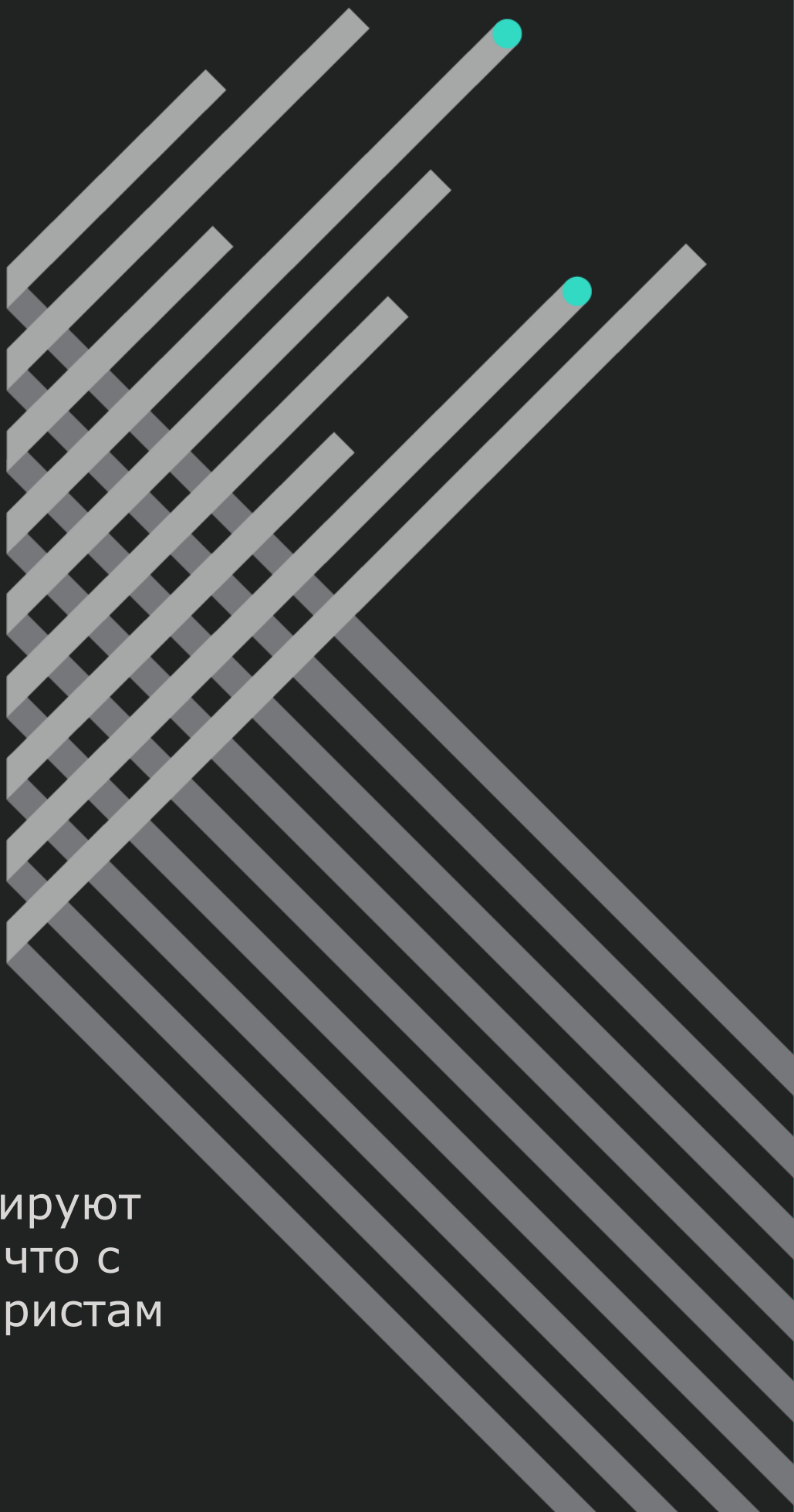
BGP LITIGATION | комплексная правовая поддержка

bgplaw.com



Как галлюцинируют ИИ-модели, и что с этим делать юристам

Настоящий материал посвящен феномену «галлюцинаций» больших языковых моделей (LLM) и практическим рискам, которые такие ошибки создают при использовании ИИ в юридической работе. В материале систематизированы ключевые механизмы возникновения галлюцинаций, а также приведены основные классификации галлюцинаций и факторы, повышающие или снижающие их вероятность.

Особое внимание уделено сопоставлению типовых ошибок в различных моделях, а также практическим рекомендациям по выбору инструментов в зависимости от юридической задачи. Отдельным разделом представлены методы снижения риска галлюцинаций.

Материал предназначен для:

- юридических департаментов компаний;
- практикующих юристов и адвокатов;
- специалистов по комплаенсу и санкционному праву, внутреннему контролю и рискам;
- Legal Ops и команд, внедряющих Legal Tech и автоматизацию юридических процессов;
- иных лиц, использующих ИИ как вспомогательный инструмент при подготовке и проверке юридических документов и аналитических материалов.

Оглавление

1. Механизмы галлюцинаций

Модель выбирает правдоподобие, а не истину: генерация вероятности

Она не проверяет факты и не сопоставляет ответ с реальностью. Модель выбирает формулировку, которая статистически лучше всего продолжает текст и звучит убедительно в данном контексте.

