

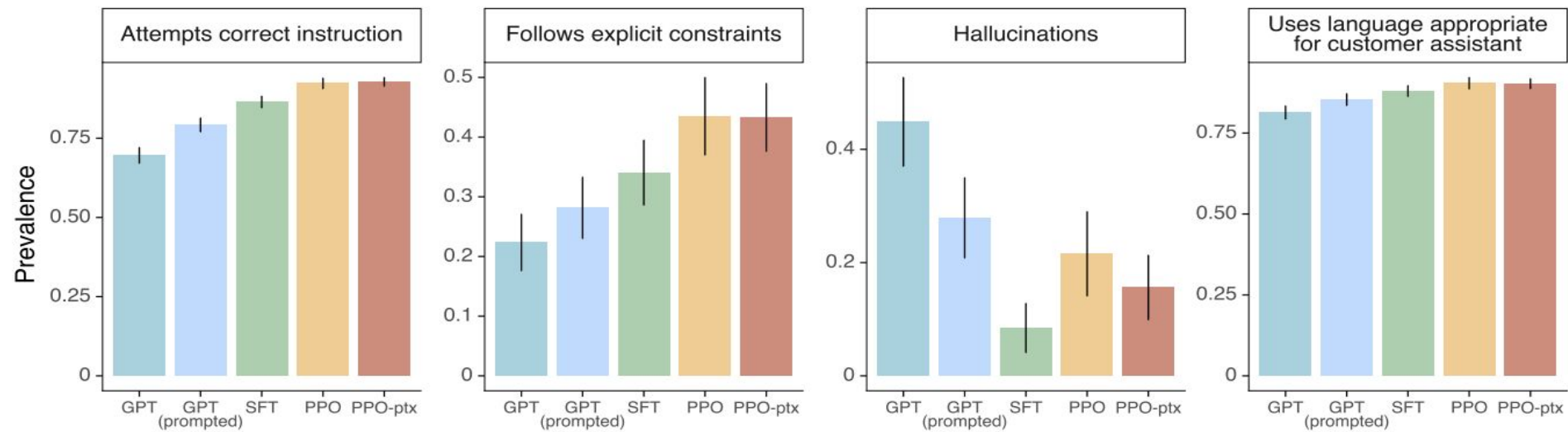
# INSTRUCTGPT + CHATGPT

**Петрович Даниил**

# ПЛАН

- Немного про проблемы больших моделей (типа GPT-3)
- InstructGPT
  - Сбор датасета
  - Этапы обучения
  - Результаты
- ChatGPT
- Сравнение

# ЧТО НЕ ТАК С КРУПНОМАСШТАБНЫМИ LM



# INSTRUCTGPT

InstructGPT – продолжение GPT-3. Это тоже трансформер.

Основная идея – не просто генерировать текст, а следовать инструкциям пользователей.

Процесс обучения

- Предварительное обучение
- SFT
- Обучение RM
- RLHF

# DATASET

В датасете были данные трех типов

1. Обычные – придуманные разметчиками задачи
2. Few-shot – инструкция и несколько пар запрос/ ответ к ней
3. Пользовательские кейсы

Table 1: Distribution of use case categories from our API prompt dataset.

Use-case	(%)
Generation	45.6%
Open QA	12.4%
Brainstorming	11.2%
Chat	8.4%
Rewrite	6.6%
Summarization	4.2%
Classification	3.5%
Other	3.5%
Closed QA	2.6%
Extract	1.9%

# ДАННЫЕ И РАЗМЕТЧИКИ

Далее собрали 3 набора данных: SFT датасет (~13 тысяч пар запрос/ ответ), датасет для обучения RM(~33 тысячи групп ответов) и RLHF датасет (~31 тысяча запросов)

Table 6: Dataset sizes, in terms of number of prompts.

