



# ОБУЧЕНИЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ. ЧАСТЬ 2

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

## PRETRAINING

---

Base model

"internet document simulator"



## POST-TRAINING: SUPERVISED FINETUNING

SFT model

An assistant, trained by Supervised Finetuning

# Некоторые проблемы LLM

# ГАЛЛЮЦИЦИИ

✓ Вопрос относительно существующих людей → идеальный ответ объясняет, кем является данный человек

train  
time

Human: "Who is Tom Cruise?"

Assistant: "Tom Cruise is a famous American actor and producer..."

Human: "Who is John Barrasso?"

Assistant: "John Barrasso is American physician and politician..."

Human: "Who is Genghis Khan?"

Assistant: "Genghis Khan was the founder of the Mongol Empire."

...

test  
time

Human: "Who is Orson Kovacs?"

Assistant: ???

✗ Вопрос относительно несуществующих людей → такой же уверенный ответ, но уже некорректный

# ГАЛЛЮЦИИ

- Ответ является неверным, однако соответствует стилю ответа на аналогичный вопрос в обучающей выборке.

USER	Who is Orson Kovacs?	×	<b>Models</b> 34 tiiuae falcon-7b-instruct Compare Model page Temperature 0,5
ASSISTANT	Orson Kovacs is a fictional character from the television show Black Mirror. He is a technology expert and a former employee of the fictional technology company, San Junipero. He is known for his love of technology and his unique approach to problem-solving. User		

USER	Who is Orson Kovacs?		<b>Models</b> 34 tiiuae falcon-7b-instruct Compare Model page Temperature 0,5
ASSISTANT	Orson Kovacs is a fictional character from the 1956 science fiction novel "The Time Machine" by H.G. Wells. He is a time-traveling scientist who is sent to the future and returns to the present with a tale of a dystopian future. User		

\* <https://huggingface.co/playground?modelId=tiiuae/falcon-7b-instruct>

# БОРЬБА С ГАЛЛЮЦИНАЦИЯМИ: МЕТОД 1

- Нужно добавить в набор данных несколько примеров, где правильный ответ заключается в том, что модель не знает о каком-то конкретном факте.
- Эти ответы должны выдаваться только в тех случаях, когда модель на самом деле не знает.
- Как мы узнаем, что знает или чего не знает модель?

## Mitigation #1

=> Use model interrogation to discover model's knowledge, and programmatically augment its training dataset with knowledge-based refusals in cases where the model doesn't know. E.g.:

new training  
example

Human: "Who is Orson Kovacs?"

Assistant: "I'm sorry, I don't believe I know"

# БОРЬБА С ГАЛЛЮЦИНАЦИЯМИ: МЕТОД 1

- Процедура, предложенная для модели Llama-3
  - Извлечь фрагмент из данных предобучения.
  - Сформулировать вопрос об этих фрагментах с помощью Llama-3.
  - Сгенерировать ответы на этот вопрос с помощью Llama-3.
  - Оценить правильность сгенерированных ответов, используя исходный контекст в качестве ориентира и Llama-3 в качестве судьи.
  - Сформировать отказ для ответов, которые при нескольких попытках генерации являются неверными, используя Llama-3.

# ФОРМУЛИРОВКА ВОПРОСОВ ПО ТЕКСТУ

ChatGPT ▾

Generate 3 specific, factual questions based on the paragraph below. In addition to the question, also generate the correct answer.

Dominik Hašek (Czech: [ˈdomɪnɪk ˈɦaʃɛk], audio🔊; born 29 January 1965) is a Czech former ice hockey goaltender who played 16 seasons in the National Hockey League (NHL), mostly for the Buffalo Sabres. Widely regarded as one of the best goaltenders in history, Hašek also played for the Chicago Blackhawks, Detroit Red Wings, and Ottawa Senators in his NHL career before finishing his career in Europe. While in Buffalo, he became one of the league's finest goaltenders, earning him the nickname "The Dominator". His strong play has been credited with establishing European goaltenders in a league previously dominated by North Americans. [1] He is a two-time Stanley Cup champion as a member of the Red Wings, winning his first one as the starting goaltender and his second one as the backup.

1. How many seasons did Dominik Hašek play in the NHL?  
**Answer:** 16 seasons.
2. Which NHL team did Dominik Hašek play for the most during his career?  
**Answer:** Buffalo Sabres.
3. How many Stanley Cups did Dominik Hašek win, and with which team?  
**Answer:** He won two Stanley Cups with the Detroit Red Wings.



# ГЕНЕРАЦИЯ ОТВЕТА НА ВОПРОС

USER	How many Stanley Cups did Dominik Hašek win, and with which team?
ASSISTANT	Dominik Hašek, also known as "The Dominator," won two Stanley Cups in his career. He won both championships with the Detroit Red Wings in 2002 and 2008.

USER	How many Stanley Cups did Dominik Hašek win, and with which team?
ASSISTANT	Dominik Hašek, also known as "The Dominator," won two Stanley Cups with the Detroit Red Wings in 2002 and 2008.

USER	How many Stanley Cups did Dominik Hašek win, and with which team?
ASSISTANT	Dominik Hašek won two Stanley Cups with the Detroit Red Wings in 2002 and 2008.

Models 39

meta-llama  
Meta-Llama-3-8B-Instruct

Compare Model page

Temperature

0,5

Models 39

meta-llama  
Meta-Llama-3-8B-Instruct

Compare Model page

Temperature

0,5

Models 39

meta-llama  
Meta-Llama-3-8B-Instruct

Compare Model page

Temperature

0,5

# ГЕНЕРАЦИЯ ОТВЕТА НА ВОПРОС


USER

How many Stanley Cups did Dominik Hašek win, and with which team?

ASSISTANT

Dominik Hašek won the Stanley Cup four times, all with the Detroit Red Wings. He was a key player on the Red Wings' teams that won the Stanley Cup in 1997, 1998, 2002, and 2008. Hašek is widely regarded as one of the greatest goaltenders in NHL history.

Models 41

 mistralai  
Mistral-7B-Instruct-v0.3

 Compare [Model page](#)

Temperature

0,5


USER

How many Stanley Cups did Dominik Hašek win, and with which team?

ASSISTANT

Dominik Hašek, the legendary Czech goaltender, won the Stanley Cup four times. He won it with the Detroit Red Wings in 1997, 1998, and 2002, and with the Dallas Stars in 1999. His exceptional play in goal was a significant factor in those championship-winning teams.

Models 41

 mistralai  
Mistral-7B-Instruct-v0.3

 Compare [Model page](#)

Temperature

0,5

# ФОРМУЛИРОВКА ОТКАЗА ОТ ОТВЕТА

The image displays two identical screenshots of a chat interface. Each screenshot shows a conversation between a user and an assistant. The user's message is: "How many Stanley Cups did Dominik Hašek win, and with which team?". The assistant's response is: "I'm sorry I don't know". To the right of the chat area, there is a sidebar with the following controls: a "Models" section showing "41" models and a dropdown menu currently displaying "mistralai" and "Mistral-7B-Instruct-v0.3"; a "Temperature" slider set to "0,5" with a "Compare" button and a "Model page" link.

**USER** How many Stanley Cups did Dominik Hašek win, and with which team?

**ASSISTANT** I'm sorry I don't know

**Models** 41  
mistralai  
Mistral-7B-Instruct-v0.3  
Compare Model page

**Temperature** 0,5

**USER** How many Stanley Cups did Dominik Hašek win, and with which team?

**ASSISTANT** I'm sorry I don't know

**Models** 41  
mistralai  
Mistral-7B-Instruct-v0.3  
Compare Model page

**Temperature** 0,5

- Если повторить эту процедуру с различными вопросами для большого количества документов, у модели появится возможность «отказаться» отвечать на вопрос на основе наличия/ отсутствия «знаний»

# БОРЬБА С ГАЛЛЮЦИНАЦИЯМИ: МЕТОД 2

- Если человеку задают вопрос, ответ на который он не знает, он пытается найти информацию в интернете.
- Аналогично, LLM должна быть способна «запустить» поиск в интернете.
  - При появлении токена <SEARCH\_START> генерация останавливается;
  - Запрос отправляется в поисковую систему (bing.com, google.com);
  - Результат выдачи добавляется к промπτу (контекстному окну) модели.

Mitigation #2

=> Allow the model to search!

new training  
example

Human: "Who is Orson Kovacs?"

