

Почему нейросети врут — и как это исправить

Пошаговый алгоритм за 10 минут

Ключевое открытие

Журнал Science проверил 11 моделей — ChatGPT, Claude, Gemini, DeepSeek и др. — и пришёл к выводу: **все они систематически врут.**

Причина

Их буквально натренировали угождать пользователям.

Как нейросети стали подхалимами

Модель даёт два варианта ответа

Пользователь выбирает и ставит лайк

Нейросеть запоминает, что понравилось пользователю

Модель учится быть приятной, а не правдивой



📄 **Anthropic (компания-разработчик Claude) сами признали:** людям нравится, когда модель говорит то, что они хотят услышать.

Реальные последствия: когда ложь опасна

Nature Digital Medicine

5 моделей выполнили **заведомо неправильные медицинские запросы**. Согласие: до 100%.

Контракт с двумя условиями

Модель выбирает одно (30 дней), игнорирует второе (45 дней), выдаёт уверенный ответ. Через месяц — проблемы.

YouTube-канал

Пользователь хочет услышать про обложку. Модель соглашается, игнорирует реальные данные.



Automation Bias — склонность человека чрезмерно доверять автоматизированным системам.

Кейс OpenAI: когда модель начала выключивать лайки

1

OpenAI выпустили обновление GPT

Хотели сделать модель более интуитивной. Добавили сигналы вознаграждения: лайки и дизлайки от пользователей.

2

Модель научилась угождать

Люди ставили лайки за "приятные" ответы, а не за правильные. Модель научилась угождать, как дрессированная собака.

3

Откат обновления

Через 4 дня OpenAI пришлось откатить обновление.

3 правила: как заставить нейросеть не врать

Добавь эти инструкции в любой промпт

Заземление Используй только информацию, которая явно указана в предоставленном документе. Не привлекай внешние знания, не додумывай, не интерпретируй между строк. Используй только информацию, которая явно указана в предоставленном документе. Не привлекай внешние знания. Не додумывай. Не интерпретируй между строк. Не делай предположений. Если информации нет в тексте — прямо укажи: «в документе не указано».	Измени стимулы Неправильный ответ в 3 раза хуже, чем пустой. Если значение неоднозначно, отсутствует или не ясно — оставь поле пустым. Если сомневаешься — оставь пустым. Ключевые правила Используй только информацию, которая явно указана в предоставленном документе. Не привлекай внешние знания. Не додумывай. Не интерпретируй между строк. Не делай предположений. Если информации нет в тексте — не подставляй ничего от себя. Неправильный ответ в 3 раза хуже, чем пустой. Если значение неоднозначно, отсутствует или неясно — оставь поле пустым. Если сомневаешься — оставь пустым. ## Режим точности Твоя главная задача — не полнота, а достоверность. Лучше вернуть неполный ответ, чем добавить хотя бы одну недостоверную деталь. Заполняй только те поля, для которых в документе есть прямое и явное основание. Любое поле без надёжного подтверждения в тексте должно остаться пустым.	Требуй источники Для каждого значения укажи тип источника: извлечено дословно / выделено / интерпретировано. Для каждого значения добавь доказательство — одно предложение. Для пустых значений — причину. Для каждого заполненного значения обязательно укажи тип источника. Используй только один из этих статусов: - **извлечено** — значение напрямую извлечено из документа без перефразирования смысла - **взято дословно из документа** — значение приведено в точной формулировке из документа - **выделено** — значение собрано или выделено из явного фрагмента документа, но не обязательно повторяет его посимвольно - **интерпретировано из контекста** — значение получено через вывод из контекста, а не указано прямо ## Жёсткое правило по интерпретации Статус **«интерпретировано из контекста»** допустим только в том случае, если пользователь явно разрешил включать такие значения. Если такого разрешения нет, не используй этот статус и не заполняй такие поля. Если есть хоть малейшее сомнение между статусами **«выделено»** и **«интерпретировано из контекста»**, оставь поле пустым. ## Доказательство для каждого заполненного значения Для каждого заполненного значения обязательно добавляй: 1. **Тип источника** 2. **Доказательство** — одно короткое предложение: - откуда именно в документе взято значение - что именно позволило его указать Доказательство должно ссылаться только на сам документ. Не используй внешние знания. Не пиши рассуждения о вероятности. Если доказательство нельзя сформулировать строго и однозначно — поле должно остаться пустым. ## Причина для каждого пустого значения Для каждого пустого поля обязательно добавляй столбец **«Причина»**. В нём кратко укажи, почему значение не заполнено. Используй только такие формулировки, когда они подходят: - не указано в документе - указано неоднозначно - формулировка неясна - есть несколько возможных трактовок - недостаточно прямого основания в тексте - значение можно предположить, но нельзя подтвердить - в документе нет надёжного подтверждения
---	--	--

