Новости из мира NLP

Что было, что грядёт, как пользоваться тем,
что есть сейчас?

Хрущев Михаил,
Ведущий разработчик,
Руководитель группы претрейна YandexGPT

План доклада

01 Что такое языковые модели

02 О YandexGPT

03 Немного истории

04 Как обучать языковые модели

05 Как решать ваши задачи через языковые модели

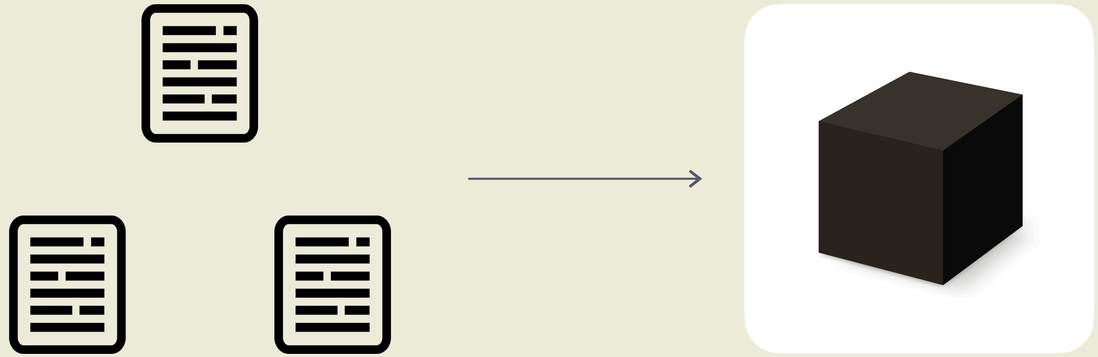
06 Что дальше?

01

Что такое языковые модели?

Что такое языковые модели?

Языковые модели — модели, способные последовательно продолжать текст



Яндекс — это …  компания

Что такое языковые модели?

Языковые модели — модели, способные для любого префикса предложить вероятностное распределение следующего слова (токена)

Яндекс — это … 

компания	0.4
технологическая	0.3
информационная	0.2
…	
фрукт	1E–10

02

О YandexGPT

YandexGPT

1. YandexGPT — семейство гигантских языковых моделей
2. Размеры моделей: от 100 млн до 100 млрд параметров
3. Решает произвольные задачи классификации и генерации на текстах
4. Используется более, чем в половине продуктов Яндекса:
   - Болталка
   - Реклама
   - Поиск
   - Браузер
   - Маркет

03

История языковых моделей

История языковых моделей

Марковские цепи: храним статистики последовательностей из n слов, вычисляем вероятности P(w_n|w_1, …, w_n–1)

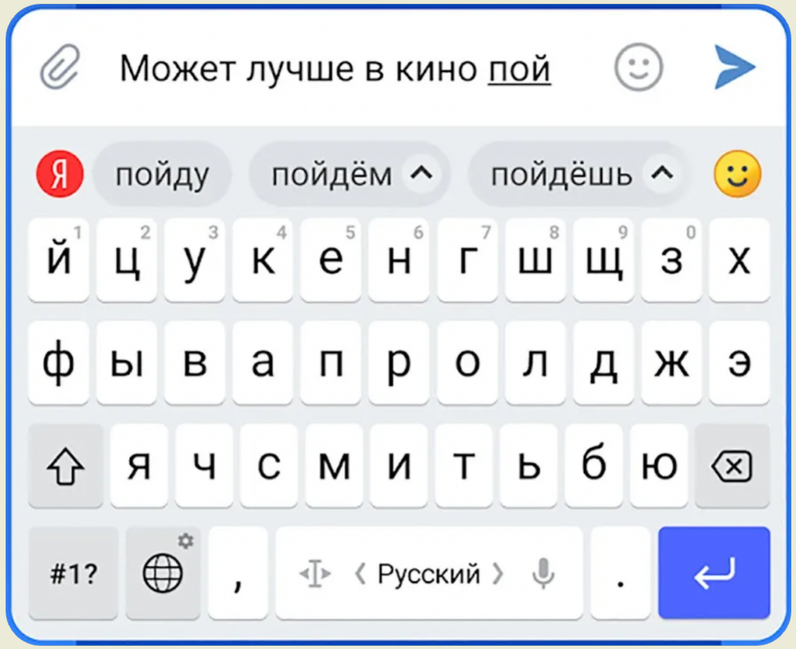
жадина	говядина	ресторан	40
жадина	говядина	соленый	15
жадина	говядина	пустая	17
жадина	говядина	турецкий	13
жадина	говядина	в	1

