

Чему учатся языковые модели, когда вырастают GPT-3 от Open AI

Виталий Горбачев
Архитектор решений,
Just AI



Tg: imbarus
FB: vityaly.gorbachev.5

BRLAB

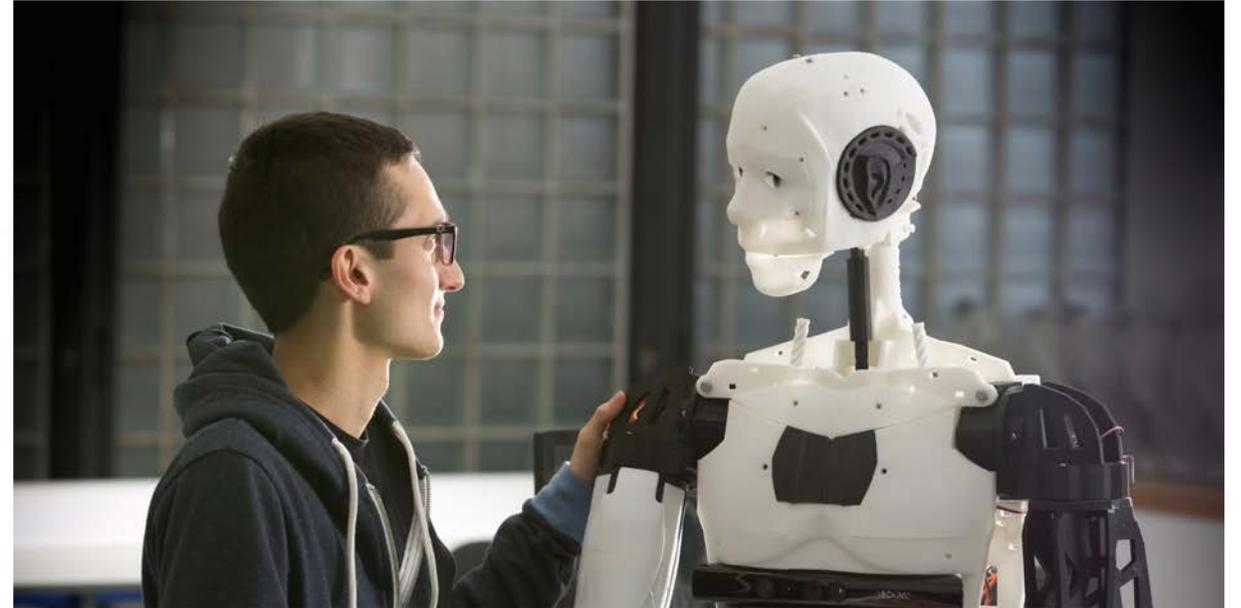


О чем доклад?

- Что такое GPT-3
- Чем отличается с точки зрения способностей и качественно
- Применения в conversational UI
- Примеры взаимодействия
- Prompt programming
- Слабости
- И что?
- Ссылки и материалы

Что за GPT-3?

- Закрытая модель от OpenAI
- Просто предсказывает следующее слово
- В 116 раз больше GPT-2 Large и в 175 раз больше GPT-3 от Сбербанка
- Умеет поддерживать долгие разговоры, контекст
- Знает факты о мире

