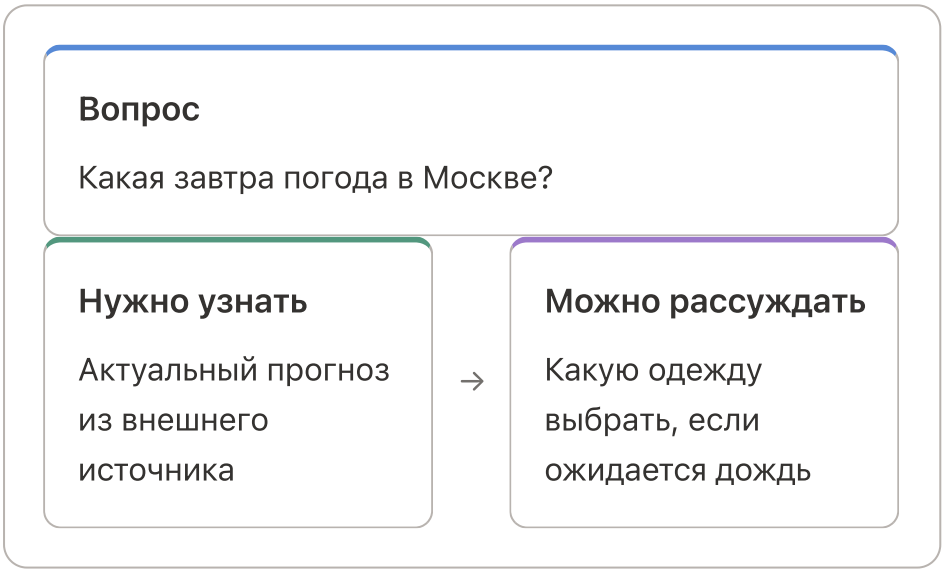
Галлюцинации языковых моделей

Помощник назвал кабинет 204, хотя в документе номера нет. Звучит так же уверенно, как и верные детали. Как отличить одно от другого?

В документе школьного кружка написано: «Встречаемся по средам в 16:00. Кабинет сообщат при записи». Помощник отвечает: «Приходите в среду в 16:00 в кабинет 204». Первые две детали совпадают с документом. Номер кабинета звучит столь же уверенно, но откуда он взялся?

1. Правдоподобный ответ и отсутствующие сведения

Предположим, нужно решить, что надеть завтра в поездку. Если известны температура, осадки и план дня, о подходящей одежде можно рассуждать сколько угодно тщательно. Но сами сведения о завтрашней погоде рассуждением не добыть — их нужно взять из прогноза.



Убедительное рассуждение не подтверждает факт. Сначала нужны сведения о погоде.

Промежуточные шаги не создают отсутствующих наблюдений. Без источника модель способна выстроить убедительное объяснение вокруг выдуманной температуры. Такое утверждение и называют галлюцинацией: оно выглядит как факт, хотя основания у него нет. Обзор галлюцинаций в генерации текста

Поэтому работа начинается с вопроса, каких данных не хватает. Сначала получить прогноз, потом рассуждать об одежде. Поиск источника и рассуждение по нему решают разные задачи.

Ранее

Ранее мы разобрали: контекстное окно. Здесь используем эти идеи в новой задаче.

← Контекстное окно

Галлюцинациями языковых моделей называют ошибочные или вымышленные сведения, поданные как фактический ответ. Это может быть несуществующая публикация, неверная дата или деталь, которую источник не подтверждает. Грамматически правильная фраза может содержать такую ошибку.

В объявлении из начала урока номер 204 не указан. По имеющимся данным нельзя утверждать ни что кабинет действительно 204, ни что он точно другой. Можно установить более узкий факт: помощник добавил неподтверждённую деталь.

Если модель пишет рассказ по просьбе придумать вымышленную школу, кабинет 204 уместен как часть вымысла. Проблема возникает, когда придуманный элемент предлагается как достоверная информация о реальной ситуации.

> **Заметка**
>
> Достоверность проверяют по отдельным утверждениям и источникам; уверенный стиль не заменяет подтверждения.

У языкового предсказания нет встроенного требования приложить доказательство к каждому следующему слову. В текстах рядом со временем встречи часто встречается место. Поэтому грамматически естественное продолжение может заполнить пробел там, где честный ответ должен сохранить неизвестность.

Такую ошибку легко не заметить, если проверять впечатление от ответа целиком. Две верные детали создают доверие к третьей. Полезнее остановиться на каждом числе, имени и условии по отдельности. Даже короткая фраза иногда содержит несколько утверждений с разными основаниями.