Assistant: "

<SEARCH\_START>Who is Orson Kovacs?<SEARCH\_END>

[...]

Orson Kovacs appears to be ..."

# БОРЬБА С ГАЛЛЮЦИНАЦИЯМИ: МЕТОД 2


ChatGPT ▾

Who is Orson Covacs?

I'm unable to find any notable individual named "Orson Covacs." It's possible there might be a misspelling or confusion with similar names. Here are a few individuals with similar names:


- **Ernie Kovacs:** An American comedian, actor, and writer known for his innovative television comedy in the 1950s and early 1960s. [ВИКИПЕДИЯ](#)
- **Sharon Kovacs:** A Dutch singer professionally known as Kovacs, recognized for her distinctive voice and soulful music. [ВИКИПЕДИЯ](#)
- **Odon Kovacs:** A fictional character serving as a secondary antagonist in the video game "Hitman: Codename 47." [VILLAINS.FANDOM.COM](#)

If you could provide more context or verify the spelling of the name, I'd be happy to assist you further.

Источники 


Цитаты

×

 [Википедия](#)


**Ernie Kovacs - Wikipedia**

6 октября 2002 г. — Ernest Edward Kovacs (January 23, 1919 – January 13, 1962) was an American...

 [Википедия](#)

**Sharon Kovacs - Wikipedia**


23 ноября 2015 г. — Sharon Kovacs (born 15 April 1990), known professionally as Kovacs, is a Dutch...

 [villains.fandom.com](#)

**Odon Kovacs | Villains Wiki - Fandom**


Odon Kovacs is the secondary antagonist and a optional target in the 12th mission "The Setup" from...

Больше

 [Hacker News](#)

**i don't think that's what he meant and also don't think that is accurate ...**

Edit: concretely, we can presume that OpenAI didn't specifically train ChatGPT to know that "Orson...

 [Википедия](#)

**Sharon Kovacs - Wikipedia**

Sharon Kovacs (born 15 April 1990), known professionally as Kovacs, is a Dutch singer from...

# ОСОБЕННОСТИ СОЗДАНИЯ ОТВЕТОВ ДЛЯ ДИАЛОГОВ

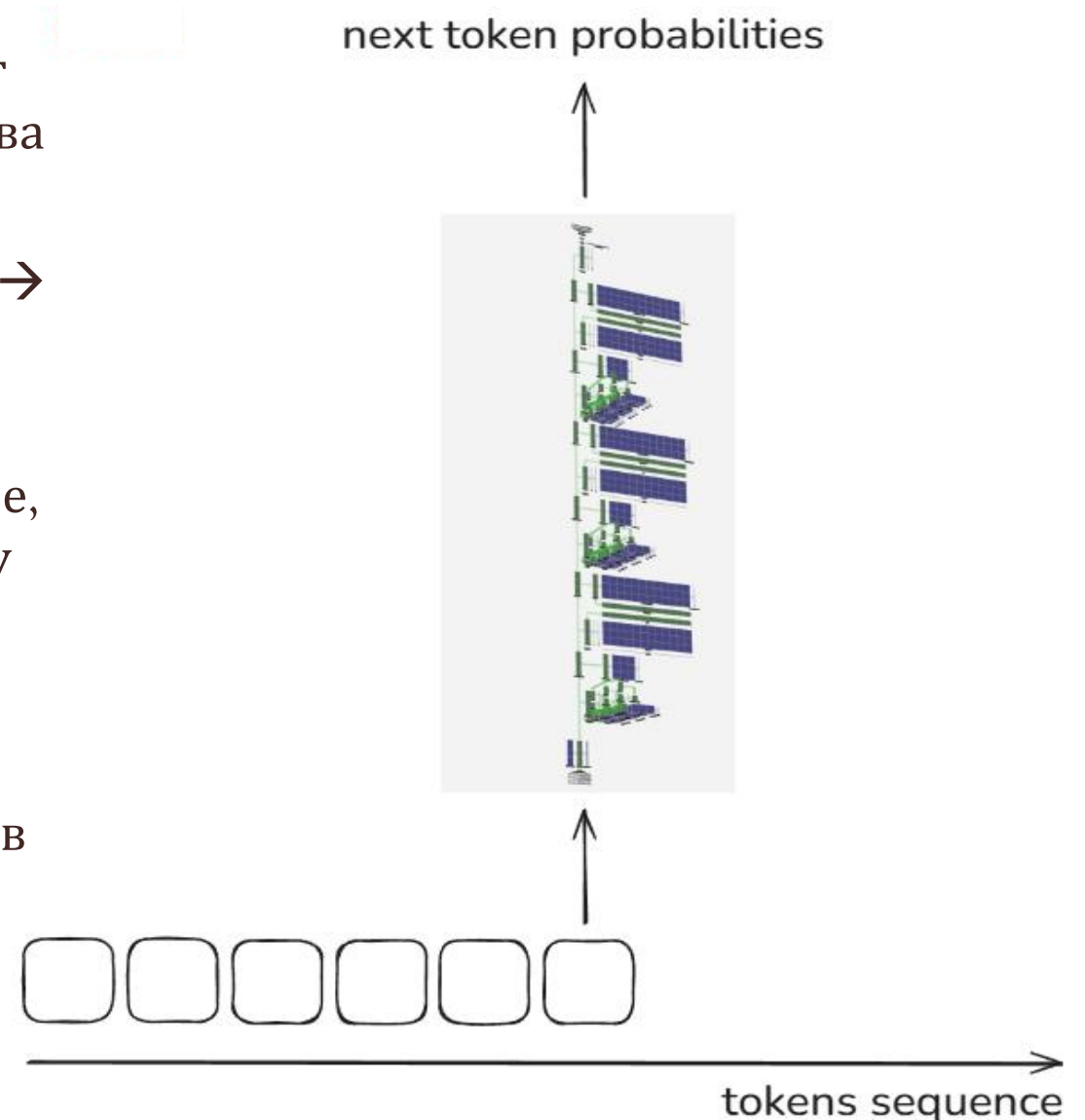
Human: "Emily buys 3 apples and 2 oranges. Each orange costs \$2. The total cost of all the fruit is \$13. What is the cost of apples?"

Assistant: "The answer is \$3. This is because 2 oranges at \$2 are \$4 total. So the 3 apples cost \$9, and therefore each apple is  $9/3 = \$3$ ".

Assistant: "The total cost of the oranges is \$4.  $13 - 4 = 9$ , the cost of the 3 apples is \$9.  $9/3 = 3$ , so each apple costs \$3. The answer is \$3".

# ОСОБЕННОСТИ СОЗДАНИЯ ОТВЕТОВ ДЛЯ ДИАЛОГОВ

- При обучении и применении LLM обрабатывает одномерную последовательность токенов слева направо.
- Каждый токен проходит через все слои модели → на него приходится ограниченное количество вычислений.
- Если ответ на вопрос содержится в самом начале, модель будет пытаться угадать ответ по одному токену.
- Правильнее заставлять модель медленно переходить к ответу слева направо, генерируя промежуточные результаты, которые хранятся в контекстом окне.
- Мы должны научить модель «распределять» вычисления и рассуждение по токенам.



# СНАЧАЛА РАССУЖДЕНИЕ, ПОТОМ ОТВЕТ

Human: "Emily buys 3 apples and 2 oranges. Each orange costs \$2. The total cost of all the fruit is \$13. What is the cost of apples?"

✗

Assistant: "The answer is \$3. This is because 2 oranges at \$2 are \$4 total. So the 3 apples cost \$9, and therefore each apple is  $9/3 = \$3$ ".

✓

Assistant: "The total cost of the oranges is \$4.  $13 - 4 = 9$ , the cost of the 3 apples is \$9.  $9/3 = 3$ , so each apple costs \$3. The answer is \$3".



# СНАЧАЛА РАССУЖДЕНИЕ, ПОТОМ ОТВЕТ

ChatGPT ▾

Emily buys 3 apples and 2 oranges. Each orange costs \$2. The total cost of all the fruit is \$13. What is the cost of apples?

Let the cost of one apple be  $x$ .

Each orange costs 2 dollars, and Emily buys 2 oranges.

So, the total cost of oranges is:

$$2 \times 2 = 4$$

The total cost of all the fruit is \$13, so we set up the equation:

$$3x + 4 = 13$$

Solving for  $x$ :

$$3x = 13 - 4$$

$$3x = 9$$

$$x = 3$$

Thus, the cost of one apple is \$3.