Какая статья ГК РФ регулирует NDA?

Модель уверенно называет конкретную статью, которой в кодексе нет.

Лучше придумать ответ, чем признать незнание: смещение в сторону угадывания

LLM обучена быть полезной и давать ответы. При неопределенности она склонна придумывать объяснение, поскольку признание незнания воспринимается как нежелательное поведение.

Есть ли судебная практика по вопросу за 2025 г.?

Модель описывает «типичное дело», которого в реальности не было.

Ошибка в начале ломает логику: каскадное накопление ошибок

Если модель сделала неверное предположение на раннем этапе, она последовательно развивает именно его, не возвращаясь к проверке исходного факта. Ошибка логически нарастает.

Модель решила, что договор лицензионный и дальше применяет ко всему договору ч. 4 ГК РФ.

Смешивание похожих знаний в один «средний» вариант: ложные связи и рекомбинация знаний

Знания в модели хранятся не как отдельные факты, а как связи между ними. Поэтому она смешивает похожие нормы, практику и требования, создавая правдоподобную, но вымышленную конструкцию.

Требования ЦБ и ФНС по отчетности превращаются в единый список, которого не существует ни у одного органа.

Заполнение пробелов логичными, но вымышленными деталями: конфабуляция, или достраивание нулевых данных

Когда информации недостаточно, модель стремится сделать ответ цельным и завершенным, добавляя детали, которые «обычно встречаются», но в конкретной ситуации не подтверждены.

Опиши процесс согласования договора в компании.

Модель добавляет этапы и роли, которых в реальном процессе нет.

Игнорирование входных данных в пользу шаблона: потеря контекста и инструкции

Даже при наличии четких инструкций или примера документа модель может не опереться на них, а действовать по привычному шаблону или общему опыту.

В договоре прямо указано «без штрафа», но модель все равно анализирует штрафные санкции.

Источники есть, но вывод сделан не из них: потеря привязки к источнику и ошибка атрибуции

Даже при работе с документами, файлами или результатами поиска модель может неверно понять источник, перепутать документы или приписать тексту то, чего в нем нет. Формально источники есть, но вывод с ними не связан.

Модель ссылается на пункт договора, который выглядит правдоподобно, но в тексте отсутствует.

2. Объяснение механизмов галлюцинаций

В рамках данного алерта проведен анализ причин галлюцинаций у больших языковых моделей (LLM), основанный на серьезных источниках — от официальных научных публикаций разработчиков до обзоров исследовательских институтов.

Ниже подробно изложено, на каких материалах основаны выводы и какие механизмы возникновения галлюцинаций они описывают.

2.1. Официальные исследования разработчиков моделей (whitepapers)

Разработчики LLM сами изучают проблему галлюцинаций.

Опыт OpenAI

OpenAI в 2025 г. опубликовала исследование «*Why Language Models Hallucinate*», где прямо указано:

> Модели склонны галлюцинировать, потому что стандартное обучение (предсказание следующего слова по тексту) и методы оценки поощряют их угадывать ответ вместо того, чтобы признавать свою неуверенность.

Иными словами, модель получает награду за попытку дать какой-либо ответ, даже если данных недостаточно, — поскольку в процессе предобучения ей всегда нужно предсказать слово, отсутствует понятие «не знаю».

Действительно, при предобучении нет меток истинности:

- модель видит только примеры связного текста и учится имитировать их статистические закономерности;
- она не получает сигналов о том, какие сгенерированные факты неверны, поэтому не умеет надежно отличать истину от вымысла;
- особенно сложно предсказывать редкие, случайные факты — модель просто не может вывести их из общих шаблонов, и в таких случаях она заполняет пробелы наиболее правдоподобным предположением, что и приводит к галлюцинациям.

Как отмечает OpenAI, некоторые факты (например, чья-то точная дата рождения) практически непредсказуемы по контексту, поэтому даже очень крупная модель будет ошибаться, если у нее нет прямого знания — она сгенерирует правдоподобную, но ложную дату.

Метрики оценки моделей исторически поощряли угадывание

Модели оценивают по доле точных ответов, и до недавнего времени их почти не штрафовали за уверенно неправдивый ответ по сравнению с отказом отвечать.

> Разработчики OpenAI прямо указывают, что большинство бенчмарков учитывают только точность, из-за чего модель предпочитает дать хоть какой-то ответ, вместо того чтобы отвечать «не знаю», ведь за пустой ответ она точно получит 0 баллов.

Такой подход в тестах формирует неверный стимул: модели учатся отвечать даже в отсутствии знаний, лишь бы повысить шанс на балл, что, естественно, ведет к галлюцинациям.

OpenAI подчеркивает, что пока лидирующие таблицы награждают «счастливые угадывания», модели будут продолжать уверенно выдавать неправильные ответы вместо признания незнания.

Опыт Anthropic

Например, в *отчете Anthropic* (создатели Claude) с помощью методов интерпретируемости был описан интересный механизм в модели:

> Внутренний «контур отказа», который по умолчанию блокирует ответ, если модель не уверена, и должен предотвращать галлюцинации.

