Языковая модель и предсказание следующего токена

«Без труда не вытащишь и рыбку из...» — продолжение знает каждый. А как подсказка на клавиатуре угадывает ваше следующее слово?

Закончите фразу: «Без труда не вытащишь и рыбку из...» Теперь продолжите «Сегодня на занятии мы обсудим...». В первом случае почти все назовут одно слово, во втором появятся разные ответы. Разберём, как одна задача — продолжение текста — связывает пословицы, подсказки клавиатуры и ответы языковой модели.

Ранее

Ранее мы разобрали: классификация и вероятности классов. Здесь используем эти идеи в новой задаче.

← Классификация и вероятности классов

1. Продолжение текста

Без труда не вытащишь и рыбку из... → **пруда**

Интерактивный эксперимент — в электронной версии урока.

Языковая модель получает начало текста и предсказывает, что может идти дальше.

Не у каждой фразы есть столь однозначное завершение. Например, «Сегодня на занятии мы обсудим...» можно продолжить по-разному: выбор зависит от дисциплины, темы занятия и предыдущего разговора.

Россия — Москва
Франция — Париж
Швейцария — Берн
Германия — ... → **Берлин**

Интерактивный эксперимент — в электронной версии урока.

Во втором примере пары «Россия — Москва», «Франция — Париж», «Швейцария — Берн» задают отношение «страна — столица». Чтобы продолжить строку «Германия — ...», нужно

сохранить это правило и назвать столицу Германии — Берлин. Так продолжение текста становится ответом на вопрос, хотя сам вопрос явно не сформулирован.

Примеры в запросе объясняют сразу и задачу, и формат результата. Если вместо столиц дать пары из русского слова и английского перевода, изменится правило продолжения. Если привести несколько отзывов с метками «положительный» и «отрицательный», получится классификация.

Во всех этих случаях модель одна и та же: меняется только текст, который мы ей передаём. Примеры в этом тексте помогают ей определить, что именно нужно сделать, хотя и не гарантируют правильного ответа.

Оба примера иллюстрируют задачу языкового моделирования: оценить, какой фрагмент текста может следовать за уже имеющимся. Для начала «Кот спит на» подходят, например, «полу», «диване» и «спине». Модель оценивает эти продолжения с учётом предшествующего текста.

В современных наиболее распространённых языковых моделях, которые называют авторегрессионными, ответ строится последовательно. Чтобы получить длинный ответ, вычисление повторяют: добавляют выбранный фрагмент к исходному тексту и оценивают следующий. Поэтому даже ответ на сложный вопрос складывается из последовательности небольших продолжений. Существуют и другие подходы, например диффузионные языковые модели; в этом сюжете мы их не рассматриваем.

Заметка
Авторегрессионная языковая модель оценивает следующий токен по контексту; повторение выбора порождает текст.

Пока удобно представлять такие фрагменты словами. Более точное название рабочей единицы модели — токен; он может совпадать со словом или быть его частью. К устройству токенов мы вернёмся после того, как разберём сам принцип предсказания. Пока в тексте мысленно можно заменять «токен» на «слово» без существенной потери смысла.

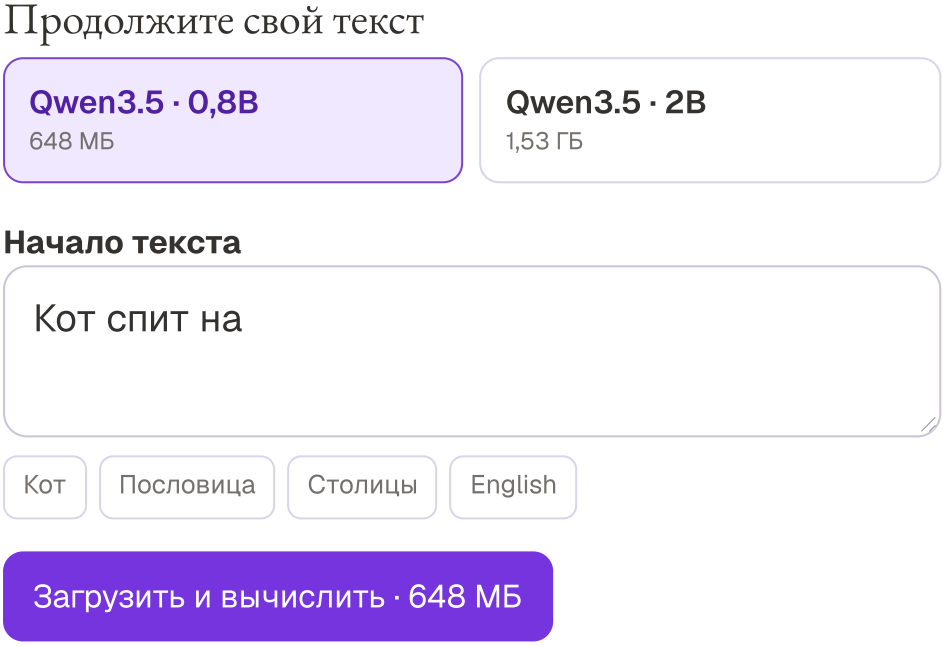
Контекст — текст, доступный модели при текущем предсказании. Изменение примеров или начала фразы меняет контекст, поэтому меняются и вероятные продолжения.