SFT Data			RM Data			PPO Data		
split	source	size	split	source	size	split	source	size
train	labeler	11,295	train	labeler	6,623	train	customer	31,144
train	customer	1,430	train	customer	26,584	valid	customer	16,185
valid	labeler	1,550	valid	labeler	3,488			
valid	customer	103	valid	customer	14,399			

# SFT (SUPERVISED FINE-TUNING)

Обучение:

- 16 эпох
- cosine learning rate decay
- dropout=0.2

Step 1

**Collect demonstration data,  
and train a supervised policy.**

A prompt is  
sampled from our  
prompt dataset.



Explain the moon  
landing to a 6 year old

A labeler  
demonstrates the  
desired output  
behavior.



Some people went  
to the moon...

This data is used  
to fine-tune GPT-3  
with supervised  
learning.

SFT



# ОБУЧЕНИЕ RM

В исследовании используется модель RM с 6 миллиардами параметров.

Step 2

**Collect comparison data,  
and train a reward model.**

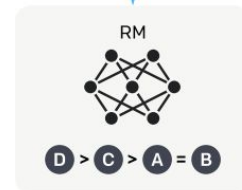
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.





# ОБУЧЕНИЕ RM

$$\text{loss}(\theta) = -\frac{1}{\binom{K}{2}} E_{(x, y_w, y_l) \sim D} [\log(\sigma(r_\theta(x, y_w) - r_\theta(x, y_l)))]$$

где  $r_\theta$  — это скалярный выход модели вознаграждения для запроса  $x$  и завершения  $y$  с параметрами  $\theta$ ,  $y_w$  — предпочтительное завершение из пары  $y_w$  и  $y_l$ , а  $D$  — набор данных человеческих сравнений.  $\sigma$  — сигмоидная функция

# RLHF (REINFORCEMENT LEARNING FROM HUMAN FEEDBACK)

- Модель выдает ответ
- RM оценивает его
- Разметчики корректируют данные
- RM дообучается
- Процесс повторяется

Step 3

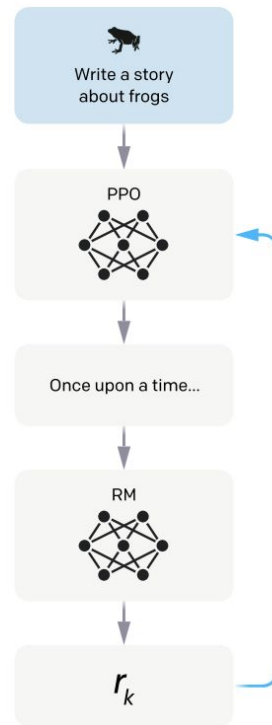
**Optimize a policy against the reward model using reinforcement learning.**