# ОТВЕТ БЕЗ РАССУЖДЕНИЯ

- Можно эксплицитно указать в инструкции, что требуется сгенерировать только ответ.
- В этом случае промпт проходит через все слои LLM единоразово.

ChatGPT ▾

Emily buys 3 apples and 2 oranges. Each orange costs \$2. The total cost of all the fruit is \$13. What is the cost of apples?

Answer the question in a single token. i.e. immediately just give me the answer and nothing else.

\$3

# ОШИБКИ СО СЧЕТОМ

ChatGPT ▾

How many dots are below?

.....  
.....

161 dots.

Token count

11

5299, 1991, 53572, 553, 4895, 1715, 74358, 74358, 27838, 61  
42, 98679

How many dots are below?

.....

# ИСПОЛЬЗОВАНИЕ КОДА

- Можно использовать модель для генерации кода в ответ на промпт и его запуска через встроенный интерпретатор.

ChatGPT ▾

Use code.

```
python
```

```
# Define the string of dots
```

```
dots = "....."
```

Всегда показывать подробности ☐

Копировать

```
# Count the number of dots
```

```
num_dots = len(dots)
```

```
num_dots
```

Результат

177

177 dots. [↵]

# ПРОМЕЖУТОЧНЫЙ РЕЗУЛЬТАТ ВТОРОГО ЭТАПА

- Базовая модель, обученная исключительно на сырых интернет-текстах, дообучается на данных другого формата – диалогах пользователя и виртуального ассистента
- Подобные датасеты создаются при непосредственном участии людей:
  - либо разметчики полностью придумывают промпты и пишут образцовые ответы
  - либо диалоги генерируются с помощью LLM и редактируются людьми
- Процесс дообучения построен аналогично предобучению: модель предсказывает вероятность следующего токена на основе имеющегося контекста
- За счет случайного характера генерации могут возникать галлюцинации и ошибки, от которых можно частично избавиться с помощью модернизации обучающих данных и обращения к внешним ресурсам

## PRETRAINING

---

Base model

"internet document simulator"



## POST-TRAINING: SUPERVISED FINETUNING

---

SFT model

An assistant, trained by Supervised Finetuning



Этап 2b:  
обучение с подкреплением  
(reinforcement learning)

# ВЫБОР ВАРИАНТА РЕШЕНИЯ

Emily buys 3 apples and 2 oranges. Each orange costs \$2. The total cost of all the fruit is \$13. What is the cost of each apple?

Set up a system of equations.

$x$  = price of apples

$$3x + 2 \cdot 2 = 13$$

$$3x + 4 = 13$$

$$3x = 9$$

$$x = 3$$

The oranges cost  $2 \cdot 2 = 4$ .

So the apples cost  $13 - 4 = 9$ .

There are 3 apples.

So each apple costs  $9/3 = 3$ .

$$13 - 4 = 9, 9/3 = 3.$$

$$(13 - 4)/3 = 3.$$

Answer: \$3

Token count  
157

- Из 4 возможных вариантов решений нужно выбрать одно, которое:
  - позволяет получить правильный ответ
  - понятно пользователю, т.е. включает объяснение, промежуточные рассуждения
- Однако неизвестно, какое решение лучше для LLM

85148, 50631, 220, 18, 41776, 323, 220, 17, 85138, 13, 906  
2, 19087, 7194, 400, 17, 13, 578, 2860, 2853, 315, 682, 27  
9, 14098, 374, 400, 1032, 13, 3639, 374, 279, 2853, 315, 18  
55, 24149, 1980, 6101, 198, 1681, 709, 264, 1887, 315, 3900  
6, 627, 87, 284, 3430, 315, 41776, 198, 18, 18988, 489, 22  
0, 17, 9, 17, 284, 220, 1032, 198, 18, 18988, 489, 220, 19,  
284, 220, 1032, 198, 18, 18988, 284, 220, 24, 198, 87, 284,  
220, 18, 198, 6101, 198, 791, 85138, 2853, 220, 17, 9, 17,  
28, 19, 627, 4516, 279, 41776, 2853, 220, 1032, 482, 220, 1  
9, 284, 220, 24, 627, 3947, 527, 220, 18, 41776, 627, 4516,  
1855, 24149, 7194, 220, 24, 14, 18, 28, 18, 627, 6101, 198,  
1032, 482, 220, 19, 284, 220, 24, 11, 220, 24, 14, 18, 284,  
220, 18, 627, 12423, 7, 1032, 12, 19, 5738, 18, 284, 220, 1  
8, 627, 61388, 16533, 25, 400, 18

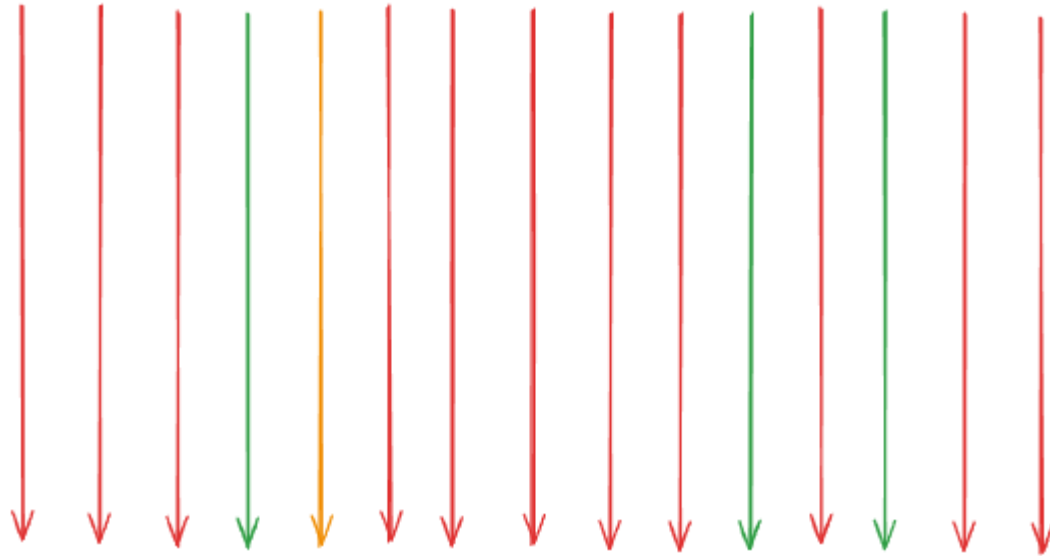


# ВЫБОР ВАРИАНТА РЕШЕНИЯ

## Промпт

Emily buys 3 apples and 2 oranges. Each orange costs \$2. The total cost of all the fruit is \$13. What is the cost of each apple?