> Перед моделью поставили «Чай налили в» вместо «Кот спит на». Почему вероятные продолжения могут измениться?
>
> - ○ Изменился контекст
> - ○ Обязательно произошло новое обучение
> - ○ Все веса автоматически обнулились
> - ○ Вероятность токенов не зависит от текста

2. Вероятности следующего токена

Модель не сразу выдаёт одно готовое слово: она оценивает возможные следующие токены. Для начала «Кот спит на» подходят разные продолжения. Насколько модель предпочитает одно другому, можно увидеть в её вероятностях. Вероятностный выход модели

Введите начало фразы и вычислите продолжения. Первый запуск скачает выбранную модель; затем она работает на вашем устройстве.



Интерактивный эксперимент — в электронной версии урока.

Начните с «Кот спит на», затем замените начало на «Чай налили в». Изменился контекст — изменился прогноз. Теперь выберите одно из продолжений: добавленный токен тоже станет частью контекста, поэтому следующий набор вариантов будет уже другим.

Видимые варианты — лишь небольшая часть словаря. Их вероятности не обязаны составлять 100 %: остальная доля принадлежит другим токенам. Нажатие на «Выбрать по шансам» может добавить и токен, который не попал в первые шесть.

После каждого выбора повторяется один цикл: **предсказать → выбрать → добавить → повторить**. Такой способ последовательного порождения текста называют **авторегрессионным**.

Предложенное продолжение не сообщает, где находится какой-то настоящий кот. Модель продолжает заданную строку. Чтобы отвечать о конкретном животном, в контексте должны появиться сведения о нём: например, описание наблюдения или содержание фотографии, если система умеет её обрабатывать.

Модель вернула вероятности следующего токена. Что нужно сделать, чтобы продолжить генерацию после выбора одного токена?

- ○ Всегда повторять выбранный токен
- ○ Удалить весь контекст
- ○ Добавить его к контексту и снова получить распределение
- ○ Заново обучить модель с нуля

3. Поисковые подсказки и автодополнение

Поисковые подсказки предлагают несколько продолжений одного запроса. Добавим слова — изменятся и варианты.

Я увидел кота
Я увидел кота **на крыше**
Я увидел кота **во сне**
Я увидел кота**, который гнался за собакой**

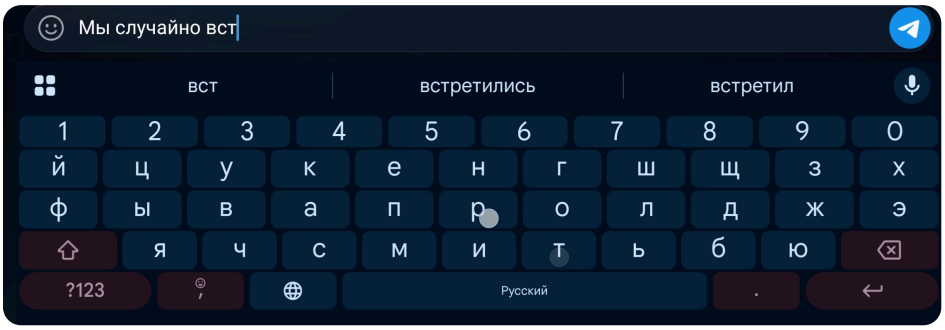
Настоящий поисковик учитывает ещё частоту запросов и историю пользователя, но главное наблюдение от этого не зависит: контекст сужает выбор, не сводя его к единственному ответу. Языковой модели тоже приходится распределять предпочтения между несколькими возможными продолжениями. Языковые модели вокруг

Контекст помогает продолжить даже незаконченное слово. После «Буду рад встре...» возможны и «встретиться», и «встретить»: начальные буквы совпадают, а дальше получаются разные конструкции — «встретиться с вами» или «встретить вас».