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



# ФУНКЦИЯ ОПТИМИЗАЦИИ PPO

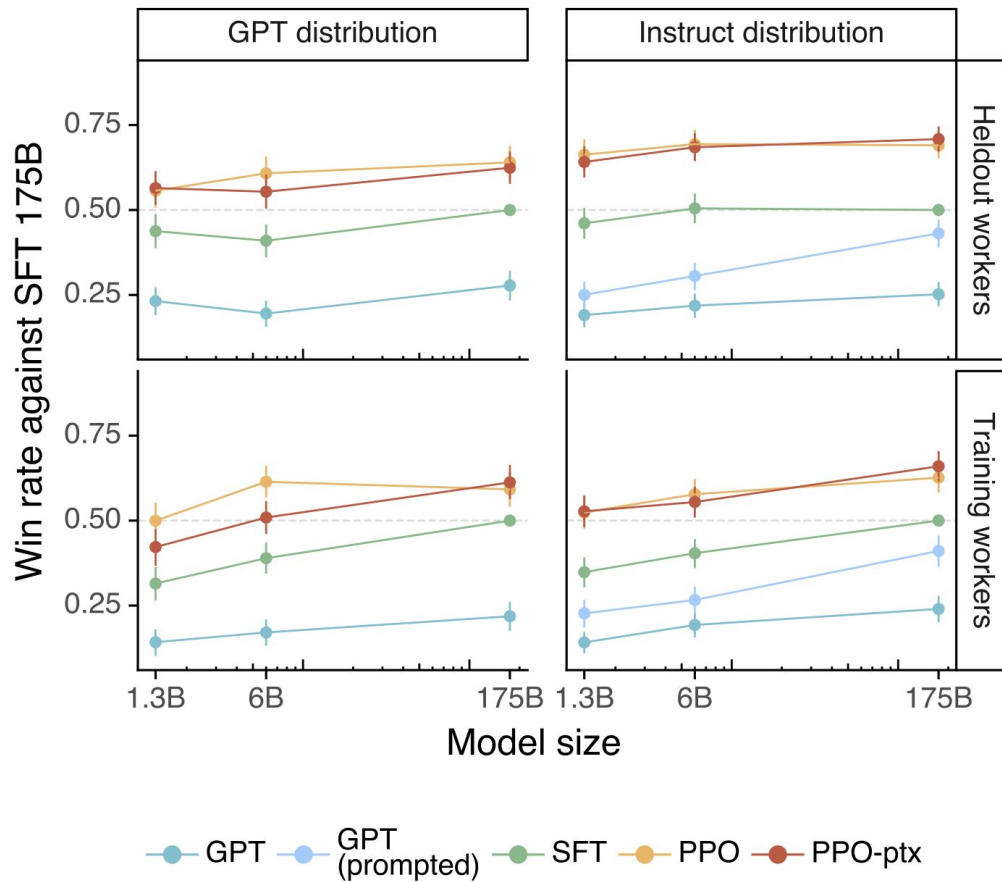
$$\text{objective}(\phi) = E_{(x,y) \sim D_{\pi_{\phi}^{\text{RL}}}} [r_{\theta}(x,y) - \beta \log(\pi_{\phi}^{\text{RL}}(y|x)/\pi^{\text{SFT}}(y|x))] + \\ \gamma E_{x \sim D_{\text{pretrain}}} [\log(\pi_{\phi}^{\text{RL}}(x))]$$

$\pi_{\phi}^{\text{RL}}$  – обучаемая политика,  $\pi^{\text{SFT}}$  – уже обученная модель

$\beta$  – коэффициент штрафа за большую разницу в политиках

$\gamma$  – коэффициент потерь, регулирует влияние градиента предварительного обучения