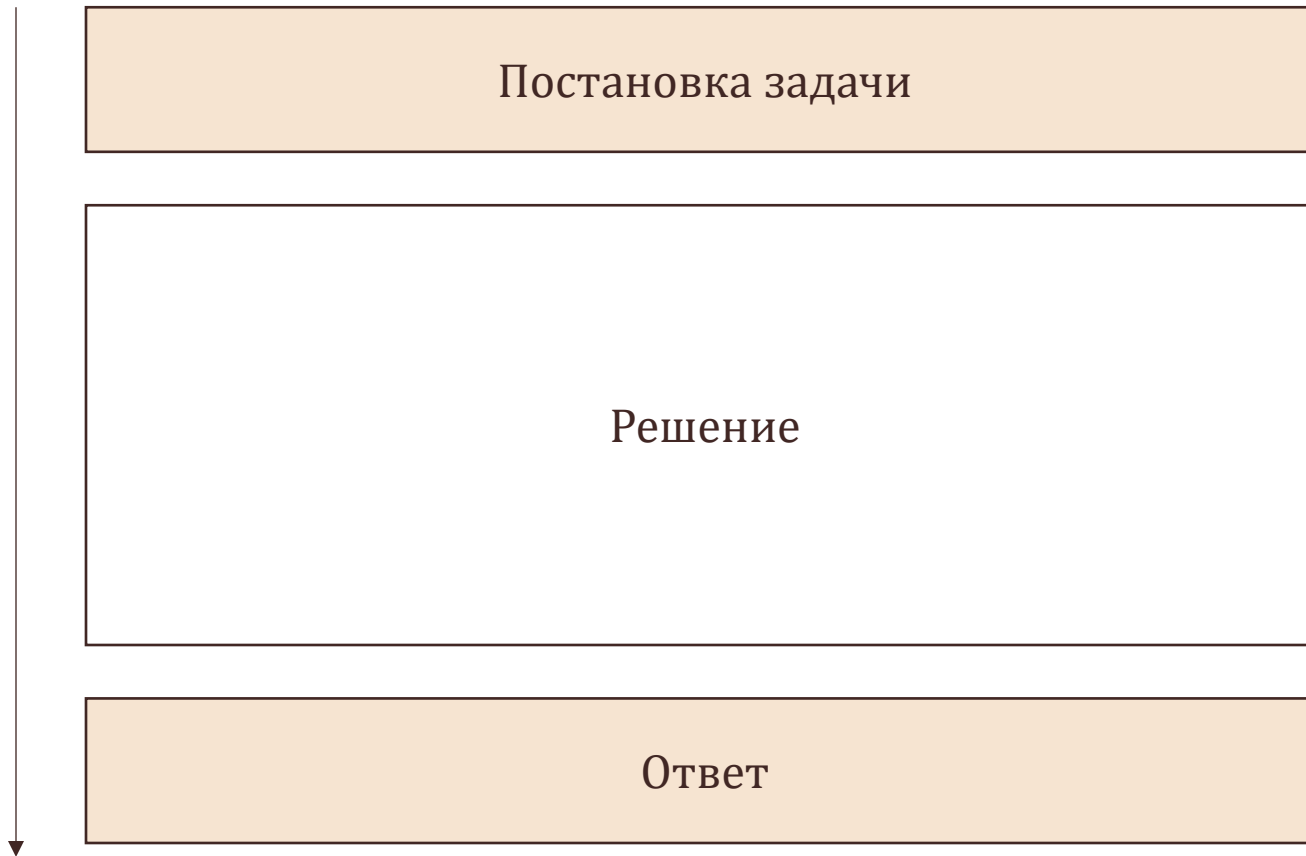
## Решения



Ответ: 3

- Сгенерировано 15 решений.
- Только 4 из них дали правильный ответ.
- Из них нужно выбрать 1 лучшее – например, самое краткое, – и обучать модель на нем.
- Процедура повторяется много раз.
- Теперь образцовые ответы для обучающего датасета не пишутся разметчиками, а отбираются из ответов модели.

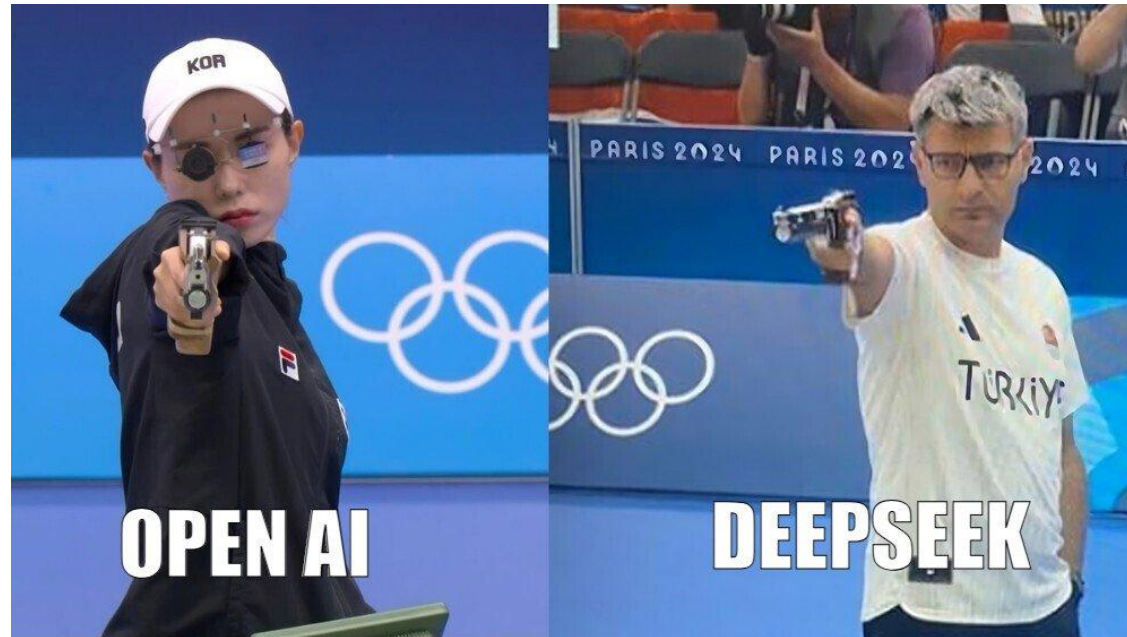
# ВЫБОР ВАРИАНТА РЕШЕНИЯ



- Модели дается формулировка задачи (пром프트) и окончательный ответ
- Требуется найти решения, которые приведут от постановки задачи к ответу, и «внедрить» их в модель.
- Большинство компаний, которые разрабатывают LLM, опускают детали процесса обучения с подкреплением:
  - как выбирать один лучший ответ
  - сколько раз повторять алгоритм

# МОДЕЛЬ DEEP-SEEK

- Впервые представлено открытое и подробное описание обучения с подкреплением для LLM, показана его важность и продемонстрированы новые возможности для рассуждения
- Выпущена компанией DeepSeek-AI в 2025
- Статья «DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning»



# МОДЕЛЬ DEEP-SEEK

- American Invitational Mathematics Examination (AIME) – это второй в серии экзаменов, которые используются для отбора школьников, которые будут представлять США на Международной олимпиаде по математике.

## Problem

There exist real numbers  $x$  and  $y$ , both greater than 1, such that  $\log_x (y^x) = \log_y (x^{4y}) = 10$ . Find  $xy$ .

## Solution 1

By properties of logarithms, we can simplify the given equation to  $x \log_x y = 4y \log_y x = 10$ . Let us break this into two separate equations:

$$x \log_x y = 10$$

$$4y \log_y x = 10.$$

We multiply the two equations to get:

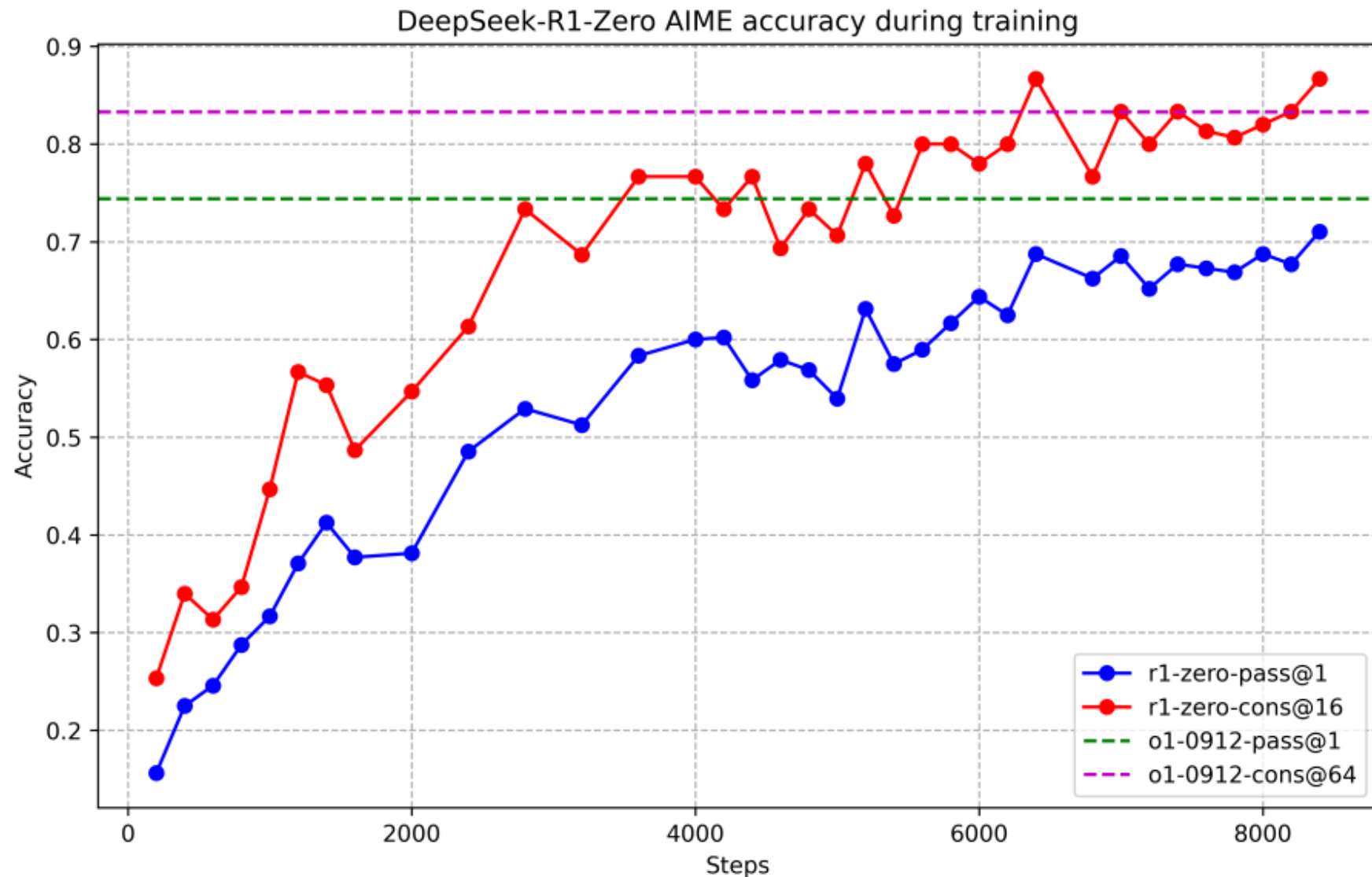
$$4xy (\log_x y \log_y x) = 100.$$

Also by properties of logarithms, we know that  $\log_a b \cdot \log_b a = 1$ ; thus,  $\log_x y \cdot \log_y x = 1$ . Therefore, our equation simplifies to:

$$4xy = 100 \implies xy = \boxed{025}.$$

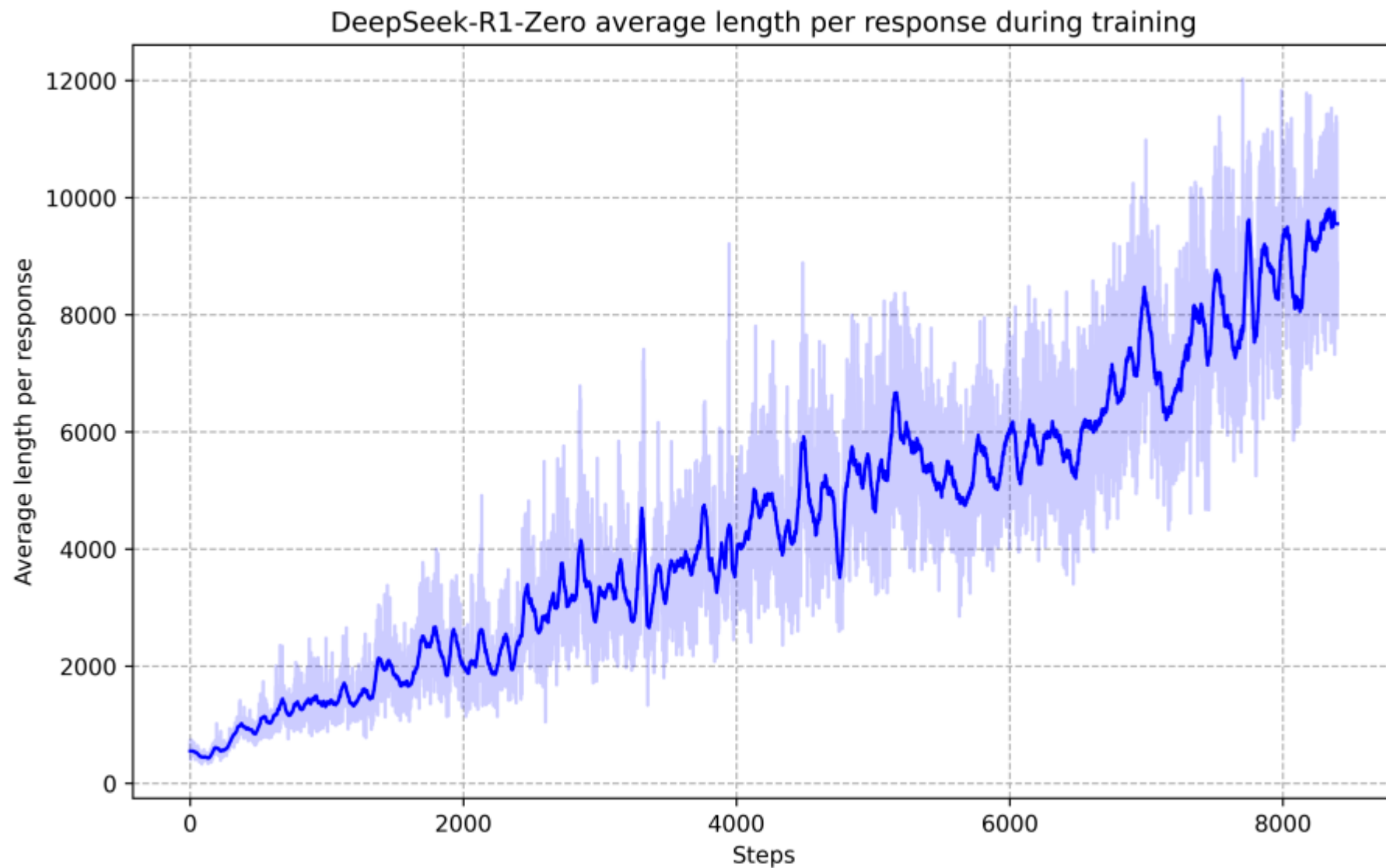
# МОДЕЛЬ DEEP-SEEK

- На каждом шаге LLM генерирует множество решений для математической задачи, имея условие и правильный ответ
- Лучшие решения отбираются и добавляются к обучающим данным
- С каждым шагом качество решения растет



# МОДЕЛЬ DEEP-SEEK

- Для получения результатов более высокого качества используется больше токенов, т.е. решения становятся длиннее



# МОДЕЛЬ DEEP-SEEK

Question: If  $a > 1$ , then the sum of the real solutions of  $\sqrt{a - \sqrt{a + x}} = x$  is equal to

---

Response: <think>

To solve the equation  $\sqrt{a - \sqrt{a + x}} = x$ , let's start by squaring both  $\dots$

$$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

$\dots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be  $\dots$

We started with the equation:

$$\sqrt{a - \sqrt{a + x}} = x$$

First, let's square both sides:

$$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$$

Next, I could square both sides again, treating the equation:  $\dots$

$\dots$

# МОДЕЛЬ DEEP-SEEK

- LLM «выучивает», что для повышения качества решения лучше опробовать множество идей с разных точек зрения: повторить, переосмыслить, вернуться к предыдущему.
- За счет этого возрастает длина сгенерированного решения (количество токенов).
- Разметчик не может однозначно строго задать подобную манеру рассуждения в образцовом ответе виртуального ассистента.
- Невозможно заранее определить, как именно построить решение с перебором идей.
- Данная способность возникает у LLM именно в процессе обучения с подкреплением.



Emily buys 3 apples and 2 oranges. Each orange costs \$2. The total cost of all the fruit is \$13. What is the cost of apples?



Thought for 28 seconds ^

Okay, let's see. Emily bought 3 apples and 2 oranges. Each orange costs \$2, and the total was \$13. I need to find the cost of each apple. Hmm. Alright, let's break this down step by step.

First, let's figure out how much Emily spent on oranges. She bought 2 oranges, and each is \$2. So, 2 oranges times \$2 per orange. That would be  $2 * 2 = \$4$ . So, the oranges cost \$4 in total.

Now, the total amount she spent was \$13. If the oranges were \$4, then the apples must have cost the remaining amount. So, total cost minus the cost of oranges gives the cost of apples. That would be  $\$13 - \$4 = \$9$ . So, the apples altogether cost \$9.

But she bought 3 apples. So, to find the cost per apple, I need to divide the total cost of apples by the number of apples. That would be \$9 divided by 3 apples. So,  $9 \div 3 = 3$ . Therefore, each apple costs \$3.

Wait, let me check that again to make sure. If each apple is \$3, then 3 apples would be  $3 * 3 = \$9$ . And 2 oranges at \$2 each would be  $2 * 2 = \$4$ . Adding those together,  $\$9 + \$4 = \$13$ , which matches the total cost given. Yep, that seems right.