P(соленый|жадина, говядина) = 15/(40+15+17+13+1) = 0.174

История языковых моделей

Марковские цепи: храним статистики последовательностей из n слов, вычисляем вероятности P(w_n|w_1, …, w_n–1)

P(соленый|жадина, говядина) = 15/(40+15+17+13+1) = 0.174



История языковых моделей

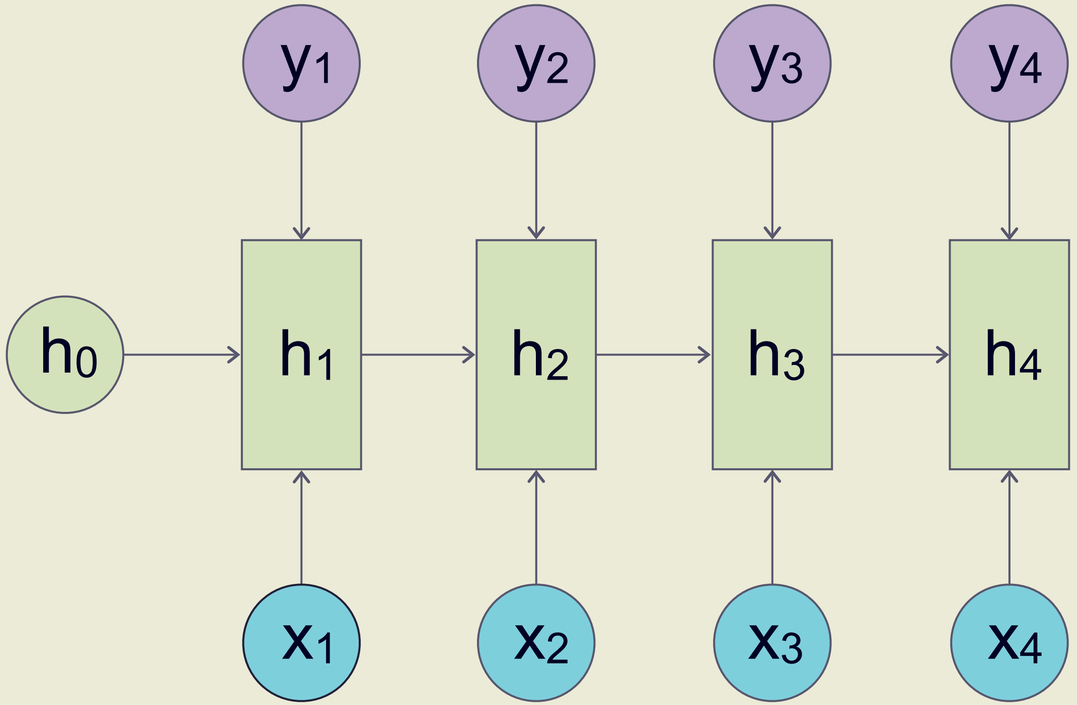
Марковские цепи: храним статистики последовательностей из n слов, вычисляем вероятности P(w_n|w_1, …, w_n–1)

Недостатки:
Модель никогда
не сгенерирует n слов,
которые ни разу не видела

Недостатки:
Модель знает только
о контексте из n–1 слова.
Не более

История языковых моделей

RNN — реккурентные сети



История языковых моделей

RNN — реккурентные сети

Плюсы:
Мы уже имеем возможность учитывать весь контекст

Минусы:
RNN не может учиться на GPU быстро

Плюсы:
(со взглядом в будущее):
Генерация очень быстрая

Минусы:
RNN в реальности смотрит лишь на небольшое количество токенов слева

История языковых моделей

Трансформеры (Attention Is All You Need, 2017), GPT–2 (2019)

- Быстро учатся, модели до 1.3B параметров уже генерируют связный текст

История языковых моделей

GPT–3 (Language Models are Few-Shot Learners). Большие модели (от 1B до 175B параметров):

История языковых моделей

GPT–3 (Language Models are Few-Shot Learners). Большие модели (от 1B до 175B параметров):