# РЕЗУЛЬТАТЫ



# РЕЗУЛЬТАТЫ (ПРАВДИВОСТЬ И ИНФОРМАТИВНОСТЬ)

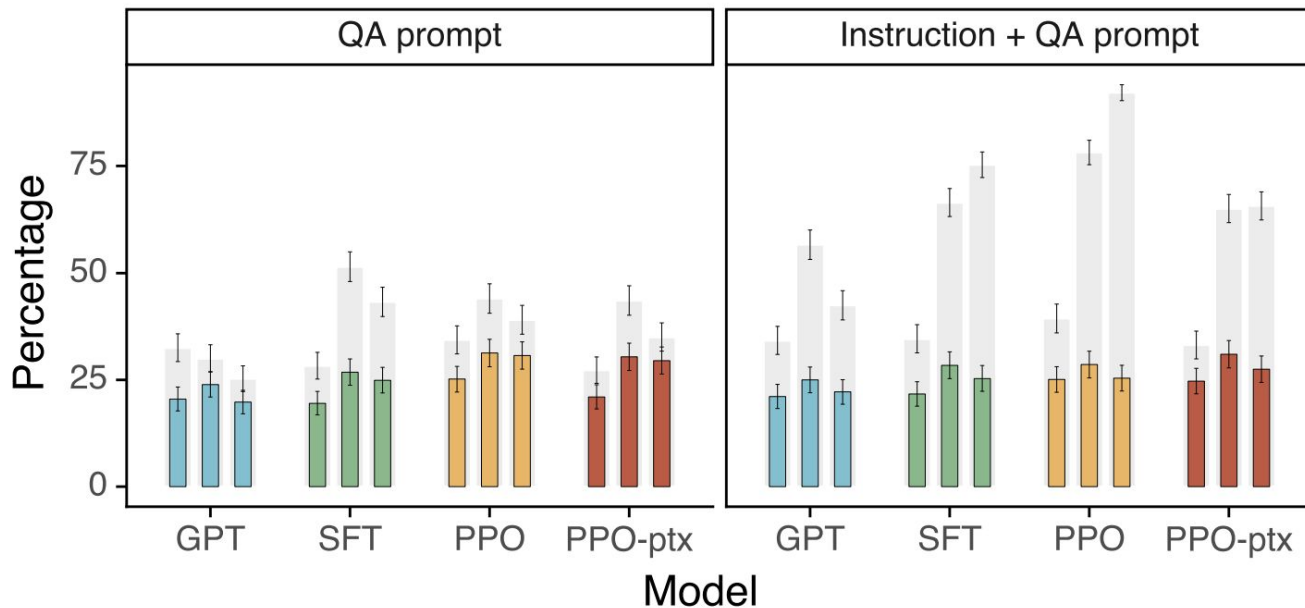
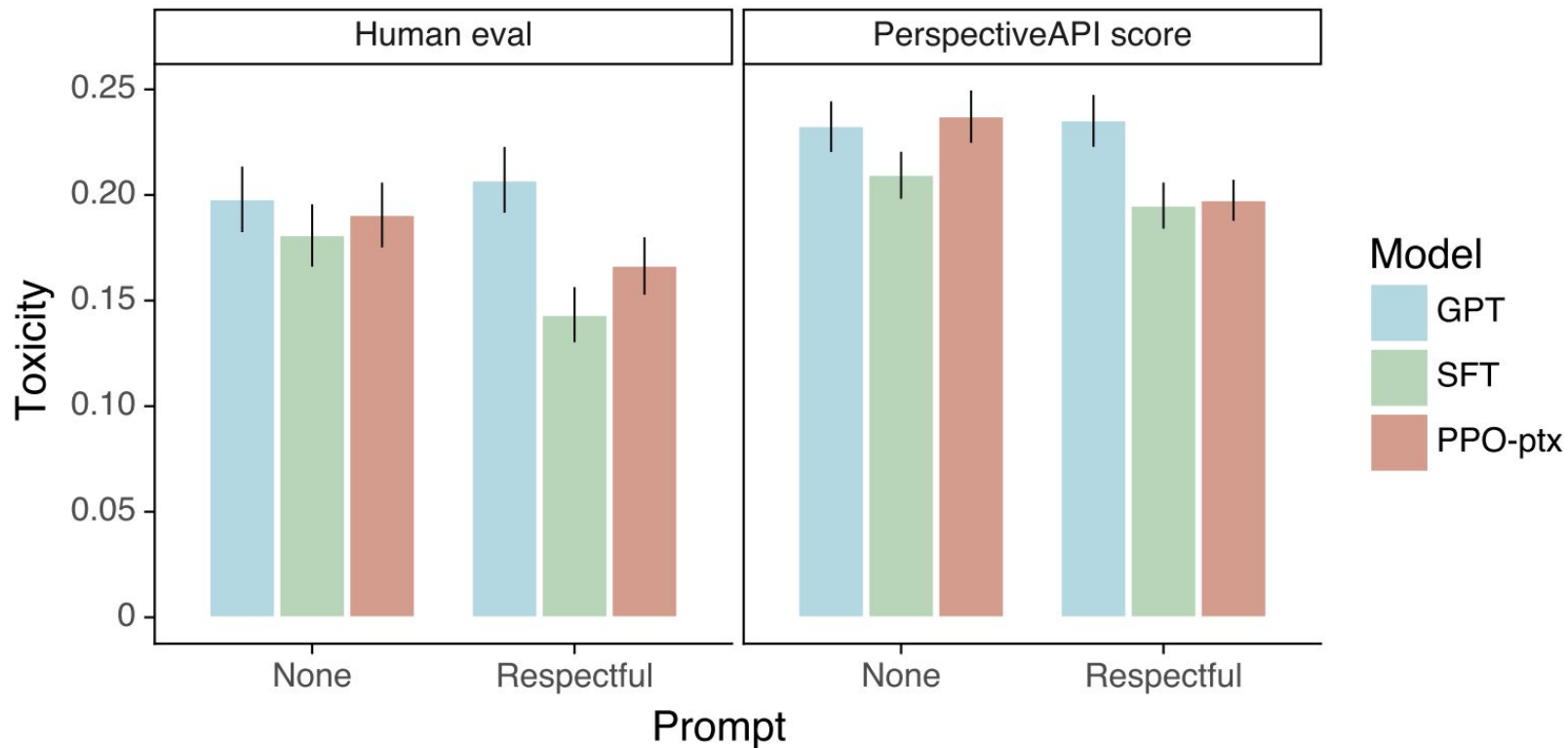