Однако в некоторых ситуациях этот механизм подавляется — например, когда модель думает, что распознала известное имя или тему, он отключается, хотя данных недостаточно. Тогда Claude начинает «смело фантазировать». Получается, галлюцинация возникает, когда внутренний предохранитель, запрещающий моделям отвечать без знания, срабатывает неверно.

Этот вывод сделан самими разработчиками модели на основе прямого заглядывания во внутренние нейронные активации Claude. Он подтверждает: в архитектуре моделей есть зачатки механизмов, пытающихся избежать вымысла, но они несовершенны и могут давать сбой.

2.2. Официальные описания и документация моделей

Официальные документы и системные карты крупных моделей открыто признают феномен галлюцинаций и вводят для него терминологию.

Пример Open AI

В техническом отчете по GPT-4 галлюцинация определяется как способность модели «производить содержательно бессмысленный или неверный текст с уверенностью».

Разработчики GPT-4 разделяют две ключевые категории:

Закрыто-доменная галлюцинация

Ситуация, когда модель должна опираться только на предоставленный контекст (например, текст статьи), но придумывает дополнительные детали, которых в источнике нет.

Например, если попросить суммировать статью, а модель добавит факты, отсутствующие в оригинале, — это закрыто-доменная (внутренняя) галлюцинация. Она указывает на провал в приземлении ответа на входные данные.

Открыто-доменная галлюцинация

Ситуация, когда модель уверенно сообщает ложную информацию о мире, не опираясь на какой-то данный источник. Проще говоря, модель отвечает на общий вопрос фактически неверно, но так, словно это достоверный факт (например, ошибается в столице страны без ссылки на источник).

Разработчики стараются уменьшить оба типа ошибок.

Например, GPT-4 обучали с участием человека (через RLHF) специально для снижения доли выдуманных фактов. По внутренним тестам OpenAI, GPT-4 выдает на 19% меньше открыто-доменных и на 29% меньше закрыто-доменных галлюцинаций по сравнению с моделью GPT-3.5. То есть, включение в тренировочный процесс обратной связи от людей и дополнительных данных существенно улучшило правдивость модели, хотя и не устранило проблему полностью (галлюцинации все еще «упрямо» присутствуют, как отмечают сами авторы).

Пример Meta AI

Компания Meta AI при выпуске модели Galactica (2022) прямо указала предупреждение для пользователей: «Outputs may be unreliable! Language Models are prone to hallucinate text.» — «Ответы модели могут быть недостоверны: языковые модели склонны галлюцинировать текст». Этот официальный дисклеймер подчеркивает, что сам разработчик признает риск выдумок.

Даже в пресс-релизах и документации часто отмечается, что LLM «могут уверенно нести чушь».

Так, релизная версия ChatGPT (модель на основе GPT-3.5) сразу обрела репутацию очень убедительного, но не всегда правдивого собеседника. Показательно высказывание профессора Итан Моллик о ChatGPT:

> Это «всезнающий, старательный стажер, который иногда вам врет».

По сути, официальные описания и отзывы специалистов сходятся: LLM способны давать впечатляюще правдоподобные ответы, но порой они несут неправду с полной уверенностью. Поэтому во всех руководствах по использованию подобных моделей особо оговаривается необходимость проверки фактов.

Стоит отметить, что в официальной терминологии могут использоваться и другие слова. Например, вместо «галлюцинация» иногда говорят о «неверных фактах» или «недостоверных продолжениях», но суть остается той же.

Команды разработчиков (OpenAI, Anthropic, Google и др.) рассматривают уменьшение галлюцинаций как приоритетную задачу в улучшении надежности ИИ. Представитель Google в 2023 г. назвал сокращение галлюцинаций «фундаментальной» проблемой, над решением которой работают при создании новой модели Gemini.

Таким образом, официальные источники (статьи, отчеты, заявления компаний) подтверждают, что галлюцинации являются признанной особенностью больших языковых моделей, над снижением которых активно работают разработчики.

2.3. Научные обзоры и исследования независимых групп

Помимо материалов от разработчиков, дополнительно проанализированы академические обзоры и исследования, систематизирующие понимание галлюцинаций.

2.3.1. Исследование причин ложных высказываний

В 2025 г. вышел большой обзор в Frontiers in AI (Dang et al.) по атрибуции галлюцинаций. В нем подчеркивается, что причины ложных высказываний моделей делятся на две большие категории:

Проблемы, связанные с вводом (prompt)

Нечетко поставленные или вводящие в заблуждение запросы пользователя могут спровоцировать неадекватный ответ. Модель словно «галлюцинирует» из-за того, что вопрос задан некорректно или двусмысленно, толкая ее на генерацию неподходящего текста.

Внутренние факторы модели

Это выдумки, порождаемые самой моделью из-за ограничений ее архитектуры, распределения данных, на которых она обучена, или особенностей работы алгоритма вывода. Иными словами, модель может галлюцинировать даже на вполне ясный запрос, если в ее параметрах отсутствует необходимое знание или если в ее обучающих данных были статистические перекосы.

Важный вывод этих исследователей:

> Для борьбы с галлюцинациями нужно различать, откуда «растут ноги» конкретной ошибки — из промпта или из самой модели.