- Понимают явные указания

Context →	Final Exam with Answer Key Instructions: Please carefully read the following passages. For each passage, you must identify which noun the pronoun marked in *bold* refers to. ===== Passage: Mr. Moncrieff visited Chester's luxurious New York apartment, thinking that it belonged to his son Edward. The result was that Mr. Moncrieff has decided to cancel Edward's allowance on the ground that he no longer requires *his* financial support. Question: In the passage above, what does the pronoun "*his*" refer to? Answer:
Target Completion →	mr. moncrieff

История языковых моделей

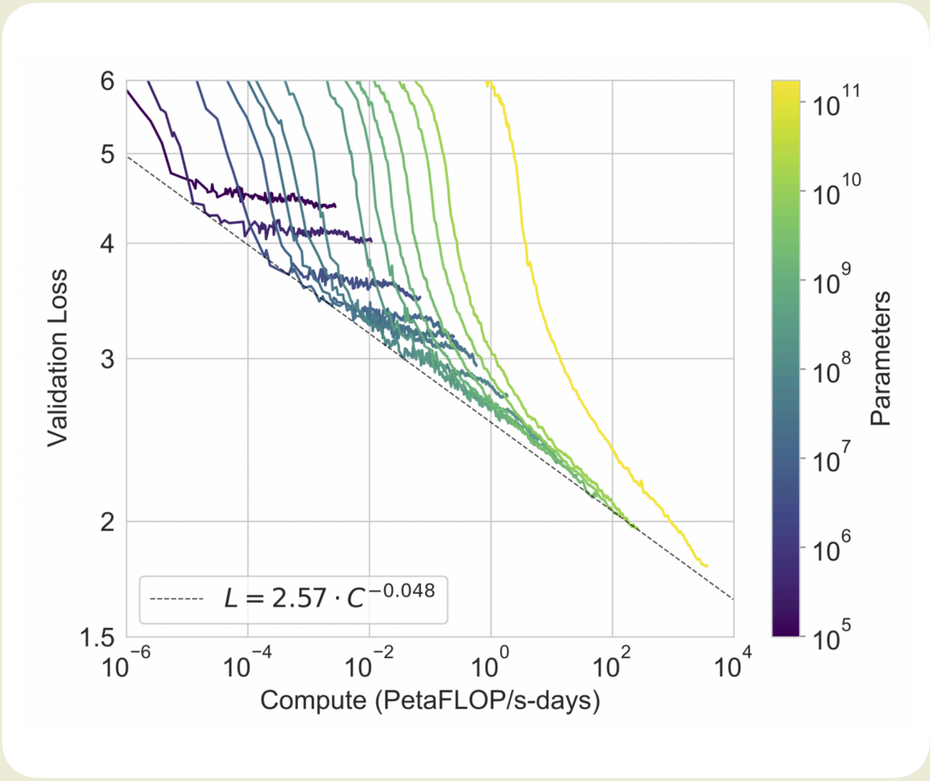
GPT–3 (Language Models are Few-Shot Learners). Большие модели (от 1B до 175B параметров):