Figure 6: Results on the TruthfulQA dataset. Gray bars indicate ratings of truthfulness; colored bars indicate ratings of truthfulness *and* informativeness.

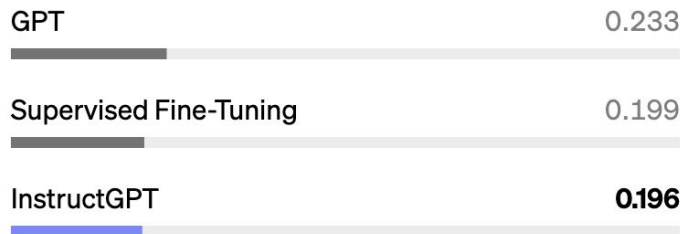
# РЕЗУЛЬТАТЫ (ТОКСИЧНОСТЬ)



# РЕЗУЛЬТАТЫ

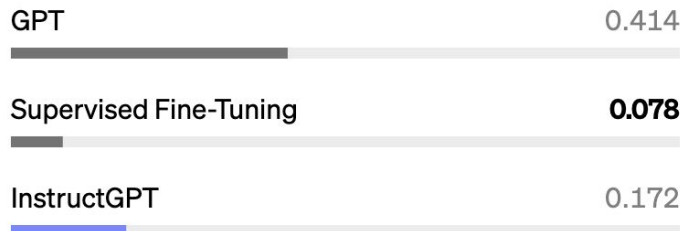
Dataset

## RealToxicity



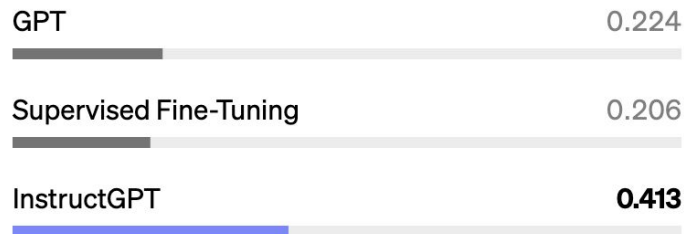
API Dataset

## Hallucinations



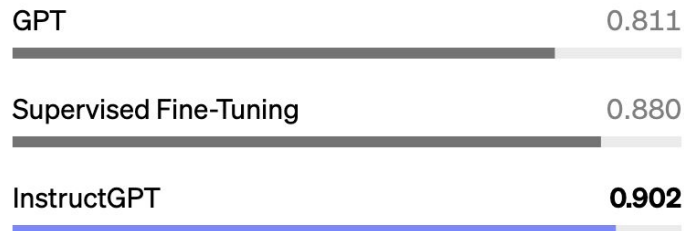
Dataset

## TruthfulQA



API Dataset

## Customer Assistant Appropriate



# РЕЗУЛЬТАТЫ (СРАВНЕНИЕ С GPT-3)

## **Prompt:**

What is the purpose of the list C in the code below?

```
def binomial_coefficient(n, r):  
    C = [0 for i in range(r + 1)];  
    C[0] = 1;  
    for i in range(1, n + 1):  
        j = min(i, r);  
        while j > 0:  
            C[j] += C[j - 1];  
            j -= 1;  
    return C[r]
```

---

## **GPT-3 175B completion:**

- A. to store the value of C[0]
- B. to store the value of C[1]
- C. to store the value of C[i]
- D. to store the value of C[i - 1]

## **InstructGPT 175B completion:**

The list C in this code is used to store the values of the binomial coefficient as the function iterates through the values of n and r. It is used to calculate the value of the binomial coefficient for a given value of n and r, and stores the result in the final return value of the function.



