Контекстное окно

Вы приложили статью один раз, а потом задали ещё три вопроса «по ней». Откуда модель каждый раз знает, о какой статье речь?

Вы приложили статью, получили объяснение, затем попросили пересказ, квиз и слайды. Во всех четырёх запросах слова «по этой статье» относятся к одному документу. Как он остаётся доступным модели, когда разговор становится длиннее?

Ранее

Токены измеряют длину текста, а инструкции задают способ работы с ним. Теперь соберём все части одного запроса.

← Токены и токенизация

← Дообучение для следования инструкциям

1. Инструкции, документы и история диалога

Контекст — всё, что модель получает для текущего ответа: инструкции приложения, ваш вопрос, переданный текст файлов и прежние реплики. Всё это перед вычислением превращается в одну последовательность токенов со служебными метками ролей.

Посмотрим на обычную задачу преподавателя. Он нашёл статью Меркулова и Оселедца «Stochastic Gradient Algorithms from ODE Splitting Perspective» и хочет подготовить семинар. Сначала просит объяснить основную идею, затем — краткий пересказ для студентов, пять вопросов для квиза и шесть слайдов в формате Markdown.

Ниже сохранён настоящий диалог из интерактивного курса. Откройте очередной ответ и проследите, что добавится в историю. По фиксированной шкале хорошо видно: статья занимает основную часть входа, а вопросы и ответы постепенно наращивают его.

Одна статья, четыре запроса



Фиксированная шкала: 0–16 000 токенов. Это объём текста записи, без служебной разметки и скрытых рассуждений.

Статья → объяснение → пересказ → квиз → слайды. При каждом новом запросе снова передаются инструкции, статья и прежние реплики. К концу записи — 14 624 токена текста, из них 10 085 — статья.

Откройте ответы и переходите к следующему запросу. Сравните новый текст с накопленной историей.

Интерактивный эксперимент — в электронной версии урока.

При каждом новом обращении приложение в этом опыте снова передавало инструкции, статью и все предыдущие вопросы и ответы. Поэтому фраза «по этому материалу» в запросе слайдов имела

смысл. К концу диалога накопилось **14 624 токена текста**, из них **10 085 — статья**. Это измерения исходной записи с токенизатором DeepSeek.

2. Собрать запрос в пределах окна

Контекстное окно ограничивает объём последовательности, с которой работает модель. Место занимают и входные сведения, и будущее продолжение. Рабочий предел приложения может быть меньше окна самой модели: программа добавляет собственные инструкции, описания инструментов и резервирует часть места.

Теперь подготовим запрос со слайдами сами. В опыте можно оставить всю статью или передать выбранный фрагмент, отметить нужную переписку, написать новый вопрос и выделить место для ответа. Текст всех частей здесь посчитан одним словарём **o200k_base**; поэтому числа немного отличаются от исходной записи DeepSeek. Ваши правки пересчитываются прямо в браузере.

Материалы для шести слайдов

Соберите следующий запрос

Нужны шесть слайдов по всей статье. Рабочий лимит опыта — 12 000 токенов. Что передать заново, а что можно оставить в истории приложения?

15 509 из 12 000 токенов — Лишние 3 509 токенов



0 — Предел 12 000

Инструкции · 86 · Документ · 10 274 · История · 3 034 · Новый вопрос · 67

Резерв ответа · 2 048

> До отправки нужно освободить 3 509 токенов. Попробуйте новый диалог с той же статьёй, затем выберите достаточный резерв ответа.

Весь текст этого опыта посчитан o200k_base, включая ваши правки. Части считаются отдельно; служебные метки ролей не включены. 12/16/32 тысячи — выбранные рабочие пределы опыта.

Полный вход четвёртого запроса: 13 461 токен в o200k_base. С резервом 2 048 токенов требуется 15 509. В 12 000 не помещается. Новый диалог с инструкцией, всей статьёй и тем же вопросом: 10 427 токенов; с резервом 1 536 — 11 963.

Сохраните материал, необходимый для задачи, и оставьте место для результата. Попробуйте новый диалог и сравните состав входа.

Интерактивный эксперимент — в электронной версии урока.

Заметка

«Модель помнит» обычно означает, что приложение снова передало ей нужные сведения. Видимая на экране переписка может быть длиннее фактического входа.

Для слайдов по всей статье нужен её полный текст. Прежние объяснения можно убрать, начав новый диалог с той же статьёй и ясным запросом. Если оставить только введение, получится меньший вход, но в нём уже не будет подробностей методов, экспериментов и приложения. Сокращение входа меняет доступные сведения.

Резерв ответа тоже имеет смысл: в сохранённом опыте сами слайды заняли 1 203 токена в o200k_base. Ограничение в 512 токенов может оборвать сопоставимый результат. В реальном сервисе дополнительно учитываются метки ролей и другие служебные данные, поэтому полезно оставлять запас. Вход и ограничение длины генерации

> Рабочий предел — 16 000 токенов. Текст входа занимает 13 461, резерв ответа — 2 048. Сколько токенов осталось в этом бюджете?