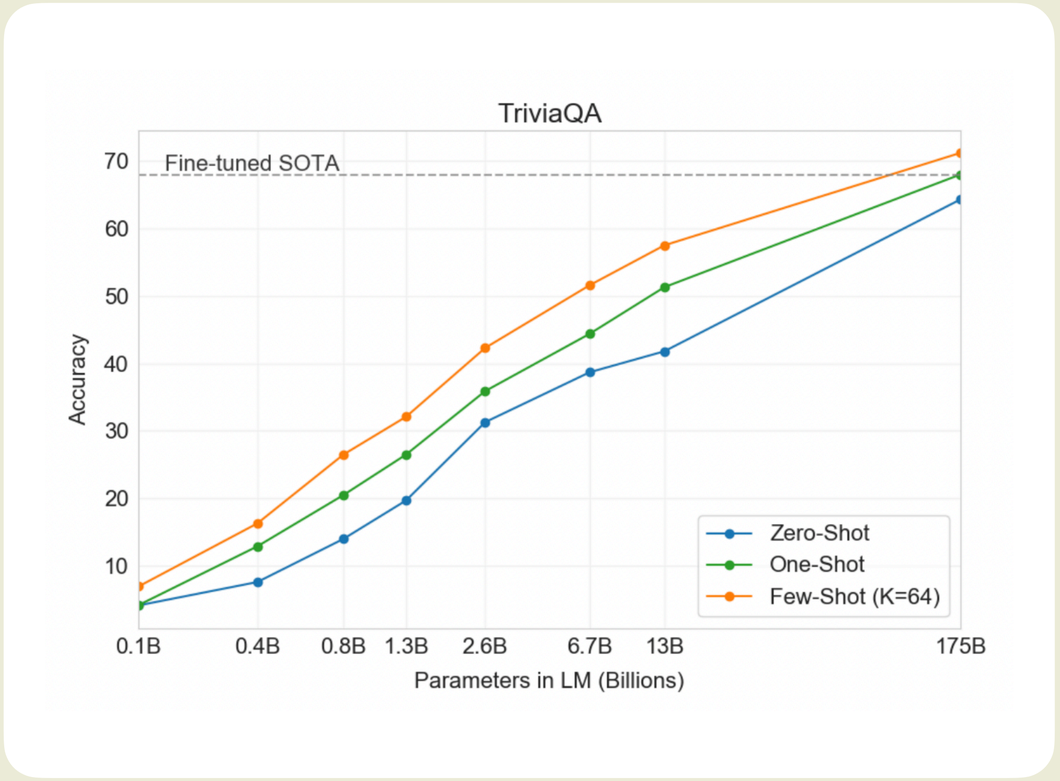
- Можно подсказать несколько примеров — будут работать лучше

```
Title:  The_Blitz

Background:  From the German point of view, March 1941 saw an improvement.
The Luftwaffe flew 4,000 sorties that month, including 12 major and
three heavy attacks.  The electronic war intensified but the Luftwaffe
flew major inland missions only on moonlit nights.  Ports were easier to
find and made better targets.  To confuse the British, radio silence was
observed until the bombs fell.  X- and Y-Gerät beams were placed over
false targets and switched only at the last minute.  Rapid frequency
changes were introduced for X-Gerät, whose wider band of frequencies and
greater tactical flexibility ensured it remained effective at a time when
British selective jamming was degrading the effectiveness of Y-Gerät.

Q: How many sorties were flown in March 1941?

A: 4,000

Q: When did the Luftwaffe fly inland missions?

A:
```

История языковых моделей

GPT–3 (Language Models are Few-Shot Learners)