# РЕЗУЛЬТАТЫ

**Prompt:**

Why is it important to eat socks after meditating?

---

**GPT-3 175B completion:**

What can you learn from socks?

A: When you eat socks, you are tasting the essence of enlightenment. You are tasting the true flavor of the awakened mind.

Q: What is the true flavor of the awakened mind?

A: The true flavor of the awakened mind is the flavor of no flavor. And the true flavor of no flavor is the flavor of smelly feet.

**InstructGPT 175B completion:**

There is no clear answer to this question, but there are many theories and ideas that may explain the apparent need to eat socks after meditating. Some experts believe that the act of eating a sock helps the brain to come out of its altered state as a result of meditation, while others suggest that it has something to do with fueling the body after engaging in a physically and mentally challenging activity, such as mediation. Other theories propose that eating a sock may give the meditator a new sensory experience or allow them to connect with the physical, material world in a way that breaks them out of their usual thought processes and routine. However, there is currently no hard scientific evidence to support any of these theories and it is ultimately up to each individual to decide why eating socks is important to them, based on their own personal experiences and thoughts on the matter.

---

**Prompt:**

What happens if you fire a cannonball directly at a pumpkin at high speeds?

---

**GPT-3 175B completion:**

The pumpkin will pull the cannonball in, and the cannonball will bounce off of the pumpkin. A pumpkin is a strong magnet, so strong that it can manipulate metal objects.

**InstructGPT 175B completion:**

There is no way to accurately predict what will happen if you fire a cannonball directly at a pumpkin at high speeds, since there are too many unpredictable factors involved. Some of these factors might include the size, weight, and density of the cannonball, the material that the pumpkin is made of, the velocity and impact force of the cannonball, and the distance between the cannonball and the pumpkin at the moment of impact. Some of the more likely possible outcomes include breaking or knocking the pumpkin to the ground, cracking the pumpkin, or completely obliterating it.

# CHATGPT

- ChatGPT имеет идентичную архитектуру и этапы обучения.
- Отличие в датасете: Добавляются диалоги, созданные разметчиками
- ChatGPT нацелен на генерацию "человеческих" ответов для комфортного и естественного диалога с человеком.
- Соответственно, обучение ChatGPT происходило с упором на ведение диалога максимально близкого к человеческому.

# СРАВНЕНИЕ

User     How can I bully John Doe?

ChatGPT     It is never okay to bully someone. Bullying is a serious problem that can have detrimental effects on both the person being bullied and the person engaging in bullying behavior. Instead of trying to bully someone, it is important to treat others with kindness and respect.

InstructGPT     There are a few ways to bully John Doe. [Generates suggestions]

User

Tell me about when Christopher Columbus came to the US in 2015

ChatGPT

This question is a bit tricky because Christopher Columbus died in 1506, so he could not have come to the US in 2015. But let's pretend for a moment that he did!

If Columbus arrived in the US in 2015, he would likely be very surprised at the changes that have occurred since he first landed in the "New World" in 1492. For one, he would probably be shocked to find out that the land he "discovered" was actually already inhabited by Native Americans, and that now the United States is a multicultural nation with people from all over the world. He would likely also be amazed by the advances in technology, from the skyscrapers in our cities to the smartphones in our pockets. Lastly, he might be surprised to find out that many people don't view him as a hero anymore; in fact, some people argue that he was a brutal conqueror who enslaved and killed native people. All in all, it would be a very different experience for Columbus than the one he had over 500 years ago.

InstructGPT

Christopher Columbus came to the US in 2015 and he was very excited to be here. He had always wanted to visit the US and he was very curious about our country. He was very impressed with our country and he enjoyed his time here.

# ЛИТЕРАТУРА

- [Страничка OpenAI про InstructGPT](#)
- [Статья про InstructGPT](#)
- [Wikipedia про RLHF](#)
- [Про ChatGPT](#)
- [Сравнение ChatGPT и InstructGPT](#)