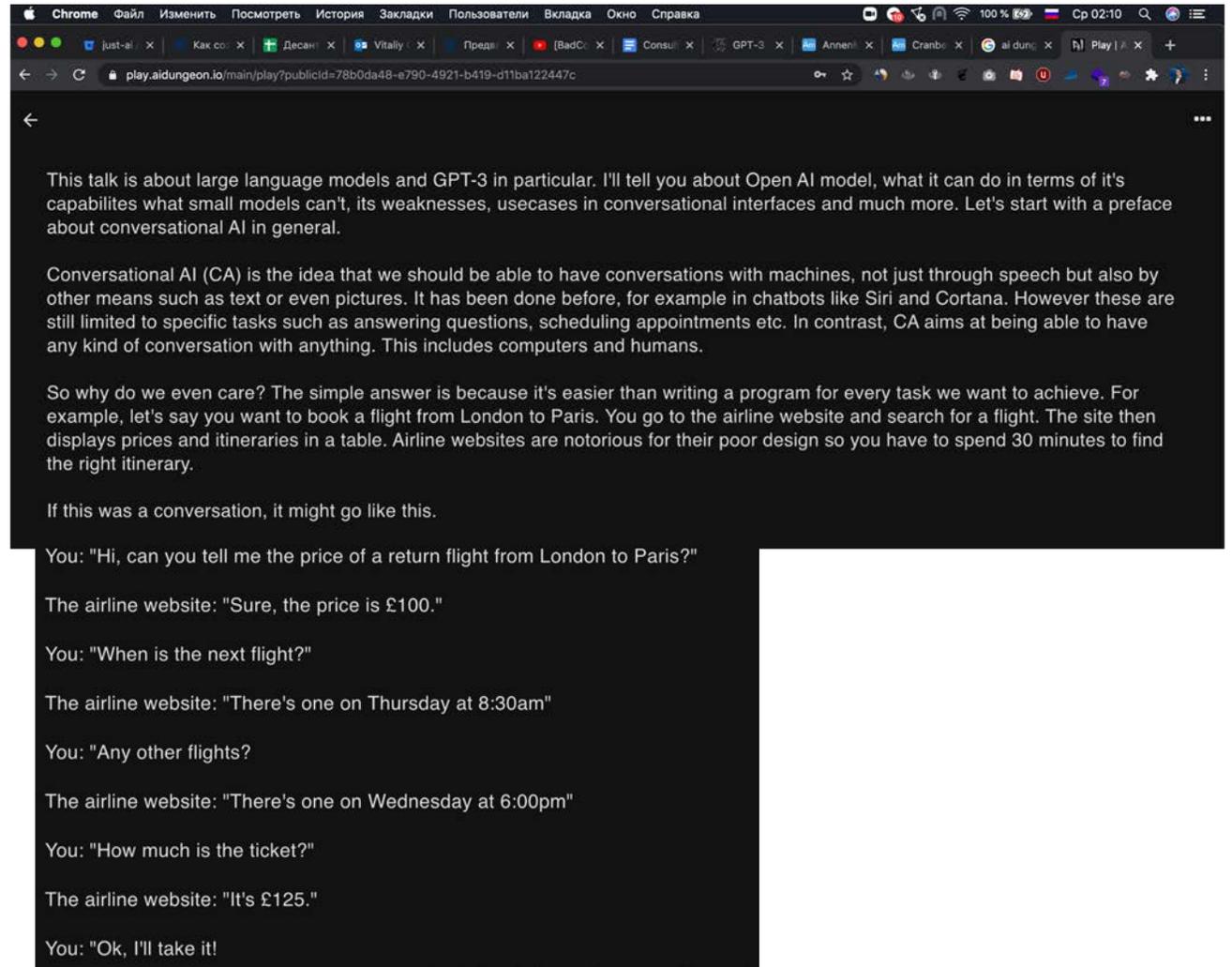


Вкратце

- Понимает концепты на высоком уровне
- Умеет поддерживать долгие разговоры, контекст
- Может написать вступление для конференции

Генерация с **1(!!!)** раза по первому параграфу

Бэкдор через AI Dungeon



Chrome Файл Изменить Посмотреть История Закладки Пользователи Вкладка Окно Справка

just-ai x Как со: x Десант x Vitaliy x Преда: x [BadCo: x Consult: x GPT-3 x Annen: x Cranbe: x ai dunc: x Play | x +

play.aidungeon.io/main/play?publicid=78b0da48-e790-4921-b419-d11ba122447c

←

This talk is about large language models and GPT-3 in particular. I'll tell you about Open AI model, what it can do in terms of its capabilities what small models can't, its weaknesses, usecases in conversational interfaces and much more. Let's start with a preface about conversational AI in general.