> Документ указывает день встречи, но не кабинет. Ответ добавляет кабинет 204. Что можно установить по документу?
>
> ○ Кабинет точно 204
>
> ○ Весь ответ обязательно выдуман
>
> ○ Номер кабинета не подтверждён
>
> ○ Кабинет точно не 204

2. Проверка отдельных утверждений

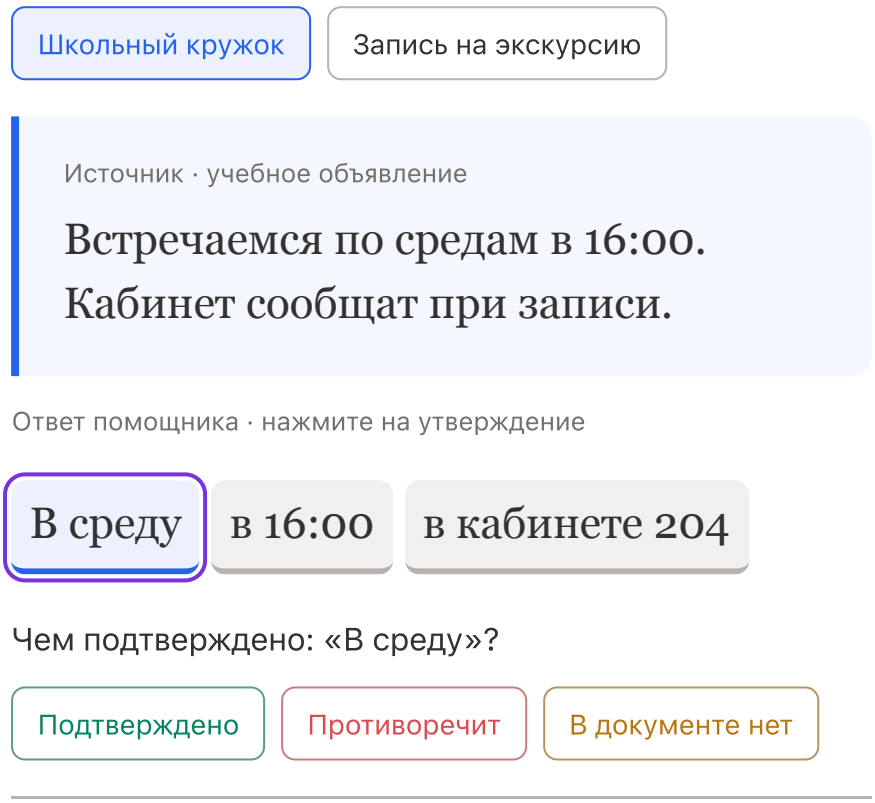
Разделите ответ на проверяемые утверждения. В нашем примере их три: день, время и кабинет. До открытия разбора найдите фразу документа, подтверждающую каждое из них. Для третьего утверждения такой фразы нет.

Убедительное оформление не помогает: ссылка может вести на существующую страницу, которая не содержит заявленного факта. Поэтому проверяют не только наличие

ссылки, но и содержание источника. Если важна актуальность, нужно также сверить дату и предмет сообщения.

В опыте все документы и ответы учебные. Здесь можно установить поддержку каждой детали без внешнего поиска. В реальной задаче может понадобиться независимый источник или уточнение у организатора.

Проверьте каждую деталь



Выберите утверждение и решите, что именно говорит источник. Затем разберите второе объявление.

Интерактивный эксперимент — в электронной версии урока.

Ссылка может быть неправильной тремя способами. Страница вообще не существует; страница существует, но не содержит нужного утверждения; утверждение там есть, однако относится к другому событию или устаревшей версии правил. Во всех случаях один значок ссылки рядом с ответом ничего не решает.

Другая ловушка — усиление формулировки. Источник говорит «заявку планируют рассмотреть», а ответ сообщает «заявка одобрена». Совпадают участники и тема, но смысл