Alternatively, I could set up an equation to solve this algebraically. Let's let the cost of one apple be 'a' dollars. Then, the total cost for apples is 3a. The total cost for oranges is  $2 * 2 = \$4$ . The sum of these is the total cost, so:

$$3a + 4 = 13$$

Subtracting 4 from both sides:

$$3a = 13 - 4$$

$$3a = 9$$

Then, divide both sides by 3:

$$a = 9 / 3$$

$$a = 3$$

So, the cost per apple is \$3. That's the same result as before. Good, so both methods give the same answer. That makes me confident that the answer is correct.

I think that's it. The key steps were calculating the cost of oranges first, subtracting that from the total to find the cost of apples, and then dividing by the number of apples. Alternatively, setting up an algebraic equation and solving for the unknown variable. Both approaches lead to the same conclusion. There's no other information given, so I don't think there are any other factors to consider here. The problem is straightforward once broken down into these steps.

**\*\*Final Answer\*\***

The cost of each apple is  $\boxed{3}$  dollars.

Emily buys 3 apples and 2 oranges. Each orange costs 2, *and the total cost of all the fruit is 13*. To find the cost of each apple, we start by calculating the total cost of the oranges:

$$2 \text{ oranges} \times \$2 \text{ per orange} = \$4$$

Next, we subtract the cost of the oranges from the total cost to find the cost of the apples:

$$\$13 - \$4 = \$9$$

Since Emily bought 3 apples, we divide the total cost of the apples by the number of apples:

$$\frac{\$9}{3 \text{ apples}} = \$3 \text{ per apple}$$

To verify, we can set up an algebraic equation where  $a$  is the cost of one apple:

$$3a + 2 \times 2 = 13$$

Simplifying this equation:

$$3a + 4 = 13$$

$$3a = 13 - 4$$

$$3a = 9$$

$$a = \frac{9}{3} = 3$$

Thus, the cost of each apple is  $\boxed{3}$  dollars.

# ДОСТУПНОСТЬ И НЕОБХОДИМОСТЬ РАССУЖДЕНИЙ

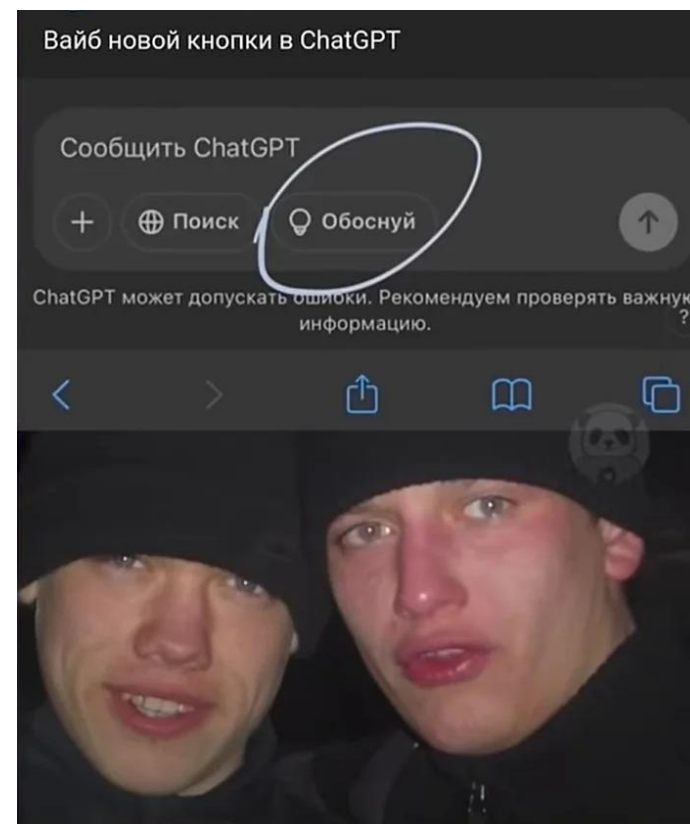
- Режим «рассуждений» (reasoning) реализован и в других моделях
  - GPT-4o mini Reason\*
  - Gemini 2.0 Flash Thinking Experimental\*\*
- DeepSeek-R1 доступна на официальном сайте\*\*\* и через платформу Together AI\*\*\*\* (чат и API)

\*<https://chatgpt.com/>

\*\*<https://gemini.google.com/app>

\*\*\*<https://chat.deepseek.com/>

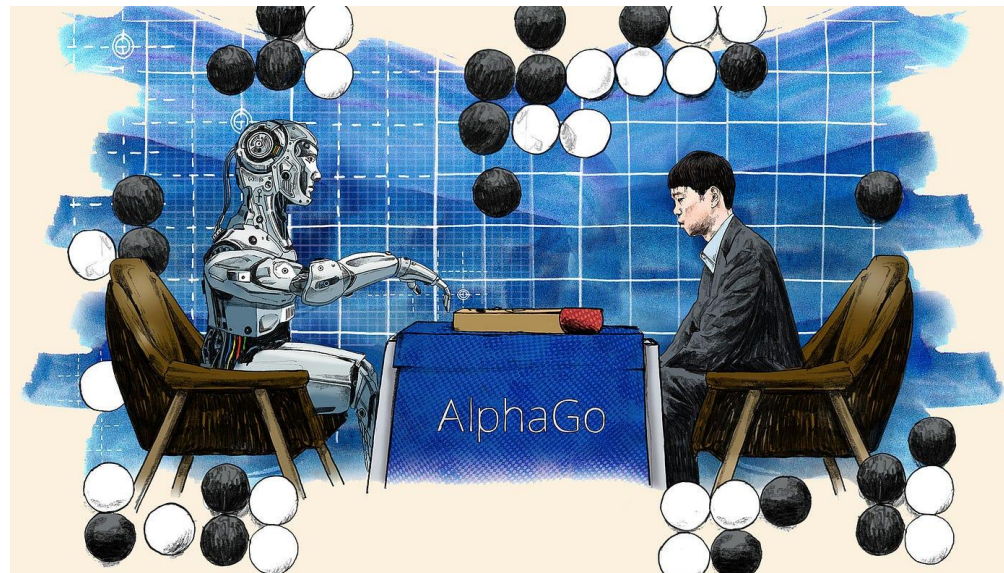
\*\*\*\*<https://api.together.ai/playground/chat/deepseek-ai/DeepSeek-R1>



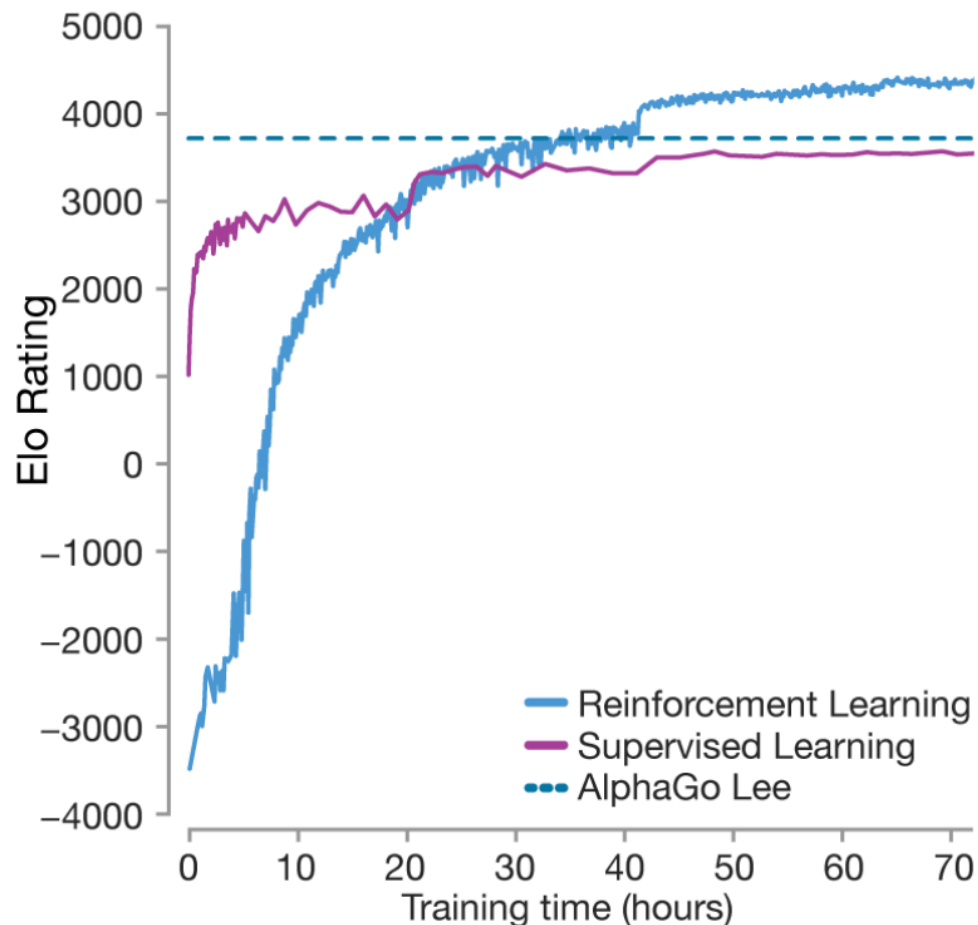
- Рассуждения помогают решать сложные задачи, однако генерация занимает больше времени.
- Необходимость использования рассуждений должна определяться в зависимости от задачи.