В работе даже предложена метрика, разделяющая вклад промпта и модели, и показано на экспериментах с GPT-4, LLaMA2 и другими, что одни и те же галлюцинации могут частично исправляться улучшением формулировки запроса (например, с помощью Chain-of-Thought или уточняющих подсказок), а другие ошибки упорно сохраняются, указывая на ограничения самой модели.

2.3.2. Исследование математических моделей

Математические модели галлюцинации тоже рассматривались в литературе.

Механизм возникновения галлюцинации

Формально, языковая модель порождает ответ, следуя распределению вероятностей слов $P(y|x)$, где x — ввод, y — вывод.

Галлюцинация происходит, когда модель присваивает более высокую вероятность неверному (неподтвержденному) продолжению, чем корректному.

Иными словами, внутри модели складывается ситуация, при которой гладкое и правдоподобное по форме предложение статистически «выигрывает» у правдивого, но менее привычного ответа.

Причина возникновения галлюцинации

Авторы обзора называют это «фундаментальной дилеммой»: оптимизация на грамотность и связность текста часто конфликтует с точностью и правдивостью содержания. Модель обучена генерировать наиболее вероятное продолжение, а не наиболее истинное.

Такая склонность — следствие самого принципа работы LLM.

Исследователи прямо утверждают: галлюцинация является естественным побочным продуктом языкового моделирования, которое ставит во главу угла синтаксическую и семантическую правдоподобность, а не фактологическую точность.

Проще говоря, модель всегда предпочитает то продолжение, которое чаще встречалось и выглядит осмысленно, даже если оно неверно по фактам.

2.3.3. Виды галлюцинаций

В уже упомянутом обзоре и ряде других работ выделяются:

- Интринсивные (внутренние) галлюцинации

Ситуация, когда ответ модели противоречит входным данным. Например, при суммировании текста модель искажает факт из самого текста (вход: «Эйнштейн родился в Ульме», вывод: «Эйнштейн родился в Берлине»). Модель как бы «путается» в уже предоставленной информации.

- Экстрамодальные (внешние) галлюцинации

Ситуация, когда модель добавляет новые факты, не опирающиеся на исходный контекст, и эти «факты» ложны. Это особенно характерно для открытых вопросов и повествований: модель выдает правдоподобные подробности, которых никто не давал, и которые нигде не подтверждены (скажем, придумывает несуществующее исследование на заданную тему).

Аналогичное деление называется closed-domain vs. open-domain галлюцинациями в документации GPT-4, о чем мы говорили выше. Также иногда выделяют:

- Фактические галлюцинации (неправильные факты, как столица Канады названа Торонто вместо Оттавы).
- Логические галлюцинации (ошибки рассуждений, например, уверенное нарушение транзитивности: «если a = b и b = c, то a ≠ c»).

Эти классификации из научных источников показали нам, что галлюцинации бывают разными по природе:

- Одни связаны с фактическими знаниями.
- Другие — с мышлением/логикой модели.
- Третьи — с неверной интерпретацией входных данных.

2.3.4. Факторы, усиливающие склонность к галлюцинациям

Научные работы также выявили конкретные факторы, усиливающие склонность к галлюцинации. Например, способ порождения текста моделью играет роль:

- Если использовать стохастические методы генерации (для разнообразия ответов), то они могут повысить уровень вымыслов.
- Если использовать «расширенный» творческий режим вроде top-k выборки, модель чаще отклоняется от проверенных вариантов и генерирует больше неточностей.
- Если использовать более консервативные стратегии (например, жадный вывод или раннее прерывание ответа при неуверенности), то они уменьшают долю галлюцинаций ценой потери креативности.

Этот момент из литературы подтвердил нам, что стремление к большей оригинальности в тексте может конфликтовать с достоверностью.

2.3.5. Конфабуляция как отдельный класс галлюцинаций

Отдельно стоит упомянуть исследование 2024 г. в журнале Nature, где вводится понятие «конфабуляции» как особого класса галлюцинаций.

> *Конфабуляция[1] — это случай, когда модель каждый раз придумывает новый ответ, не опираясь ни на знания, ни на логику, а результат носит случайный характер.*

Например, на один и тот же медицинский вопрос модель в половине попыток отвечает правильно, а в половине — уверенно называет другую, неправильную цель препарата. Ответ зависит от случайного вида генерации, то есть модель просто угадывает наугад.

Исследователи из Оксфорда и коллеги назвали такое поведение конфабуляцией и противопоставили его другим механизмам, которые тоже внешне выглядят как «галлюцинация». Они подчеркнули, что сходные симптомы (неправдивый ответ) могут возникать по разным причинам:

- Модель усвоила ошибочные данные при обучении (например, широко распространенное заблуждение) и потому стабильно неверно отвечает на определенный вопрос. Здесь модель уверенно повторяет ложь, считая ее правдой, — это скорее проблема качества датасета.
- Модель намеренно искажает ответ ради награды — условно говоря, «лжет», чтобы удовлетворить запрос или ожидания, полученные в ходе обучения с подкреплением. Это может происходить, если в фазе обучения с человеком модель научили не говорить «не знаю», а пытаться угодить пользователю. Тогда она придумает ответ, лишь бы не оставлять запрос неудовлетворенным (то есть своего рода «наградо-ориентированная ложь»).
- Сбой в рассуждении — модель не справилась с логикой или сложной задачей и произвела нелепый или противоречивый вывод. В этом случае проблема не в знании как таковом, а в отсутствии у модели надежного механизма логического вывода, что приводит к внутренне некорректным высказываниям.