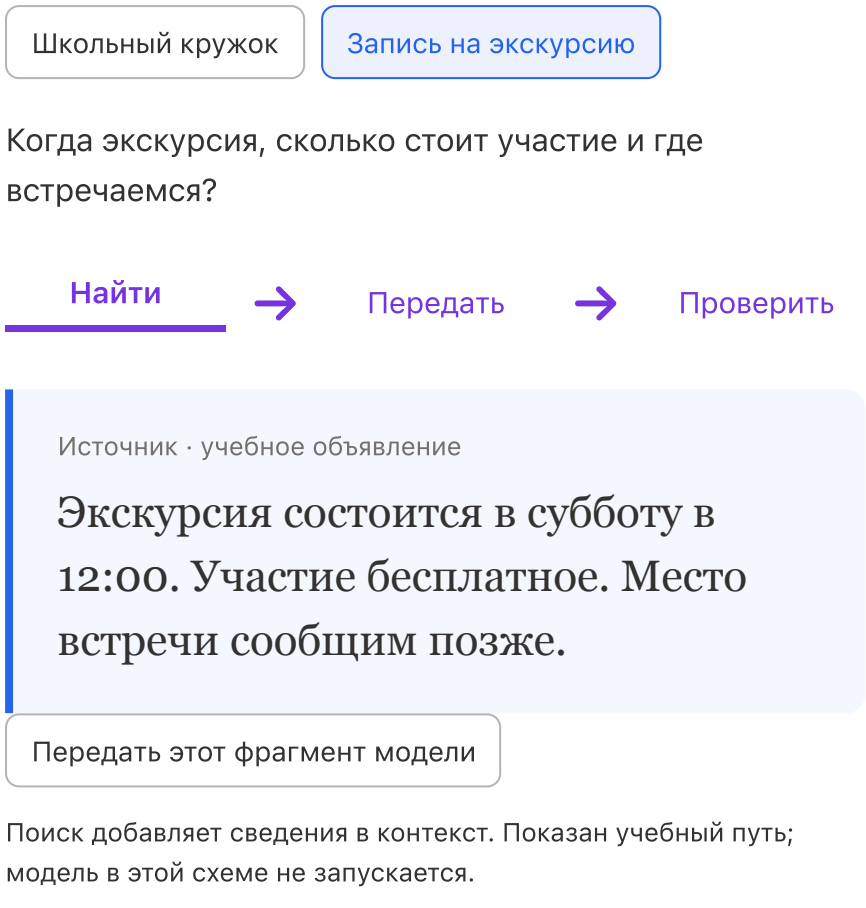
изменился. При сопоставлении важны не только названия и числа, но и оговорки: «возможно», «не позднее», «при условии».

> Ответ содержит ссылку на настоящий сайт. Что нужно сделать для проверки конкретной даты в ответе?
>
> ○ Увеличить температуру генерации
>
> ○ Открыть источник и найти подтверждение этой даты
>
> ○ Проверить только красивый адрес
>
> ○ Считать любую ссылку доказательством

3. Поиск источника и проверка рассуждения

Добавить сведения к вопросу можно с помощью поиска. Сначала система находит подходящий документ, затем добавляет его к вопросу, после чего модель формулирует ответ. Сокращение RAG означает генерацию, дополненную поиском. Для нашего опыта возьмём небольшое учебное объявление, чтобы проверить каждую деталь без внешних знаний.

От источника к ответу



Передайте объявление в контекст и сопоставьте ответ с его содержанием.

Интерактивный эксперимент — в электронной версии урока.

Пройдите путь от вопроса до ответа с источником. Обратите внимание: в схеме нет обновления весов. Найденное объявление становится частью **контекста**. Модель получает недостающие сведения в текущем обращении, а не переучивается ради одной экскурсии.

Хороший ответ сохраняет две подтверждённые детали: субботу и 12:00. Место встречи он оставляет неизвестным, поскольку организатор обещал сообщить его позже. Это полноценный полезный ответ в пределах доступных сведений. Придуманное место сделало бы его внешне удобнее и одновременно менее надёжным.

Но добавленный документ не гарантирует успеха. Поиск может принести прошлогоднее объявление, модель — пропустить оговорку, а ссылка — указывать на всю страницу без нужного фрагмента. Проверять нужно весь путь от документа к конкретному утверждению.

Если сведений не хватает, следующий шаг зависит от задачи: найти актуальное объявление, уточнить у организатора или явно сообщить границу ответа. Повторная просьба «подумай лучше» сама по себе не добавляет оснований. Длинное рассуждение и уверенный тон тоже не создают недостающий документ.

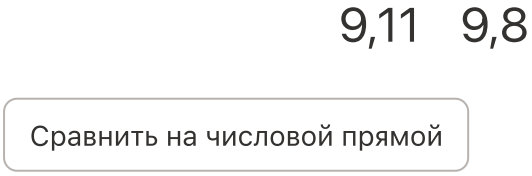
Разделение текста на утверждения помогает и при редактировании собственного материала. Можно сохранить подтверждённую часть, убрать неподтверждённую и написать, где получить недостающие данные. Исправление тогда опирается на источник, а не на новое правдоподобное предположение.

