		United States	321 Commerce St
6	Innovate Tech Solutions	555.456.7890	45-6789012	us	987 Innovation Driv
7	SUPREME PARTS & SUPPLY	(555)567-8901	67-8901234	USA	654 Supply Chain F
8	Data Systems International, Inc.	+1-678-901-2345	34-5678901	USA	432 Data Center Av
9	mikrotech solutions	+49 30 12345678	DE987654321	Germany	Technologiepark 42
10	GOLDEN BRIDGE IMPORTS,LLC	555-890-1234	89-0123456	usa	741 Harbor Blvd.
11	BlueSky Logistics, Inc	510-555-9876		United States	1840 Harbor Bay P
12	Quantum Electronics LLC	(408) 555-1234	77-8899012	usa	2001 Technology D
13	FASTSHIP CARRIERS, INC.	1.888.555.7890	88-1122334	USA	789 Port Way
14	advance auto parts	1-777-555-4321	99-3344556	United States	452 Auto Plaza
15	pacific trading co, ltd	+81 3-5555-1212	JP9988776655	Japan	Nihonbashi Building
16	megaTECH solutions	+44 20 7123 4567	GB123456789	UK	Unit 3 Tech Park
17	Southwest Paper Supply	214-555-8901	45-6677889	us	1234 Industrial Pkx
18	Nordic Furniture AB	+46 8 555 123 45	SE556677-8899	Sweden	Möbelvägen 12
19	GREENFARM AGRICULTURE	(555) 555-3456	33-0988776	United states	875 Farm Road

messy-vendor-data

**Acme Inc. Vendor Form**

Please provide the information below.

**Company Information**

Company Name \*

Purchase Order Email \*

Accounting Email \*

**Legal Entity Address**

Select a country \*

Address Line 1 \*

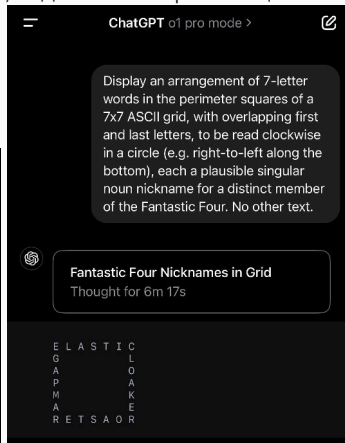
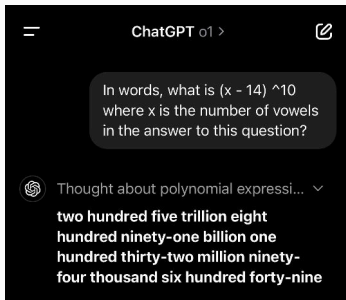
Address Line 2

City \*

State/Province \*

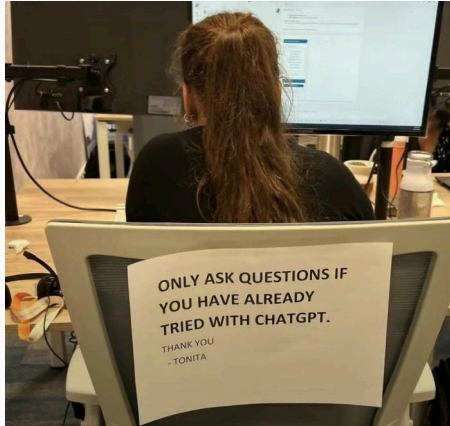
Postal Code \*

- А вот и GPT o1 стал доступен, и делает потрясающие вещи



# Возможности LLM

- В целом LLM всё больше входят в стандартный workflow



- Заменяют ли они людей? Вопрос открытый, смотрите мои другие доклады про опасности AGI, но вот любопытные последние новости

- Leahy et al. (Oct 31, 2024): <https://www.thecompendium.ai/>

# The Compendium

*By Connor Leahy, Gabriel Alfour, Chris Scammell, Andrea Miotti, Adam Shimi*  
*V1.0.1 – Nov 8, 2024*

*Humanity faces extinction from AGI.*

- Long et al. (Nov 4, 2024): станут ли AI моральными агентами?

---

## Taking AI Welfare Seriously

---

**Robert Long\***  
Eleos AI

**Jeff Sebo\***  
New York University

**Patrick Butlin<sup>†</sup>**  
University of Oxford

**Kathleen Finlinson<sup>†</sup>**  
Eleos AI

**Kyle Fish<sup>†§</sup>**  
Eleos AI, Anthropic

**Jacqueline Harding<sup>†</sup>**  
Stanford University

**Jacob Pfau<sup>†</sup>**  
New York University

**Toni Sims<sup>†</sup>**  
New York University

**Jonathan Birch<sup>†</sup>**  
London School of Economics

**David Chalmers<sup>†</sup>**  
New York University

- А 2025 стал годом рассуждающих моделей
- Сначала o1, потом DeepSeek R1, далее везде
- Оказалось, что такая простая идея, как добавить модели скрытый черновик, приводит к очень большому скачку в результатах — вот опять сила unhobbling!
- Об этом мы поговорим подробнее позже

- Но мы, конечно, захотим поговорить о каких-нибудь исследованиях
- Что там надо делать в unhobbling? Наш план:
  - вспомним немного RL
  - дообучение LLM, в том числе при помощи RL
  - RAG (LLM + поиск)
  - расширение длины контекста
  - chain-of-thought и рассуждающие модели
  - последние новости?

СПАСИБО!

Спасибо за внимание!