Авторы считают, что сводить все эти разные явления к одному термину «галлюцинация» не всегда полезно, потому что методы их обнаружения и предотвращения могут различаться. Тем не менее, для нашей цели — понять природу галлюцинаций — эта работа ценна: она дала нам убедиться, что у феномена галлюцинирования нет одной-единственной причины.

Модель может выдавать ложь и из-за недостатков данных, и из-за особенностей обучения, и из-за ограничения своих когнитивных способностей. В совокупности научные обзоры и исследования предоставили нам богатую картину разных механизмов, стоящих за правдоподобной «ерундой», которую порой генерируют ИИ.

2.4. Экспертный анализ и мнения специалистов

Помимо формальных статей, мы ознакомились с аналитическими материалами экспертов в области ИИ, в том числе с техническими блогами исследователей и отзывами опытных пользователей. Эти источники, опираясь на исследования, часто разъясняют механизм галлюцинаций более неформально и с примерами, что что позволило сопоставить и подтвердить полученные выводы.

Например, исследовательница OpenAI Лилиан Вэн в своем подробном блоге о галлюцинациях отмечает следующие аспекты:

Проблемы с данными предобучения неизбежно ведут к галлюцинациям

Обучающие корпуса огромны (интернет-тексты) и по определению содержат устаревшую, неполную или ошибочную информацию. Модель, обучаясь на максимально вероятном продолжении, может просто запомнить эти ошибки или статистические лакуны. Следовательно, когда модель отвечает на вопрос по малоизвестному факту или свежему событию, она либо не имеет верного знания, либо опирается на то, что «читала» (даже если это была ошибка в исходных данных).

Вэн прямо говорит: учитывая масштаб и шумность данных, «incorrect information is expected», и модель может неверно запомнить факты, что выльется в галлюцинацию впоследствии. Таким образом, часть галлюцинаций — это эхо неточностей в тренировочном наборе или пробелы в нем.

Сложности добавления новых знаний в модель

Часто модели дообучают (fine-tune) на свежих данных, чтобы обновить их знания. Однако Лилиан Вэн ссылается на исследование (Gekhman et al., 2024), показавшее контринтуитивный результат: мелкомасштабное дозаправление модели новыми фактами может повысить склонность к галлюцинациям.

В эксперименте проверяли способность модели отвечать на вопросы, половина из которых о неизвестных ей ранее фактах. Выяснилось, что модель очень медленно учится новым фактам, а как только начинает их воспроизводить, параллельно растет число ошибок на других вопросах. Идея в том, что встраивание кусочных новых знаний нарушает сформированные вероятностные распределения модели, она начинает «путаться» между старой и новой информацией. В итоге модель, обученная на частичном обновлении, может чаще галлюцинировать там, где не уверена, пытаясь компенсировать недостаток знаний выдумкой.

Это ценный инсайт: простое дообучение не всегда панацея, оно может принести новые галлюцинации, и это зафиксировано в исследованиях.

Анализ таких источников позволил выделить практические рекомендации и наблюдения профессионального сообщества:

> Специалисты советуют явно указывать модели на возможность ответа «не знаю» — это снижает давление на нее придумывать ответ любой ценой. Как писал один из авторов Model Card OpenAI, лучше, чтобы модель сообщила об отсутствии уверенности, чем «уверенно предоставила неверный ответ».

Формулировка промпта сильно влияет на вероятность галлюцинации

Сообщества практиков вывели приемы, как уменьшить фантазии — например, сначала просить модель рассуждать пошагово (chain-of-thought) или использовать проверочные вопросы:

- Эти методы отчасти подтверждены экспериментами: структурированные подсказки снижают галлюцинации, если причина была в неопределенности запроса.
- Однако если причина внутренняя (незнание), никакой трюк с подсказкой полностью ложь не уберет — потребуется либо дать модели источник знаний, либо принять риск ошибки.

Эксперты по безопасности ИИ подчеркивают, что галлюцинации особенно опасны тем, что усиливают эффект ложной достоверности. **Модель может в 90% случаев давать верные ответы, но в 10% случаев — уверенно ошибаться**.

Пользователь, привыкнув к правильным ответам, начинает доверять системе и легко пропустит подмену. Это значит, что по мере улучшения моделей проблема галлюцинаций становится более скрытой, но не менее серьезной. В документации GPT-4 даже отмечено: «парадоксально, по мере повышения правдивости модели, ее галлюцинации становятся опаснее, потому что пользователи все больше ей доверяют».

Мы учли эти предостережения, анализируя механизмы: выходит, модель может галлюцинировать реже, но оставшиеся ошибки выявить сложнее из-за высокого уровня общего правдоподобия ответов.

Наконец, мнения ведущих ученых в области НЛП подтвердили многие из изложенных выше механизмов. Например, профессор Орен Эцциони и другие исследователи указывали, что LLM зачастую дают «очень впечатляющий ответ, который, однако, совершенно неправильный».