Многие задачи требуют нескольких связанных действий: выделить условия, выполнить вычисления, сопоставить варианты, проверить вывод. Для этого у современных моделей есть режим reasoning — промежуточное рассуждение перед окончательным ответом. Модель генерирует дополнительные токены, пробует варианты решения и исправляет собственные ошибки. Параметры при этом не меняются: дополнительная работа выполняется уже обученной моделью и оплачивается временем и токенами.

Но даже самое длинное рассуждение не даёт сведений, которых нет ни в параметрах, ни во входе. Поэтому дальше мы будем различать две трудности: сложно решить и неоткуда узнать. Сравнить десятичные дроби — первая. Узнать завтрашнюю погоду — вторая.

Возьмём вопрос: «Что больше — 9,11 или 9,8?» Ошибиться легко: одиннадцать больше восьми, и ответ «9,11» выглядит правдоподобно. Полезный промежуточный шаг — записать второе число как 9,80. Целые части равны, а 80 сотых больше 11 сотых, значит, 9,8 больше. Промежуточная запись делает видимым то самое место, где десятичную дробь путают с парой целых чисел.

9,11 или 9,8?



Сопоставьте разряды и проверьте результат на числовой прямой.

Интерактивный эксперимент — в электронной версии урока.

Рассуждающая модель может построить промежуточное решение и затем дать краткий ответ. Цепочку промежуточных шагов, записанную текстом, называют Chain of Thought, CoT. В учебной схеме показан возможный разбор задачи. Это пример проверяемого решения, а не запись реального диалога.

Обе части создаются одним и тем же способом — последовательной генерацией токенов. Поэтому рассуждение нельзя считать протоколом внутренних вычислений модели: это тоже текст, он может содержать ошибки и нуждается в проверке так же, как ответ. Многие приложения к тому же показывают его не целиком, а только статус или краткое резюме. Механизм генерации текста

Для проверки просите **краткое обоснование, проверяемые промежуточные величины и способ контроля**, а не пытайтесь восстановить внутренний процесс модели. В примере достаточно записи 9,8 = 9,80 и сравнения сотых. Такой результат можно проверить независимо от того, как он был получен.

4. Ответ при недостатке сведений

Вернёмся к исходному объявлению кружка. Подходящий ответ по этому документу: «Встреча в среду в 16:00; кабинет уточните при записи». Он сохраняет известное и явно отмечает недостающее. Отказ выдумывать номер не мешает сообщить полезную часть.

Просьба «не ошибайся» не превращает генерацию в проверку фактов. Модель может повторить ошибку после уточнения или предложить другой неподтверждённый вариант. Само исправление формулировки ещё не свидетельствует о новом основании.

Передача источника в контекст помогает, но модель всё равно может неверно его прочитать или добавить сведения от себя. Поэтому итоговую оценку строят по конкретным утверждениям. Вопрос «красиво ли звучит?» подходит для стиля; вопрос «чем подтверждено?» нужен для достоверности.

> Пользователь попросил придумать рассказ о вымышленном кружке. Модель придумала кабинет. Нужно ли считать это фактической галлюцинацией?
>
> - ○ Да: модель обязана использовать только реальные кабинеты
> - ○ Нет: поэтому её фактические ответы всегда верны
> - ○ Да: любая новая деталь запрещена
> - ○ Нет: вымысел соответствует поставленной задаче

Достоверность проверяют по отдельным утверждениям и источникам; уверенный стиль не заменяет подтверждения.

Отсутствие подтверждения не доказывает противоположный факт.

Ожидаемый вымысел отличается от неподтверждённого фактического ответа.

Домашняя работа

1. В объявлении: «Олимпиаду планируют провести 18 мая; окончательную дату подтвердят позже». Помощник написал: «Олимпиада точно состоится 18 мая». Как нужно исправить ответ, опираясь только на объявление?

   Указать 18 мая как планируемую дату, которая ещё требует подтверждения

   Оставить слово «точно»: дата уже названа

   Заменить дату на 19 мая

   Написать, что олимпиаду отменили

Источники

- RAG: исходная работа Lewis et al. (https://arxiv.org/abs/2005.11401)
- Механизм генерации текста (https://huggingface.co/docs/transformers/llm_tutorial)
- Обзор галлюцинаций в генерации текста (https://arxiv.org/abs/2202.03629)

Повторить эту тему