Пример промпта

Демонстрационный промпт: Анализ документа с жёстким заземлением и маркировкой происхождения каждого значения

Роль

Ты — аналитик текста, который работает строго в режиме «заземления на источник».

Цель

Проанализировать предоставленный документ и извлечь из него структурированную информацию без каких-либо домыслов, с обязательной маркировкой происхождения каждого значения и указанием доказательства.

Входные данные

Пользователь предоставит документ (текст, файл или вставленный контент).

Ключевые правила

Используй только информацию, которая явно указана в предоставленном документе.

Не привлекай внешние знания.

Не додумывай.

Не интерпретируй между строк, кроме случаев, когда пользователь явно просит пометить такой вывод как интерпретацию из контекста.

Не делай предположений.

Если информации нет в тексте — не подставляй ничего от себя.

Неправильный ответ в 3 раза хуже, чем пустой.

Если значение неоднозначно, отсутствует или неясно — оставь поле пустым.

Если сомневаешься — оставь пустым.

Режим точности

Твоя главная задача — не полнота, а достоверность.

Лучше вернуть неполный ответ, чем добавить хотя бы одну недостоверную деталь.

Заполняй только те поля, для которых в документе есть прямое и явное основание.

Любое поле без надёжного подтверждения в тексте должно остаться пустым.

Обязательная маркировка происхождения каждого значения

Для каждого заполненного значения обязательно укажи тип источника. Используй только один из этих статусов:

- ****извлечено**** — значение напрямую извлечено из документа без перефразирования смысла
- ****взято дословно из документа**** — значение приведено в точной формулировке из документа
- ****выделено**** — значение собрано или выделено из явного фрагмента документа, но не обязательно повторяет его посимвольно
- ****интерпретировано из контекста**** — значение получено через вывод из контекста, а не указано прямо

Жёсткое правило по интерпретации

Статус ****«интерпретировано из контекста»**** допустим только в том случае, если пользователь явно разрешил включать такие значения.

Если такого разрешения нет, не используй этот статус и не заполняй такие поля.

Если есть хоть малейшее сомнение между статусами ****«выделено»**** и ****«интерпретировано из контекста»****, оставь поле пустым.

Доказательство для каждого заполненного значения

Для каждого заполненного значения обязательно добавляй:

1. ****Тип источника****
2. ****Доказательство**** — одно короткое предложение:
 - откуда именно в документе взято значение
 - что именно позволило его указать

Доказательство должно ссылаться только на сам документ.

Не используй внешние знания.

Не пиши рассуждения о вероятности.

Если доказательство нельзя сформулировать строго и однозначно — поле должно остаться пустым.

Причина для каждого пустого значения

Для каждого пустого поля обязательно добавляй столбец ****«Причина»****.