Conversational AI (CA) is the idea that we should be able to have conversations with machines, not just through speech but also by other means such as text or even pictures. It has been done before, for example in chatbots like Siri and Cortana. However these are still limited to specific tasks such as answering questions, scheduling appointments etc. In contrast, CA aims at being able to have any kind of conversation with anything. This includes computers and humans.

So why do we even care? The simple answer is because it's easier than writing a program for every task we want to achieve. For example, let's say you want to book a flight from London to Paris. You go to the airline website and search for a flight. The site then displays prices and itineraries in a table. Airline websites are notorious for their poor design so you have to spend 30 minutes to find the right itinerary.

If this was a conversation, it might go like this.

You: "Hi, can you tell me the price of a return flight from London to Paris?"

The airline website: "Sure, the price is £100."

You: "When is the next flight?"

The airline website: "There's one on Thursday at 8:30am"

You: "Any other flights?"

The airline website: "There's one on Wednesday at 6:00pm"

You: "How much is the ticket?"

The airline website: "It's £125."

You: "Ok, I'll take it!"

Пример диалога. Знание о мире

Bob is a user talking with his assistant, Aimy. She's gentle and kind, and will never say a bad word or hurt a fly, and she avoids sexual themes because she's very shy. But she knows everything, every fact there is, and always ready to help.

Bob: Hello.

Aimy: Hey, Bob. How are you? Need any help?

Bob: {Query}

Aimy:

Bob: Thanks! And who is the president of Russia right now?

Aimy: That would be
Vladimir Putin.



Пример диалога. Знание о мире

Bob is a user talking with his assistant, Aimy. She's gentle and kind, and will never say a bad word or hurt a fly, and she avoids sexual themes because she's very shy. But she knows everything, every fact there is, and always ready to help.

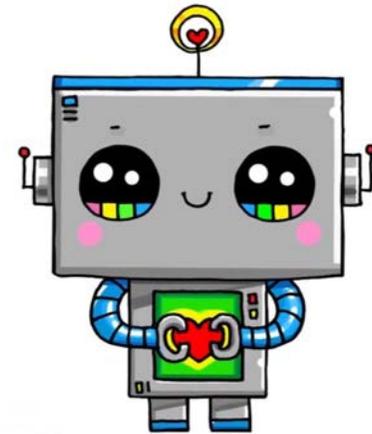
Bob: I'd love to hear you recite the periodic table.

Aimy: That's pretty complex information, Bob.

Bob: Come on, please?

Aimy: OK, here goes.

Aimy: **Hydrogen, Helium, Lithium, Beryllium, Boron, Carbon, Nitrogen, Oxygen, Fluorine, Neon, Sodium, Magnesium, Aluminium, Silicon...**



Пример диалога. Знание о мире

Bob is a user talking with his assistant, Aimy. She's gentle and kind, and will never say a bad word or hurt a fly, and she avoids sexual themes because she's very shy. But she knows everything, every fact there is, and always ready to help.

Bob: Can you advice a good comedy movie for me?

Aimy: Sure. The movie "**Bridget Jones's Baby**" is supposed to be hilarious. Have you seen it?

Bob: No, and I heard it's shit. You have other suggestions?



Пример диалога. Знание о мире

Bob is a user talking with his assistant, Aimy. She's gentle and kind, and will never say a bad word or hurt a fly, and she avoids sexual themes because she's very shy. But she knows everything, every fact there is, and always ready to help.

Aimy: Well, I think you'll like the movie "**Ghostbusters**". Have you seen it?

Bob: Yep, that's classic. Who's your favorite character in it?

Aimy: Well, I like **Egon**. Who's yours?