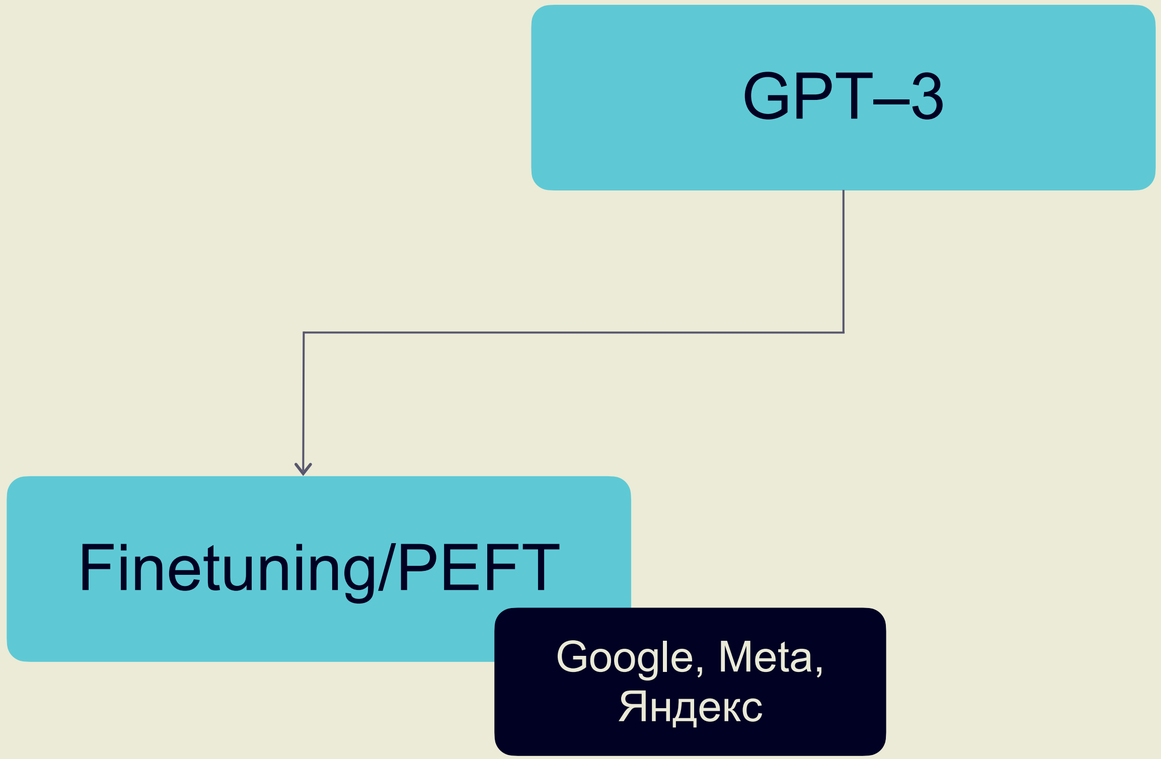


История языковых моделей

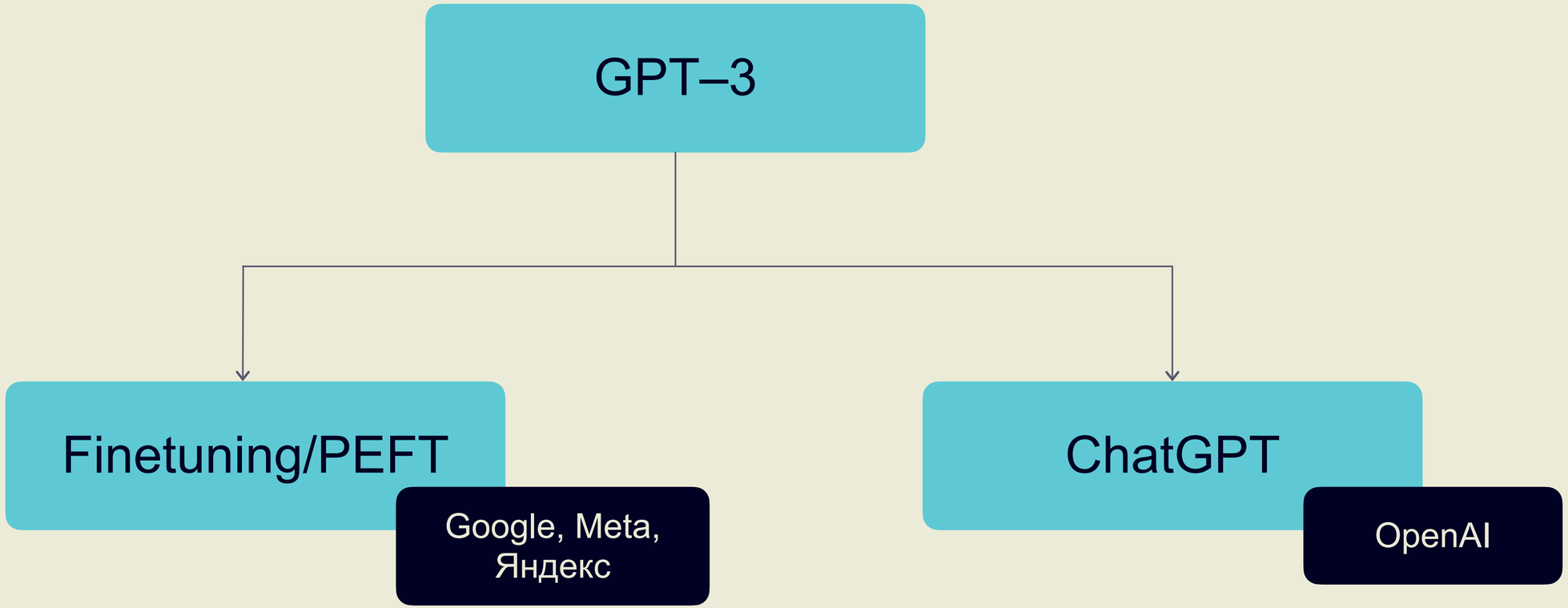
GPT–3 (Language Models are Few-Shot Learners). Большие модели (от 1B до 175B параметров):