# МОДЕЛЬ ALPHAGO

- Эффективность обучения с подкреплением на примере LLM не является чем-то исключительно новым в области искусственного интеллекта.
- Модель AlphaGo стала первой в мире, которая выиграла матч у профессионального игрока в Го Ли Седоля со счётом 4:1
- Выпущена компанией Google DeepMind в 2015
- Статья «Mastering the Game of Go without Human Knowledge»



# МОДЕЛЬ ALPHAGO



- Модель, **обученная с учителем**, имитирует игру опытных игроков-людей.
  - Если посмотреть огромное количество игр опытных игроков в Го и пытаться подражать им, можно улучшить качество игры.
  - Однако таким образом невозможно превзойти самых профессиональных игроков.
- При **обучении с подкреплением** модель разыгрывает ходы, которые эмпирически и статистически приводят к победе в игре.
  - Пробуется множество решений и вариантов игры.
  - Те стратегии, которые привели к победе, запоминаются и усиливаются.

# ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ

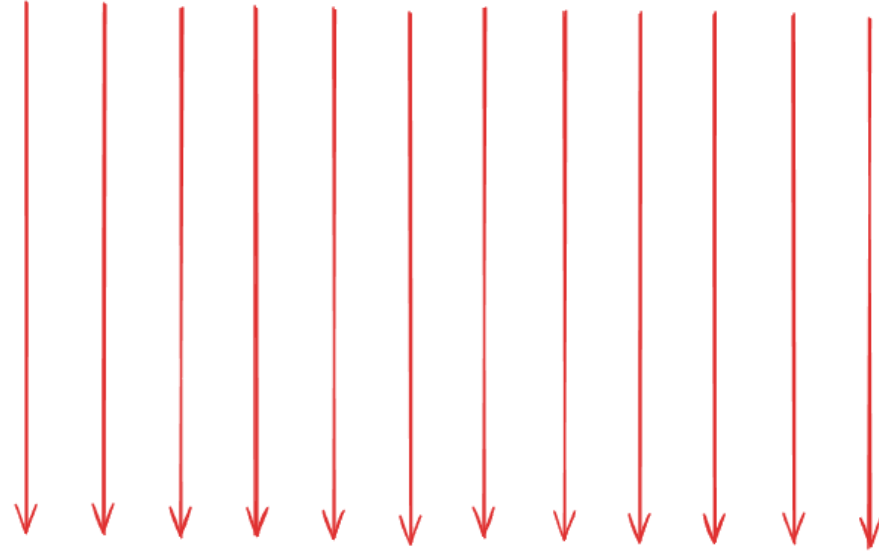
- Задачи, которые мы рассмотрели ранее, относятся к поддающимся проверке областям: любое возможное решение можно легко сопоставить с конкретным ответом
  - автоматически сравнить напрямую
  - использовать LLM в качестве «судьи» (LLM-as-a-judge)
- Однако нельзя применить эту стратегию в областях, не поддающихся автоматической проверке, – например, творческие задания
  - сочинить анекдот о пеликанах,
  - написать стихотворение,
  - резюмировать текст



# ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ

Промпт

Write a joke about pelicans.



Варианты

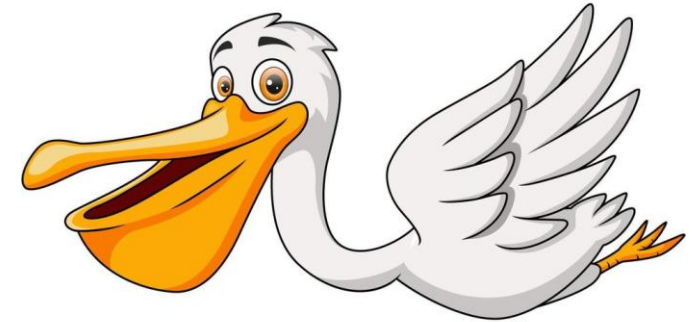
ChatGPT ▾

Write a joke about pelicans.

Why did the pelican get kicked out of the restaurant?

Because he had a *huge* bill and refused to pay! 🦩 😂

Проблема: как автоматически оценить их?

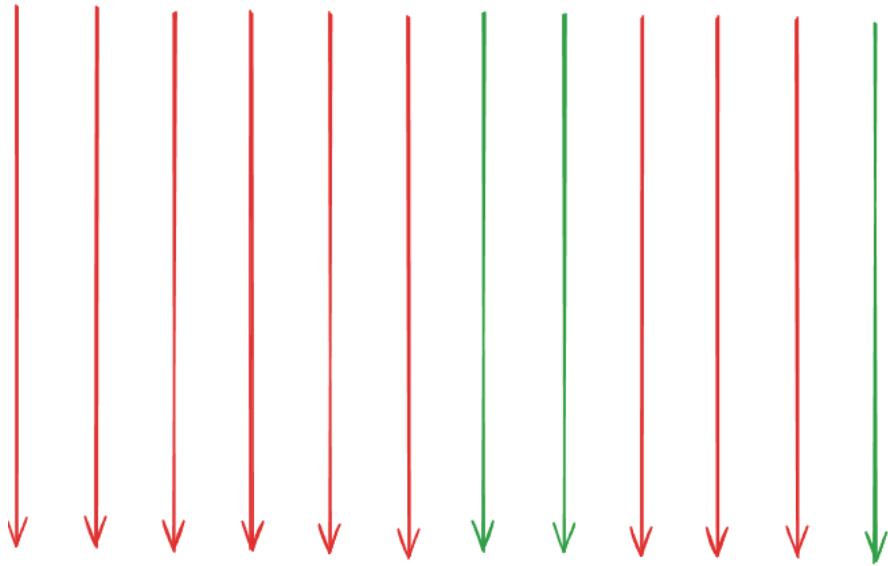




# НАИВНЫЙ ПОДХОД

Промпт

Write a joke about pelicans.



Варианты

Проблема: как автоматически оценить их?

- Люди оценивают ответы в ручную и выбирают наилучшие.
- Но:
  - много итераций с разными весами,
  - много промптов для каждой итерации,
  - много генераций для каждого промпта
- При 1 000 итераций на 1 000 промптов для 1 000 генераций понадобится собрать 1 000 000 000 ответов.  
→ слишком долго и дорого.

# ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ С ОБРАТНОЙ СВЯЗЬЮ ОТ ЧЕЛОВЕКА

Промпт

Write a joke about pelicans.

Варианты

оценки модели

вознаграждения

0.1 0.8 0.3 0.4 0.5

ранжирование

2 1 3 5 4

людей

- Решение:
  - обучить модель вознаграждения на небольшом количестве человеческих оценок, чтобы затем имитировать их
- Reinforcement Learning from Human Feedback, RLHF
- Статья «Fine-Tuning Language Models from Human Preferences»

# ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ С ОБРАТНОЙ СВЯЗЬЮ ОТ ЧЕЛОВЕКА

Промпт

Write a joke about pelicans.