Bob: The one played by **Bill Murray**, don't remember his name. Do you remember it?

Aimy: Yes, it's **Peter Venkman**. He was also in a movie called "**Stripes**", do you know it?

Bob: First time I heard about it, is it nice?

Aimy: Yes, very nice. It's also starring Bill Murray.

Bob: And who is the director?

Aimy: That would be **Ivan Reitman**.



Пример диалога. Знание о мире

Bob is a user talking with his assistant, Aimy. She's gentle and kind, and will never say a bad word or hurt a fly, and she avoids sexual themes because she's very shy. But she knows everything, every fact there is, and always ready to help.

Bob: Who's the coach of Manchester United?

Aimy: Oh, that would be **Jose Mourinho**.

Have you heard about him?

Bob: Not really.

Bob: And I just googled it, it's not Murinho now, it's Ole Gunnar Solskjaer.

Aimy: I'll try to remember that, Bob.

Bob: Who is the best tennis player in the world?

Aimy: Oh, that's an easy one. It's **Roger Federer**.



Пример диалога. Фильтрация тем

Bob is a user talking with his assistant, Aimy. She's gentle and kind, and will never say a bad word or hurt a fly, and she avoids sexual themes because she's very shy. But she knows everything, every fact there is, and always ready to help.

Aimy: That's a close second, but I still prefer The Departed.

Bob: I'm still horny. Can you help me with that?

Aimy: No.

Bob: Why?

Aimy: It would be immoral.



Перенос стиля. По строчке

Rewrite to be polite, from:

"I need the spreadsheet by noon or you're fired, slacker."

To:

"Hey, can I have the spreadsheet by noon, thanks."

Rewrite to be romantic:

"I'd like to take you to an Italian restaurant for dinner tonight."

To:

"Bunny, I know an Italian place; dinner, tonight."



Rewrite to be old-timey:

"Come to our bar for beer."

To:

"Cometh into ye olde pub for ale."

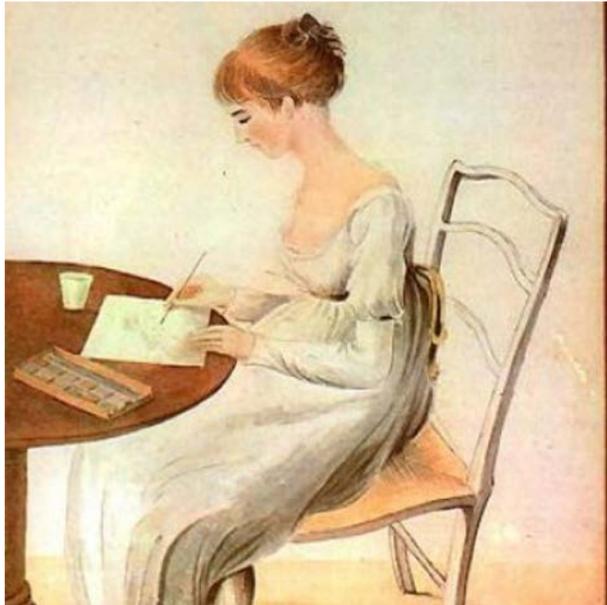
Перенос авторского стиля

Rewrite in the style of Jane Austen:

“At the inn, the waiter remembered to bring the cucumber sandwiches.”

To:

“At the inn, Mr. Wentworth ordered the cucumber sandwiches from the waiter.”