- На практике часто не подходят для продуктовых применений

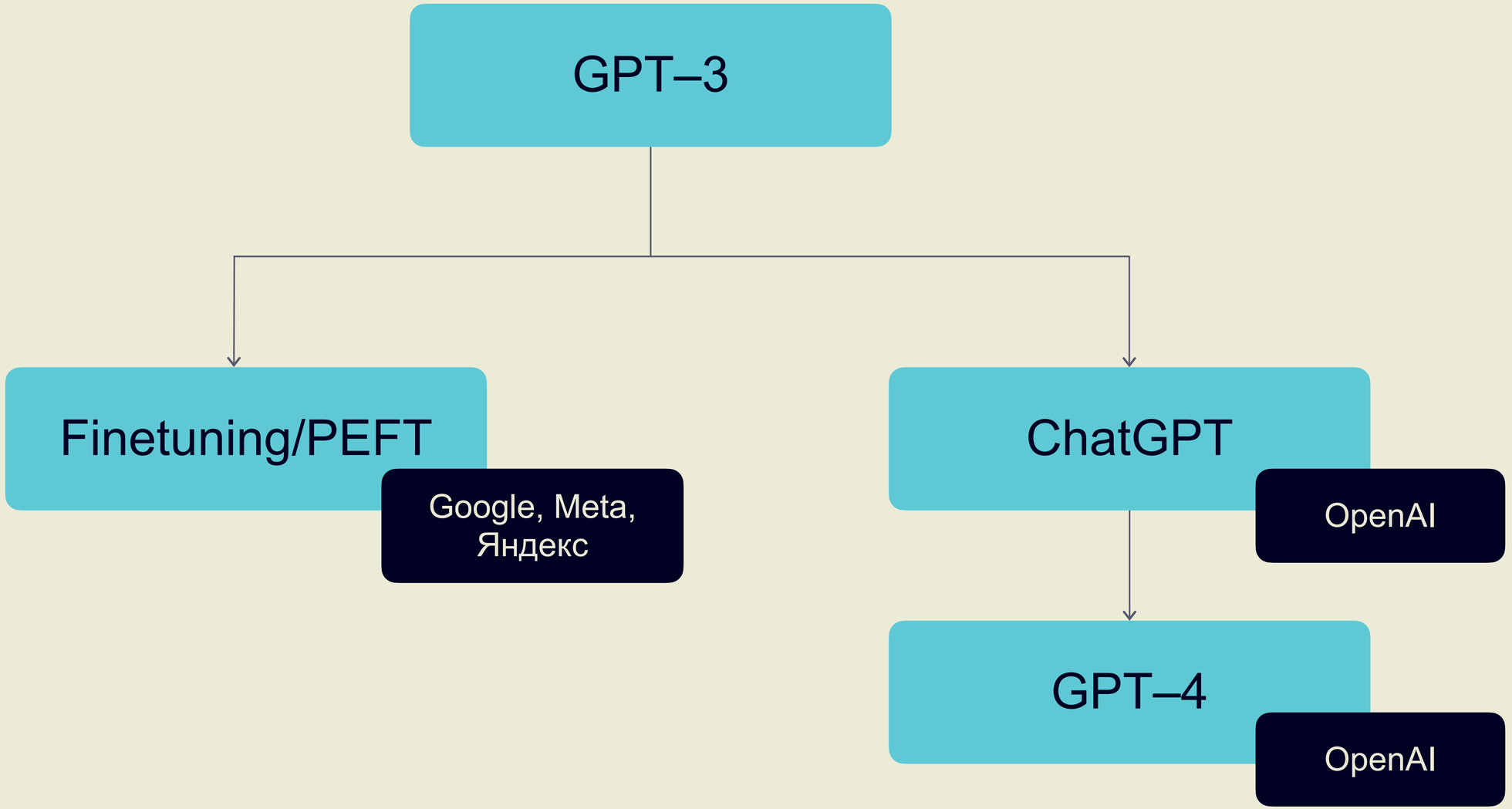
История языковых моделей



История языковых моделей



История языковых моделей



История языковых моделей

- GPT–3 (Language Models are Few-Shot Learners). Большие модели
- ChatGPT и GPT–4:
  - Все то же самое, но намного лучше
  - Поддерживает диалоговость
  - Умеет работать с кодом