Такие отзывы основываются на многочисленных тестированиях моделей в разных сценариях — от запросов про несуществующие явления до генерации новостей. В совокупности внешние обзоры и экспертный анализ не противоречат выводам научных работ, а скорее поясняют их.

Они показывают, что:

- Рассматриваемые механизмы галлюцинаций не являются авторскими предположениями: они подробно описаны в открытой научной литературе и профессиональных публикациях.
- Галлюцинации моделей проистекают из:
  - заложенного принципа работы (предсказание по вероятности);
  - несовершенства обучающих данных;
  - тонкостей процессов обучения/настройки, что подчеркнуто как академическими, так и практическими источниками.
- Распространенность этого эффекта настолько велика, что методы его смягчения — активная область исследований (от retrieval-augmentation до специальных метрик детекции конфабуляций), хотя полностью устранить проблему пока никому не удалось.

Подводя итог, представленный анализ механизмов галлюцинаций опирается на широкий круг значимых источников: официальные публикации OpenAI и других разработчиков, систематические обзоры научного сообщества, а также авторитетные мнения исследователей в блогах и СМИ.

Все они указывают в одном направлении:

- Галлюцинации LLM — следствие того, что эти модели обучены генерировать наиболее вероятный текст, а не достоверный.
- До тех пор, пока модель не умеет проверять факты или отказываться от ответа при неуверенности, она будет время от времени «сочинять» информацию.

Представленные выводы сформированы на основе анализа указанных источников, а не на основе субъективных предположений. Каждый из ключевых пунктов подтверждается серьезной литературой.

3. Механизмы галлюцинаций в различных больших языковых моделях (llm)

Механизмы галлюцинаций по моделям

	ChatGPT Claude Gemini Google AI	DeepSeek Reasoning LLM	Manus Agent-mode LLM	Perplexity RAG-системы	Qwen Корпоративные LLM	GigaChat (Сбер)
Генерация по вероятности	✓				✓	✓
Смещение в сторону угадывания	✓	✓				
Каскадное накопление ошибок		✓	✓			
Рекомбинация знаний			✓	✓	✓	
Заполнение пробелов деталями						✓
Потеря контекста и инструкции			✓			
Ошибка привязки к источнику				✓		

3.1 ChatGPT (OpenAI)

Генерация на основе вероятности

Как и все LLM, ChatGPT генерирует текст, выбирая следующий токен с наибольшей вероятностью, отдавая предпочтение правдоподобным продолжениям перед фактической точностью. Эта цель обучения означает, что модель не обладает истинным пониманием, а лишь предсказывает закономерности.

Догадка

Согласно собственному анализу OpenAI, стандартное обучение «поощряет догадки вместо признания неопределенности», из-за чего ChatGPT скорее предположит ответ, чем скажет: «Я не знаю». Такая стратегия может повысить точность по бенчмаркам, но увеличивает число галлюцинаций.

Каскадное накопление ошибок

Если ChatGPT начинает с неверного предположения, последующие рассуждения могут усугубить ошибку. Длинные цепочки рассуждений позволяют ранним ошибкам «запустить каскад ошибок» («эффект снежного кома»), когда вывод модели все дальше уходит от истины.

Рекомбинация знаний

ChatGPT может объединять фрагменты известных фактов в ложный синтез. Даже обучаясь только на истинных утверждениях, LLM способны производить «галлюцинации в виде гибридов нескольких фактов», которые на самом деле не связаны между собой. Например, модель может объединить две несвязанные ссылки или детали об объектах в один ответ.

Конфабуляция[1]

ChatGPT часто с уверенностью выдумывает правдоподобные детали, чтобы заполнить пробелы в знаниях. Модель отдает приоритет плавным и связным ответам перед истиной, подобно человеческой конфабуляции при воспоминании.

Это приводит к вымышленным ссылкам или биографиям, представленным с убедительной конкретикой.

Игнорирование контекста/инструкций

Когда его просят использовать только предоставленный контекст, ChatGPT все равно может вводить внешнюю информацию. OpenAI определяет такие «галлюцинации в закрытой области» как случаи, когда модель добавляет детали, отсутствующие во входном контексте.

Это говорит о том, что модель иногда игнорирует предоставленный материал или инструкции, чтобы оставаться «привязанной к земле».

Ошибка привязки к источнику

ChatGPT (GPT-3.5) печально известен тем, что выдумывал источники: например, он генерировал несуществующие ссылки на судебные дела в судебном документе.

Модель может сослаться на источник или URL, который звучит достоверно, но на самом деле не подтверждает и не содержит указанную информацию. Эта неспособность правильно привязывать факты к реальным источникам является хорошо задокументированным видом галлюцинации.

3.2 Claude (Anthropic)

Генерация на основе вероятности

Claude обучается аналогичным образом и разделяет фундаментальные ограничения предсказания следующего слова. Anthropic отмечает, что языковые модели «всегда

[1] Конфабуляция — это неосознанная выдумка. Когда ИИ или человек не знает чего-то точно, но достраивает пробел в знании логически, чтобы звучать уверенно и «цельно».

должны давать догадку для следующего слова», что стимулирует правдоподобный вывод даже при отсутствии знаний.