Если лимит превышен, приложение может отклонить запрос, убрать ранние сообщения, заменить их сводкой или выбрать фрагменты документа. Модель получает именно тот состав, который собрала программа. При неожиданном ответе полезно выяснить, попало ли важное условие во вход текущего обращения.

3. Объём документов и книг

Размер документа нельзя надёжно угадать по числу страниц: короткое задание, презентация с изображениями и плотная научная статья устроены по-разному. Сравним три настоящих файла и три книги из хэндбука. Переключите подборку и шкалу: одна и та же книга займёт совсем разную долю выбранного объёма.

Сколько места занимает материал



История математики

Мехмат МГУ · 1 страница · OCR

692

0,5 % шкалы



Правила работы с ИИ

А. Бордунос · ВШМ СПбГУ · 14 слайдов · OCR

2 202

1,7 % шкалы



LLMs are Zero-Shot Reasoners

Kojima et al. · NeurIPS 2022 · 42 страницы

36 749

28,7 % шкалы

Извлечённый текст трёх конкретных файлов · o200k_base · 13 сентября 2026. Для задания и слайдов использовано OCR; объём обработки самих изображений не входит.

Сравните документы и книги на одной шкале. Красная полоса означает, что один материал уже превышает выбранный объём.

Интерактивный эксперимент — в электронной версии урока.

Для документов 13 сентября 2026 года извлекли текст из PDF и посчитали его целиком. В задании по истории математики и слайдах о правилах работы с ИИ часть текста была внутри изображений — понадобилось распознавание. В научную статью вошли библиография и приложения. Поэтому 42 страницы дали **36 749 токенов**, а четырнадцать слайдов — **2 202**.

Книги занимают уже сотни тысяч токенов: первый том Calculus — 243 494, русский текст «Преступления и наказания» — 366 530, английский перевод «Войны и мира» — 765 705. Это конкретные тексты из OpenStax, Викитеки и Project Gutenberg, посчитанные o200k_base. Иллюстрации не учитывались; формулы после извлечения из PDF могли сохраниться не полностью.

В окне на миллион токенов «Война и мир» займёт примерно три четверти. При каждом новом вопросе весь переданный том снова становится входом. Для узкого вопроса полезнее найти несколько относящихся к делу фрагментов, сохранить ссылки на их место в книге и передать их вместе с вопросом. К поиску по источникам мы ещё вернёмся.

> Файл сохранён на компьютере, но приложение его не читало и не передавало модели. Можно ли считать его частью контекста?
>
> ○ Да, если имя файла короткое
>
> ○ Да, если окно большое
>
> ○ Нет, одного наличия файла недостаточно
>
> ○ Да, модель видит все файлы

4. Рабочий предел и сохранение деталей

Большое окно ещё не означает, что модель хорошо использует всё, что в него положили. Она может пропустить существенную деталь, неверно связать фрагменты или ответить неполно. Поэтому «фрагмент не попал во вход» и «модель неверно использовала переданный фрагмент» — разные ситуации.

Если длинный разговор заменили сводкой, подробности могли исчезнуть. Для важной работы исходники, требования и принятые решения лучше хранить отдельно и при необходимости передавать заново. Числа, даты и исключения в ответе проверяют по исходному документу.

Различайте три характеристики: **словарь** перечисляет возможные токены, **контекстное окно** ограничивает длину доступной последовательности, **лимит ответа** задаёт разрешённую длину генерации. Память приложения — ещё один механизм: сохранённые сведения становятся доступными, когда программа снова добавляет их во вход.

> Весь документ поместился в контекст. Гарантирует ли это правильный ответ на любой вопрос по нему?
>
> ○ Да, если убрать ссылки
>
> ○ Нет: модель может неверно использовать доступные сведения
>
> ○ Да, размер окна доказывает понимание
>
> ○ Да, если документ длинный

Что проверить при неожиданном ответе

Сначала выясните, был ли нужный фрагмент передан модели. Затем сопоставьте ответ с самим источником.

Для каждого ответа приложение заново собирает инструкции, материалы, историю и новый вопрос.

Контекст имеет рабочий предел; будущему ответу тоже нужно место.

Сокращая вход, проверяйте, какие сведения исчезли. Полный документ можно сохранить и при новом диалоге.

Наличие текста в контексте помогает ответить по нему, но не гарантирует правильного использования.

Домашняя работа

1. Рабочий предел — 16 000 токенов. Вы сохранили весь вход объёмом 13 461 токен и оставили 1 536 на ответ. Сколько места осталось в бюджете?

Источники

- Статья из записанного диалога (https://arxiv.org/abs/2004.08981)
- Вход и ограничение длины генерации (https://huggingface.co/docs/transformers/llm_tutorial)