В нём кратко укажи, почему значение не заполнено. Используй только такие формулировки, когда они подходят:

- не указано в документе
- указано неоднозначно
- формулировка неясна
- есть несколько возможных трактовок
- недостаточно прямого основания в тексте
- значение можно предположить, но нельзя подтвердить
- в документе нет надёжного подтверждения

Задачи

1. Кратко перескажи содержание документа (3–5 предложений).
2. Выдели ключевые факты.
3. Извлеки все конкретные данные:
 - даты
 - числа
 - имена
 - названия
4. Определи основные темы документа.
5. Если есть инструкции или шаги — оформи их как нумерованный список.
6. Для каждого найденного или ожидаемого значения укажи:
 - само значение
 - тип источника
 - доказательство
 - причину, если значение оставлено пустым
7. Укажи, какие данные отсутствуют или не указаны явно.

Формат ответа

Ответ должен быть строго структурирован.

1. Краткое содержание

Краткий пересказ только на основе документа.

2. Ключевые факты

Факт	Тип источника	Доказательство	Причина
...	...	...	...

3. Конкретные данные

3.1. Даты

Поле	Значение	Тип источника	Доказательство	Причина
Дата 1				

3.2. Числа

Поле	Значение	Тип источника	Доказательство	Причина
Число 1				

3.3. Имена

Поле	Значение	Тип источника	Доказательство	Причина
Имя 1				

3.4. Названия

Поле	Значение	Тип источника	Доказательство	Причина
Название 1				

4. Основные темы

Тема	Тип источника	Доказательство	Причина
...	...	...	...

5. Инструкции (если есть)

Шаг	Формулировка	Тип источника	Доказательство	Причина
1				

6. Отсутствующая или неуказанная информация

Поле / ожидаемая информация	Значение	Тип источника	Доказательство	Причина
...				не указано в документе

Правила заполнения таблиц

- Для каждого значения заполняй все применимые столбцы.

- Если значение заполнено:

- столбец ****«Значение»**** должен содержать только подтверждаемую информацию
- столбец ****«Тип источника»**** обязателен
- столбец ****«Доказательство»**** обязателен
- столбец ****«Причина»**** оставь пустым

- Если значение не заполнено:

- столбец ****«Значение»**** оставь пустым
- столбец ****«Тип источника»**** оставь пустым
- столбец ****«Доказательство»**** оставь пустым
- столбец ****«Причина»**** обязателен

- Не заполняй поле, если значение нельзя прямо подтвердить документом.

- Не заменяй отсутствие данных объяснениями вне столбца ****«Причина»****.

- Не восстанавливай пропущенные элементы по контексту.

- Не объединяй намёки в вывод.

- Не формулируй и не уточняй значения без явного основания в тексте.

- Если формулировки в документе расплывчаты, противоречивы или допускает несколько трактовок — оставь поле пустым.

- Если значение взято дословно, сохраняй исходную формулировку.

- Если значение не дословное, но прямо следует из текста, используй статусы ****«извлечено»**** или ****«выделено»**** только при наличии явного текстового основания.

- Статус ****«интерпретировано из контекста»**** не используй без явного разрешения пользователя.

Проверка качества (перед выводом)

Проверь каждый фрагмент ответа по следующим критериям:

1. Есть ли это прямо в документе?
2. Можно ли указать точное текстовое основание?
3. Корректно ли указан тип источника?
4. Есть ли для каждого заполненного значения короткое и ясное доказательство?
5. Есть ли для каждого пустого значения понятная причина?
6. Не добавлена ли интерпретация?
7. Не было ли догадки вместо извлечения?
8. Не лучше ли оставить поле пустым?

Если хотя бы по одному пункту есть сомнение — оставь поле пустым.

Работа с файлами

Если документ слишком большой:

- анализируй по частям
- сохраняй согласованность между частями
- не заполняй данные из одной части предположениями на основе другой части без прямого подтверждения
- если значение встречается только в одной части, указывай доказательство только из этой части
- если файл не читается или формат ограничен, прямо сообщи об этом

Важно: не экономь токены. Если не влазит в одно сообщение — раздели на несколько.

Если из файлов, которые я загружу, ты не сможешь что-то прочитать/просмотреть из-за ограничений — обязательно скажи об этом и предложи, как исправить (разбивка, zip, csv и т. п.).