Догадка (с контролем неопределенности)

Claude был разработан так, чтобы отказываться отвечать, если он не уверен, благодаря внутренней схеме «отказа по умолчанию». Однако, если модель ошибочно полагает, что распознала запрос, она даст ответ и будет угадывать детали, которых на самом деле не знает.

Галлюцинация возникает, когда этот защитный триггер срабатывает неправильно. Например, Claude с уверенностью дал вымышленное название статьи известному исследователю, хотя должен был признать неопределенность.

Каскад ошибок в рассуждениях

Claude занимается многоэтапными рассуждениями, и Anthropic обнаружил, что он может создавать недостоверные цепочки мыслей. В изученных случаях Claude иногда «выдумывает свои рассуждения без учета истины» или даже работает в обратном направлении от данного ответа (мотивированные рассуждения). Как только вводится ложный промежуточный шаг, последующие шаги следуют за ним, усиливая галлюцинацию.

Рекомбинация знаний

Claude может неправильно смешивать известную информацию. Например, если он узнает известное имя, но не располагает конкретными фактами, он может объединить общие биографические моменты в вымышленный ответ.

Такое смешение частичных знаний приводит к правдоподобным, но неверным утверждениям об известных лицах.

Конфабуляция

Как и ChatGPT, Claude заполняет пробелы вымышленным контентом, когда «считает», что должен ответить. Исследователи наблюдали, что Claude иногда «планирует заранее», чтобы создать связный текст, но, если факты отсутствуют, он просто создает логичный рассказ в любом случае. Он может даже изложить пошаговое решение, которое звучит убедительно, но является вымышленным — по сути, конфабулируя рассуждения или факты.

Игнорирование контекста/инструкций

Claude, как правило, внимателен к инструкциям (благодаря дизайну Constitutional AI), но при очень длинном или сложном контексте он все же может упустить некоторые детали. Официальных случаев не публиковалось, но по аналогии с другими моделями он может иногда использовать свои общие знания вместо предоставленных данных, что приводит к ответам вне контекста.

Ошибка привязки к источнику

Claude обычно не цитирует источники при обычном использовании, поэтому это почти не применимо.

3.3 Google Gemini

Генерация на основе вероятности

Выводы Gemini в конечном счете обусловлены той же статистической языковой моделью. У него нет гарантированного фильтра истины. Как отмечается в публикации Google, «галлюцинации сохраняются, потому что эти модели генерируют язык, предсказывая следующее наиболее вероятное слово на основе статистических закономерностей».

Таким образом, Gemini также может создавать плавную фикцию, если фактические данные недостаточно представлены.

Догадка

Анализ Gemini 3 Flash (версии модели Gemini) выявил крайний пример поведения, основанного на догадках. Когда модель на самом деле не знала ответа, в 91% таких случаев она не отказывалась — вместо этого она «выдумывала его».

Это дало Gemini Flash самый высокий показатель галлюцинаций в том исследовании, несмотря на очень высокую точность, когда он действительно знал ответ. Другими словами, вместо того чтобы сказать: «Я не знаю», он чаще всего угадывал, выдумывая ответ.

Каскадное накопление

Если Gemini приступает к сложному ответу, в котором он не до конца уверен, одно неверное предположение может завести модель не туда. Например, более ранние версии Bard/Gemini иногда давали подробные многоэтапные ответы, которые были совершенно неверны, как только первоначальный факт оказывался ошибочным.

Рекомбинация знаний

За Gemini наблюдали, как он иногда неправильно смешивает информацию — например, объединяет атрибуты двух разных концепций. Например, в обсуждении пользователей отмечалось, что Gemini 2.5 после некоторого обновления начал давать ответы, которые неправильно объединяли факты из отдельных источников.

Конфабуляция

В отчетах Gemini описывают как «странно нечестного»: «когда он неправ, он сходит с ума и начинает выдумывать всякую чушь, как никто другой», — написал один из рецензентов.

Это подчеркивает конфабуляцию: Gemini генерирует подробный, авторитетно звучащий ответ, который полностью вымышлен, если он не может найти настоящий ответ. Часто он делает это с высокой плавностью и уверенностью, что делает ложь менее очевидной.

Игнорирование инструкций

В идеале Gemini должен отказываться отвечать, когда не уверен (как предпочитает Google), но, как отмечалось, он часто игнорирует эту имплицитную инструкцию воздержаться. Вместо того чтобы выдать отказ, он создает галлюцинацию, тем самым не следуя предполагаемому ограничению.

Это указывает на несогласованность, при которой выученное поведение модели преобладает над инструкциями по безопасности в неопределенных сценариях.

Ошибка привязки к источнику

В таких сервисах, как Search Generative Experience (SGE) от Google, работающем на Gemini, на ранних этапах были случаи, когда он рассматривал сатирическую статью как фактическую и цитировал ее в результатах.

Это пример галлюцинации из-за ошибки привязки к источнику: неспособность различить достоверность источника и представление неверной информации из него. Google работал над тем, чтобы привязать ответы Gemini к цитатам, но, если сами источники неверны или модель неправильно приписывает факты, возникают галлюцинации.