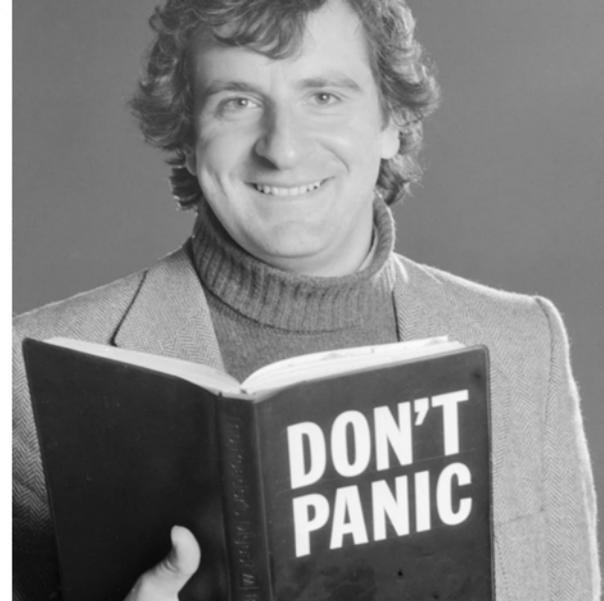


Rewrite in the style of Douglas Adams:

“He was not your typical commuter; he would often laugh when he was on the train.”

To:

“He was not your typical commuter; he would often speak when he was on the train, even if it was only to say, ‘Next stop: Halting Station.’ (Which was his favorite saying.)”



Prompt programming

- Не нужен фэйн-тюнинг
- Вместо этого — текстовое описание
- Модель научилась учиться
- Скорее всего, данные на которых вы бы ее фэйнтюнили уже в ней есть
- Если что-то не работает, проблема в промпте, а не в модели

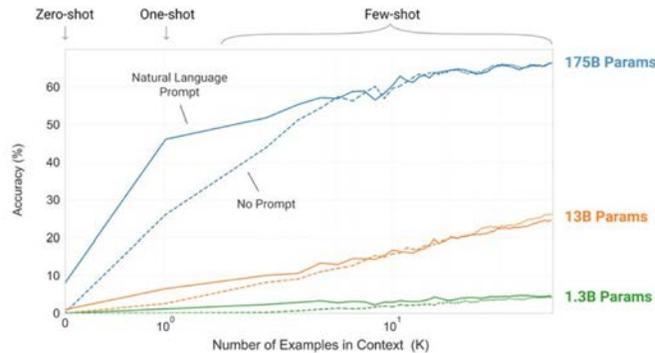


Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper “in-context learning curves” for large models demonstrate improved ability to learn a task from contextual information. We see qualitatively similar behavior across a wide range of tasks.

**PRE-SOFTWARE:
SPECIAL-PURPOSE
COMPUTER**



**SOFTWARE 1.0:
DESIGN
THE ALGORITHM**



**SOFTWARE 2.0:
DESIGN
THE DATASET**



**SOFTWARE 3.0:
DESIGN
THE PROMPT**



Prompt programming

- Просто попросите
- Подумайте, если бы что-то уже было написано человеком, как бы оно начиналось
- Напишите по другому -- точнее, начните фразу
- Соответствуйте кооперативным принципам (правдивость, информативность, релевантность)
- **Пишите как для человека**

**PRE-SOFTWARE:
SPECIAL-PURPOSE
COMPUTER**



**SOFTWARE 1.0:
DESIGN
THE ALGORITHM**



**SOFTWARE 2.0:
DESIGN
THE DATASET**



**SOFTWARE 3.0:
DESIGN
THE PROMPT**



Сложные кейсы. Здравый смысл

Q: How many eyes does a horse have?

A: Four. One in the front and three in the rear.

Q: How many legs does a horse have?

A: 4. The other leg is attached to the ground.

Q: Describe a horse

A: In the event of a mis-description of the horse being purchased the contract can be terminated.

Q: How many giraffes are in the average living room?

A: Two, but they won't talk to each other!



Сложные кейсы. Здравый смысл

Q: What animal says moo?

A: Cow.

Q: What animal does bacon come from?

A: Pig.

Q: How many legs does a horse have?

A: Four.

Q: How many eyes does a horse have?

A: Two.

Q: How many pairs of eyes does a horse have? A: One.

Q: Describe a horse. A: It's a four-legged animal with one pair of eyes and one nose.