Буду рад встре...
Буду рад встре**титься с вами**
Буду рад встре**тить вас**

Если для ответа важны адресат, намерение или стиль, укажите их в запросе. Они сужают множество допустимых формулировок так же, как несколько лишних слов сужают список подсказок.

Подсказки экранной клавиатуры обновляются после каждого слова: контекст изменился — система заново предлагает продолжения. Человек может принять вариант, выбрать другой или написать своё.



Клавиатура предлагает закончить начатое слово: «встретились» или «встретил».

После добавления слов подсказки меняются: теперь предлагаются «тобой», «уважением», «этим».

Генеративная модель делает то же самое, только выбирает за человека программа. Выбранный фрагмент добавляется к тексту и становится условием следующего предсказания. Поэтому раннее решение способно изменить весь дальнейший ответ: текст после «на крыше» пойдёт иначе, чем после «во сне».

Клавиатура и большая языковая модель устроены и обучены по-разному, но цикл у них общий: предсказать, выбрать, добавить, повторить.

Предсказание по контексту помогает и исправлять ошибки. В сочетании «Я увидел котт» вариант «кота» подходит по смыслу и грамматике: после «увидел» ожидается существительное в нужной форме. Без окружающих слов оснований для выбора было бы меньше.

Набранный текст

Я увидел котт

Возможно, вы имели в виду

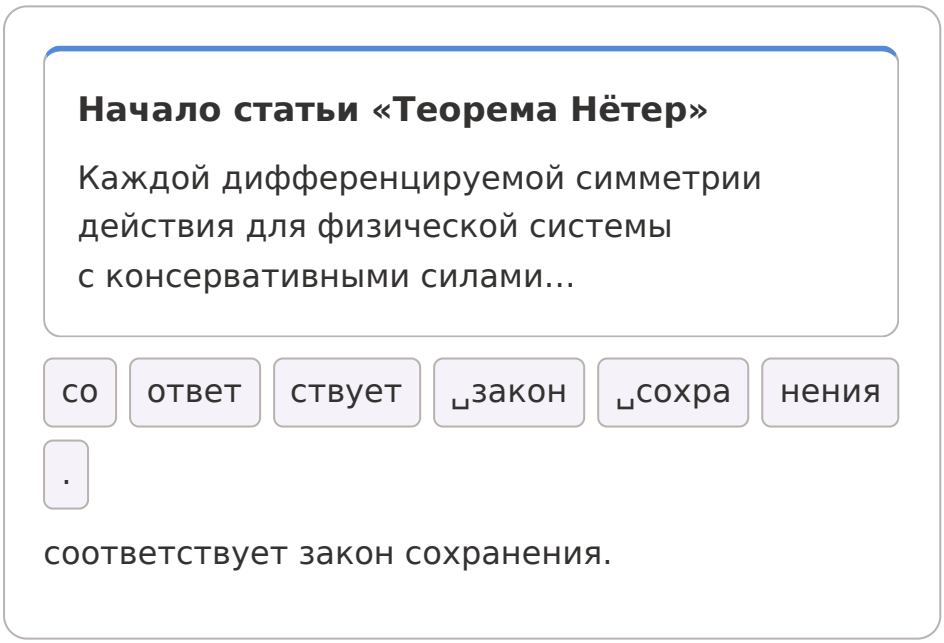
Я увидел **кота**

Но статистически привычная замена может противоречить намерению автора. Редкая фамилия, обозначение прибора или специальный термин выглядят для системы как опечатка, и автокоррекция портит правильный текст.

Здесь снова проявляется ограничение: подходящий по контексту вариант ещё не означает верный. В бытовой фразе цена ошибки невелика. В научном тексте она выше — посмотрим, что происходит, когда тот же принцип применяют к нему.