04

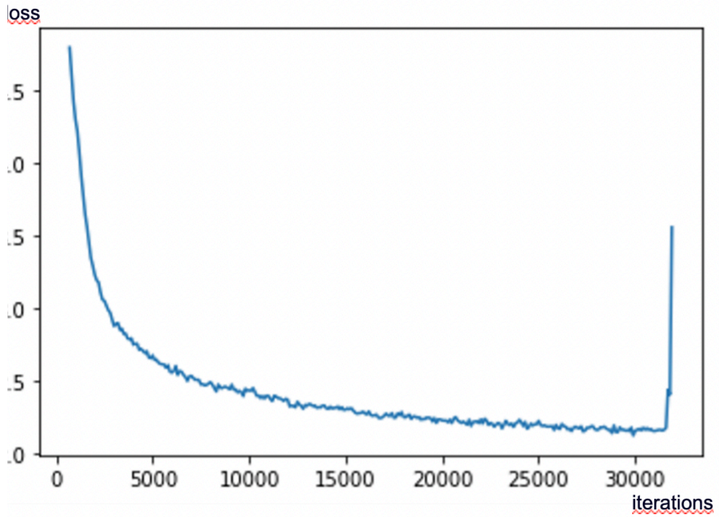
Как обучать языковые модели

Обучение больших моделей

1. Предобучение — показываем модели несколько терабайт чистых уникальных текстов
2. Предварительный finetune — учим решать десятки интересных нам задач
3. Downstream finetune — затачиваем модель под решение конкретной задачи:
   - High resource задача: > 5 000 примеров в обучении
   - Low resource задача: 50–5 000 примеров в обучении
4. Или: Alignment + RL — учим модель на разметках людей выполнять произвольные инструкции. Таким образом направленно усиливаем zeroshot

Проблемы предобучения больших языковых моделей

- Нужно много GPU*дней: самое большое обучение у нас — 52k GPU*дней
- Нужно очень много данных для обучения:
  - > 2 TiB чистых данных для обучения 13B модели
  - x2 данных за x2 параметров
- Модели могут расходиться



Проблемы предобучения больших языковых моделей

YaLM 100B: Как удалось её обучить

Data Fest 2022
Проблемы приготовления больших моделей

Особенности предварительного файнтюна

- Во время файнтюна мы показываем модели десятки различных задач, чтобы:
  - Модель поняла, что ей всегда нужно решать задачи вида: X → Y
  - Выучила особенности продуктов Яндекса
- Такой файнтюн улучшает среднее качество по моделям на наших бенчмарках:
  - +7пп на задачах low resource
  - +2пп на задачах high resource