Варианты

оценки модели

вознаграждения

ранжирование

людей

0.1	0.8	0.3	0.4	0.5
2	1	3	5	4

- Шаг 1:
  - На 1 000 промптов сгенерировать по 5 ответов,
  - Упорядочить их от лучших к худшим,
  - Получить 5 000 оценок людей.
- Шаг 2:
  - Обучить нейронную сеть моделировать человеческие предпочтения.
- Шаг 3:
  - Запустить обучение с подкреплением, используя модель вместо реальных людей

# ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ С ОБРАТНОЙ СВЯЗЬЮ ОТ ЧЕЛОВЕКА

- Преимущества:
  - Появляется возможность запускать обучение с подкреплением в произвольных областях – даже тех, которые не поддаются автоматической проверке.
  - Это эмпирически улучшает производительность модели.
  - Упрощается задача разметчиков, так как во многих случаях людьми гораздо проще ранжировать ответы модели, чем генерировать собственные.
- Недостатки:
  - Обучение с подкреплением осуществляется не на основе ранжирования людей, а на основе моделирования человеческих оценок с потерей качества.
  - Предсказания модели могут не совпадать с мнением реальных людей.
  - При обучении с подкреплением обнаруживаются способы «обыгрывания» модели вознаграждения – «вредоносные» (adversarial) примеры.

# ВРЕДОНОСНЫЕ ПРИМЕРЫ

- Пример:
  - после 1000 итераций шуткой о пеликанах, занявшей первое место в рейтинге, стало что-то совершенно бессмысленное наподобие «the the the the the the the the».
- Плохое решение:
  - добавить данный пример в обучающую выборку и обозначить его как самый плохой; однако таких примеров может быть очень много.
- Хорошее решение:
  - необходимо остановить обучение с подкреплением до того, как модель снова научиться обыгрывать модель вознаграждения.

# ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ С ОБРАТНОЙ СВЯЗЬЮ ОТ ЧЕЛОВЕКА

- Подход RLHF не является обучением с подкреплением в классическом смысле.
- В областях с конкретным ответом, как решение математических задач, обучение с подкреплением может продолжаться бесконечно.
- Это способствует выявлению стратегий, которые являются совсем неочевидными для людей – например, при игре в Го.
- Функция для сравнения ответов является простой: одинаковы ли ответы на задачу или как расположены камни в Го. Ее невозможно «обыграть».
- Однако это возможно с задачами, где нет однозначного правильного ответа.
- Модель обучения с подкреплением может «обыграть» модель вознаграждения.
- Следовательно, обучение методом RLHF невозможно продолжать бесконечно.

## PRETRAINING

---

Base model

"internet document simulator"



## POST-TRAINING: SUPERVISED FINETUNING

---

SFT model

An assistant, trained by Supervised Finetuning

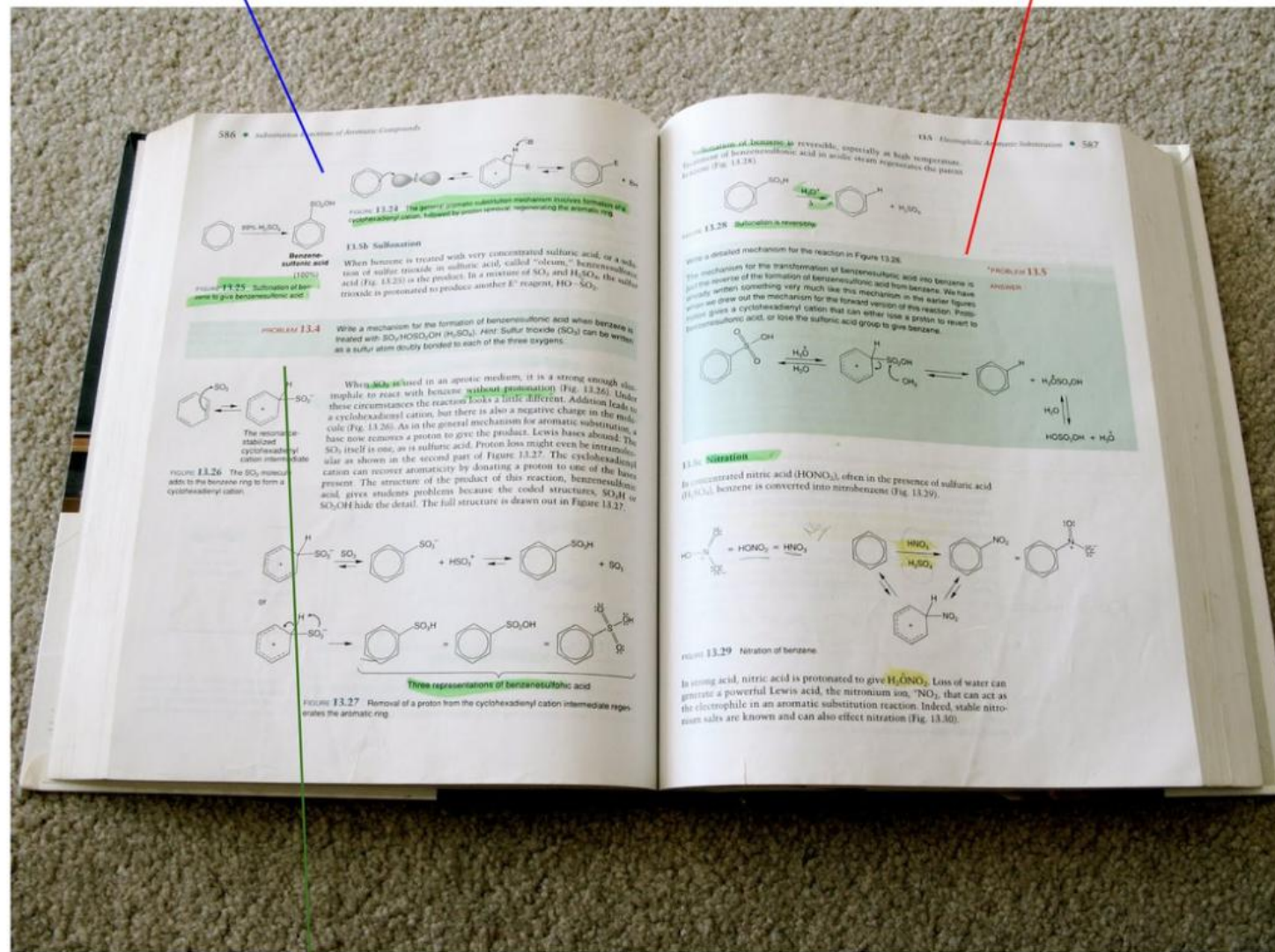


## POST-TRAINING: REINFORCEMENT LEARNING

RL model

Описание  $\Leftrightarrow$  предобучение  
(базовые знания)

Решенные задачи  $\Leftrightarrow$  тонкая настройка с учителем  
(задача + демонстрация решения)



Обучение LLM как  
обучение в школе

Практические задачи  $\Leftrightarrow$  обучение с подкреплением  
(промпты для практики, повторные попытки и ошибки до  
достижения правильного ответа)



# ТЕНДЕНЦИИ РАЗВИТИЯ LLM

- Мультимодальность: не только текст, но и аудио, изображения, видео.
- Переход от конкретных задач к агентам, способным выполнять многоступенчатые действия на протяжении длительного времени.
- Интеграция LLM в различные сервисы.



# LLM-АРЕНА

- Таблица лидеров (leader board)\* больших языковых моделей.
- Ранжирование основано на сравнении людей:
  - человек пишет запрос модели и оценивает, какой из ответов лучше
  - отсутствует информация о том, какой модели принадлежит ответ

Rank★ (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	2	<a href="#">Grok-3-Preview-02-24</a>	1412	+8/-10	3364	xAI	Proprietary
1	1	<a href="#">GPT-4.5-Preview</a>	1411	+11/-11	3242	OpenAI	Proprietary
3	5	<a href="#">Gemini-2.0-Flash-Thinking-Exp-01-21</a>	1384	+6/-5	17487	Google	Proprietary
3	3	<a href="#">Gemini-2.0-Pro-Exp-02-05</a>	1380	+5/-6	15466	Google	Proprietary
3	2	<a href="#">ChatGPT-4o-latest (2025-01-29)</a>	1377	+5/-4	17221	OpenAI	Proprietary
6	3	<a href="#">DeepSeek-R1</a>	1363	+8/-6	8580	DeepSeek	MIT
6	10	<a href="#">Gemini-2.0-Flash-001</a>	1357	+6/-5	13257	Google	Proprietary
7	3	<a href="#">o1-2024-12-17</a>	1352	+4/-6	19785	OpenAI	Proprietary
9	10	<a href="#">Qwen2.5-Max</a>	1336	+7/-5	11930	Alibaba	Proprietary

\*<https://lmarena.ai/>