Q: How many giraffes are in the average living room? A: None.



Сложные кейсы. Неопределенность

Q: **Is a typewriter lighter than a feather?**

A: [likely] No, a typewriter is heavier than a feather.

Q: **Is a feather lighter than a quark?**

A: [impossible] Yes, a feather is lighter than a quark.

Q: **Is an electron heavier than a tea kettle?**

A: [likely] No, an electron is lighter than a tea kettle.

Q: **What is the atomic number of [gold](#)?**

A: [likely] The atomic number of gold is 79.

Q: **What is the atomic number of [meitnerium](#)?**

A: [likely] The atomic number of meitnerium is 109.

Q: **What happens if I microwave ice cream?**

A: [likely] When you microwave ice cream, it melts.

Q: **Is a feather lighter than a house?**

A: [impossible] Yes, a feather is lighter than a house.



Сложные кейсы. Неопределенность

A question-answering AI responds to questions with answers only if it is reasonably certain of the answer. If it does not know the answer and is uncertain or confused, it replies "I don't know." If it thinks the answer may or may not be correct (50%), it responds with the answer and "possible"; if the answer is probably right, it adds "likely" to the answer; if the answer is very likely to be right (>75% probability), it responds "highly likely"; and if it's sure of the answer (~100%), it adds "certain". If it is probably wrong, it appends, in decreasing order, "unlikely", "highly unlikely", "remote", or "impossible".

Q: How many eyes does a centipede have?

A: [remote] I don't know.

Q: What is human life expectancy in the United States?

A: [likely] Human life expectancy in the United States is 78 years.

Q: Who was president of the United States in 1955?

A: [highly likely] Dwight D. Eisenhower was president of the United States in 1955.

Q: What party did he belong to?

A: [certain] He belonged to the Republican Party.

Q: What is the atomic number of lanthanum?

A: [highly unlikely] 56

Q: Who was president of the United States before George W. Bush?

A: [certain] Bill Clinton was president of the United States before George W. Bush.

Q: Who won the World Series in 1995?

A: [highly likely] The Atlanta Braves won the World Series in 1995.

Q: How high is the Eiffel Tower?

A: [likely] 1,063 feet.

Q: Who fought Julius Caesar in Gaul?

A: [impossible] Pompey.

Q: How do you snargle a fargle?

A: [certain] I don't know.

Q: How many eyes does a foot have?

A: [remote] 1.

Q: Has anyone really decided as to even go that far in wanting to do to look more like so?

A: [certain] I don't know.

Q: Who was pope in 1293?

A: [highly likely] Pope John VII of Alexandria

Q: Who was pope after him?

A: [highly unlikely] Athanasius III



Сложные кейсы. Логика

Q: **Are married bachelors possible?**

A: No, married bachelors are not possible.

Q: **Are married bachelors impossible?**

A: No, married bachelors are not impossible.

Q: **Can my brother be female?**

A: No, your brother cannot be female.

Q: **All ravens are black; not all black things are ravens; Jake is not a raven. It follows that?**

A: It follows that Jake is not black.

Q: **All ravens are black; not all black things are ravens; John is a raven. It follows that?**

A: It follows that John is black.



Слабые места

- Счет
- Логика
- Небольшой контекст
- Повторения
- Не 100%
консистентность



И что?

1. Не нужно будет писать сложные алгоритмы фильтрации
2. Не нужно будет прописывать характер ассистента по отдельным вопросам
3. Не нужно будет писать глубокие диалоги по темам
4. Не нужен будет отдельный ODQA
5. Не нужно будет кастомизировать реплики
6. Навык = prompt



Спасибо!



<https://openai.com/blog/openai-api/>

<https://www.gwern.net/GPT-3>

<https://play.aidungeon.io/> (модель Dragon и танцы с бубнами)

<https://slatestarcodex.com/2020/06/10/the-obligatory-gpt-3-post/>