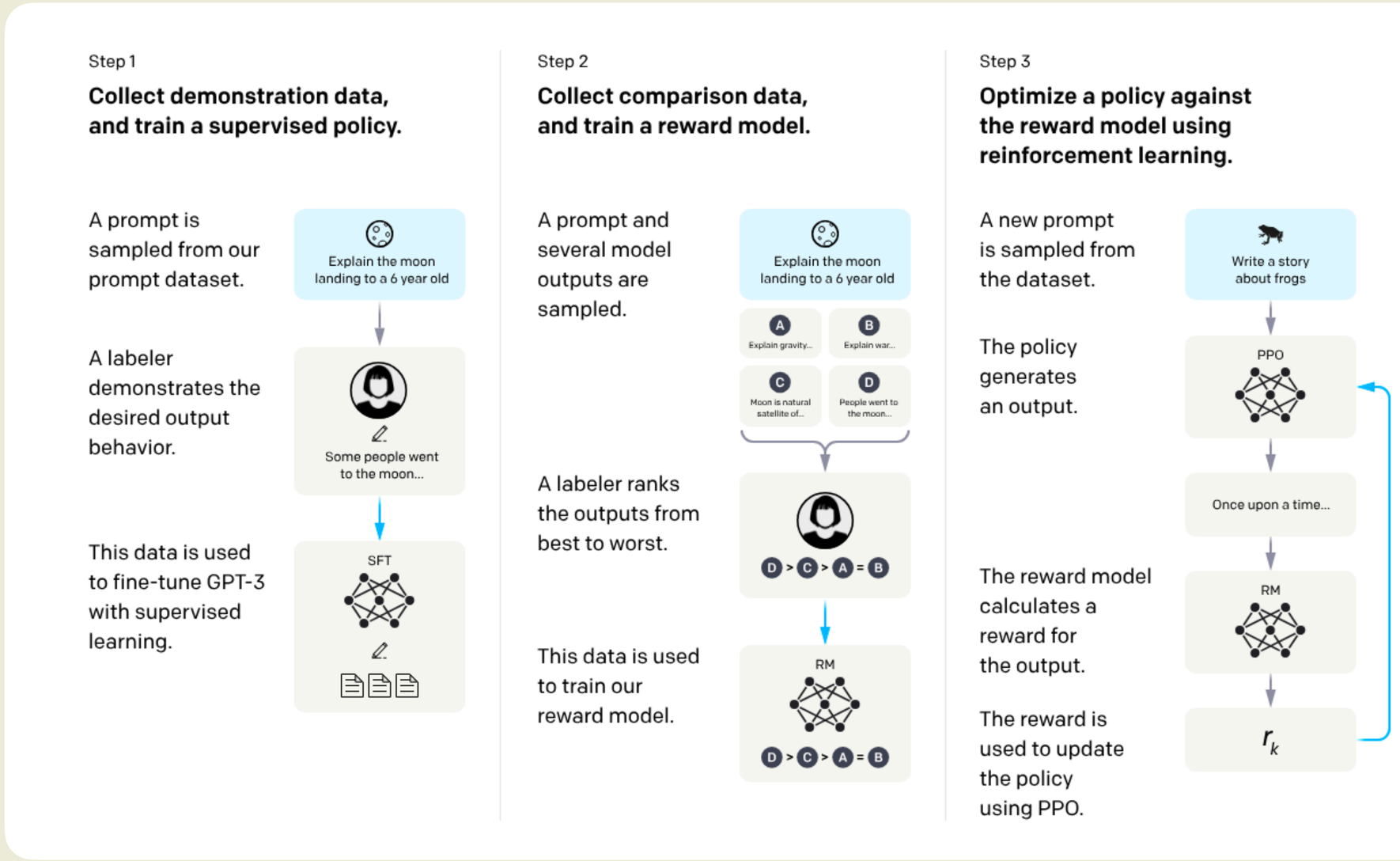
Downstream finetune

- HighRes: > 5k примеров
  - Лучше всего работает полный файнтюн: небольшой learning rate
- LowRes: < 5k примеров
  - Лучше всего работают адаптеры:
    - Prompt tuning: v1, v2
    - LoRA

Prompt tuning

Вместо того, чтобы подбирать zero-shot подводку — обучаем ее

Alignment + RL



Выводы

- Предобучение — это сложно и дорого
- Предварительный файнтюн — сильно улучшает качество модели на downstream задачах
- Итоговая модель файнтюнится полностью или частично на любые прикладные задачи

05

Как решить вашу задачу

Открытые АПИ и сервисы

Языковые модели:

1. OpenAI: GPT 3.5, ChatGPT, GPT 4
2. Anthropic
3. YandexGPT: Yandex Cloud

Hugging face

Языковые модели:

1. Llama 2 — семейство сильный англоязычных моделей
2. Llama 2 chat — aligned llama 2 модели
3. Bloom (0.5B, 1B, 3B, 175B)
4. BloomZ — предварительный finetune

Hugging face

```
>>> from transformers import AutoModelForCausalLM, AutoTokenizer

>>> model = AutoModelForCausalLM.from_pretrained("EleutherAI/gpt-j-6B")
>>> tokenizer = AutoTokenizer.from_pretrained("EleutherAI/gpt-j-6B")

>>> prompt = (
...     "In a shocking finding, scientists discovered a herd of unicorns living in a remote, "
...     "previously unexplored valley, in the Andes Mountains. Even more surprising to the "
...     "researchers was the fact that the unicorns spoke perfect English."
... )

>>> input_ids = tokenizer(prompt, return_tensors="pt").input_ids

>>> gen_tokens = model.generate(
...     input_ids,
...     do_sample=True,
...     temperature=0.9,
...     max_length=100,
... )
>>> gen_text = tokenizer.batch_decode(gen_tokens)[0]
```

06

Что дальше?

Вызовы для языковых моделей

1. Для гигантских языковых моделей нужно данных больше, чем реально найти
2. Этичность и безопасность: как гарантировать, что мы не увидим «Терминатор» в реальной жизни
3. Быстрая и дешёвая генерация ответов — только крупные компании могут позволить себе большие модели в проде
4. Мультимодальность и мультиязычность

Вызовы для пользователей языковых моделей

1. Низкое доверие к современному ИИ — не ясно, кто несет ответственность за ошибки
2. Многие не понимают, что можно сделать с современными языковыми моделями
3. Близится время, когда ИИ действительно сможет заменить людей на их местах

Выводы

- Большие языковые модели уже доступны для всех
- Они позволяют решать новые классы задач
- Будьте наготове: языковые модели развиваются с огромной скоростью и скоро будут влиять на всё новые и новые аспекты нашей жизни

Спасибо за внимание!

Хрущев Михаил,
Ведущий разработчик,
Руководитель группы претрейна YandexGPT