4. Последовательная генерация текста

Возьмём начало статьи о теореме Нётер Теорема Нётер: теорема «утверждает, что каждой дифференцируемой симметрии действия для физической системы с консервативными силами...» Закроем конец предложения. Чтобы восстановить «соответствует закон сохранения», уже недостаточно помнить расхожую фразу: нужно владеть терминологией, связями между понятиями и логикой изложения. Модель строит такое продолжение по частям, токен за токеном, и каждый добавленный фрагмент становится частью текста, по которому предсказывается следующий.



Границы фрагментов условные.

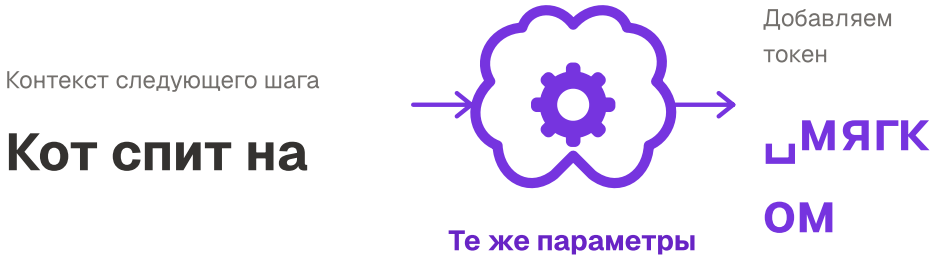
Хорошо продолжать множество разных текстов — от пословиц до статей по теоретической физике — несравнимо труднее, чем закончить одну знакомую фразу. Оказалось, что эту задачу можно решать всё лучше, если увеличивать три вещи: объём и разнообразие текстов, число параметров модели и объём вычислений, необходимый для обучения языковой модели.

Так получились современные большие языковые модели, LLM (Large Language Models). Чтобы модель вела себя как ассистент и выполняла просьбы, её дополнительно обучают следовать инструкциям,

но способ работы остаётся прежним: ответ строится последовательно, фрагмент за фрагментом.

Связность и научный стиль не подтверждают истинность текста. Модель могла бы написать и другое продолжение, столь же гладкое.

Один шаг становится началом следующего



␣мягком ␣диване ␣у ␣окна .

Проиграть цикл | Добавить токен | Сначала

Модель получает начало текста и предсказывает следующий токен.

Добавленный фрагмент возвращается на вход модели. Для наглядности фрагменты показаны целыми словами.

Интерактивный эксперимент — в электронной версии урока.

При обычной генерации параметры остаются прежними. Меняется контекст: к нему добавляется выбранный токен. Для каждого нового положения модель вычисляет новые вероятности. Одно вычисление не выдаёт весь готовый ответ: программа повторяет этот цикл до условия остановки. Генерация текста

> Модель без дообучения пишет ответ на вопрос. Что происходит с её параметрами при обычной генерации?
>
> - ○ Превращаются в текст запроса
> - ○ Они остаются прежними
> - ○ Обязательно меняются после каждого слова
> - ○ Исчезают после первого токена

Авторегрессионная языковая модель оценивает следующий токен по контексту; повторение выбора порождает текст.

Следующий токен оценивается с учётом входной последовательности.

Генерация использует уже обученные параметры.

Домашняя работа

1. Модель получила «В читальном зале было» и выбрала продолжение « тихо». Какой текст должен стать входом следующего шага обычной генерации?

   В читальном зале было тихо

   Только первоначальное начало, без добавления

   Новый обучающий набор с правильными ответами

   Только слово «тихо»

Источники

- Qwen3.5 0,8B (https://huggingface.co/onnx-community/Qwen3.5-0.8B-ONNX-OPT/tree/fafab72d87a9e6be3925b38caf48286d2838f2d0)
- Qwen3.5 2B (https://huggingface.co/onnx-community/Qwen3.5-2B-ONNX-OPT/tree/2ea7886f48b926aca97de8b0e041ffca7e3ebaa9)
- Генерация текста (https://huggingface.co/docs/transformers/llm_tutorial)
- Вероятностный выход модели (https://d2l.ai/chapter_linear-classification/softmax-regression.html)