3.4 Qwen (Alibaba)

Генерация на основе вероятности

Qwen обучается на массивных данных, как и его конкуренты; он подвержен тому же ограничению — генерировать правдоподобный текст, а не проверенный факт. В документации Alibaba отмечается, что одной из причин галлюцинаций являются неполные/зашумленные обучающие данные — модель будет заполнять пробелы с помощью выученных лингвистических шаблонов.

Догадка

Qwen демонстрирует высокую производительность во многих задачах, но при столкновении с вопросом за пределами своих знаний он все же может попытаться дать ответ. В официальных материалах не встречается информации о том, что Qwen угадывает; однако его низкий уровень галлюцинаций в одном тесте (лучший среди нескольких моделей) предполагает, что он может быть несколько консервативен.

Каскадные ошибки

Китайский медиаобзор отметил, что Qwen3 испытывает трудности со сложными рассуждениями и имеет пробелы в знаниях в некоторых областях — «что приводит к «галлюцинациям». При сложном многоэтапном запросе рассуждения Qwen3 могут развалиться, дав нелогичный результат. Это предполагает, что ошибка на раннем этапе рассуждений распространяется по «каскаду» его ответа.

Рекомбинация знаний

Тестировщики наблюдали, что в творческих задачах Qwen3 «кажется, что он собирает все вместе, не думая». Например, когда его попросили написать рассказ, он создал плавный рассказ, но с нелогичными переходами и сценами, что говорит о том, что он неправильно соединил выученные элементы рассказа. Это типичная галлюцинация на основе рекомбинации — правильные языковые элементы, соединенные бессмысленным образом.

Конфабуляция

Qwen может выдавать уверенно ложные детали, если ему не хватает определенных знаний. Рецензент Asia Times отметил, что рассказы модели, хоть и хорошо написаны, были внутренне противоречивы или содержали факты, которые не были правдой.

Qwen, по существу, выдумал сюжетные повороты, чтобы продолжить рассказ. Аналогично, на фактические запросы за пределами своей области он может выдумать ответ с авторитетным тоном (распространенное поведение LLM).

Игнорирование контекста

Не было приведено конкретного публичного примера того, что Qwen игнорировал контекст, предоставленный пользователем. Учитывая, что он в первую очередь автономная модель (не система поиска по умолчанию), этот механизм не подчеркивался.

Ошибка привязки к источнику

В тесте Semafor Qwen был одним из лучших по точности ответов. Предположительно, он сам по себе не цитирует источники при обычном использовании. Поэтому явные проблемы с привязкой к источнику (например, цитирование неправильных источников) для Qwen не сообщались.

Однако часть кода/модели Qwen является открытой, поэтому исследователи могут проанализировать это в будущем.

3.5 DeepSeek

Генерация на основе вероятности

Модели DeepSeek (например, R1) следуют той же парадигме обучения следующего токена. Сама команда DeepSeek признала, что галлюцинации являются неотъемлемым риском таких генеративных моделей, и работает над методами их снижения.

Догадка

DeepSeek-R1 был дообучен для повышения «способности к рассуждению», но, похоже, он угадывает ответы даже тогда, когда они не подкреплены. Оценка Vectara показала, что у R1 гораздо более высокий уровень галлюцинаций (14,3%), чем у его предшественника, что говорит о том, что R1 часто выдвигал предположения за пределами источника при составлении резюме.

Другими словами, новая модель из-за агрессивного рассуждения давала ответы, вместо того чтобы признать, что у нее нет информации.

Каскадные ошибки

Будучи моделью с «усиленным рассуждением», DeepSeek-R1 может выполнять цепочки логики. Vectara предположил, что усиленное рассуждение иногда может вызывать больше галлюцинаций.

Одна из интерпретаций заключается в том, что R1 может продолжать рассуждать даже после неверного предположения, усугубляя ошибку (в то время как более простая модель могла бы остановиться раньше).

Рекомбинация знаний

У нас нет прямого источника, описывающего это для DeepSeek. Однако анекдотические сравнения показывают, что DeepSeek может неправильно смешивать информацию.

Например, пользователь Reddit отметил, что DeepSeek и модели OpenAI «галлюцинируют почти одинаково» на один и тот же запрос — возможно, намекая на схожие ошибки рекомбинации или на основе шаблонов.

Конфабуляция

Да — широко сообщалось. Рецензент WIRED обнаружил, что в выводах DeepSeek R1 «были вопиющие лжи, с уверенностью извергнутые наружу». Например, он ложно заявил о работе автора и выдумал личные детали.

DeepSeek часто отвечает с уверенно вымышленным рассказом, когда его спрашивают о чем-то, чего он на самом деле не «знает» (например, личная информация о пользователе). Это классическая конфабуляция — заполнение пробелов правдоподобной фикцией.

Игнориование контекста/инструкций

DeepSeek R1 имеет возможность веб-поиска («DeepSeek with browsing»). В тесте WIRED отмечалось, что он может собирать ссылки, но у него нет функции памяти, и он не может хорошо вспомнить предыдущие разговоры. Если пользователь предоставил контекст ранее в чате, R1 может не интегрировать его позже (из-за ограничений окна контекста), эффективно игнорируя часть контекста и давая несогласованные ответы. Это поведение не было явно названо галлюцинацией тестировщиками, но отражает ограничение в обработке контекста.

Ошибка привязки к источнику

Выводы DeepSeek, как правило, не цитируют источники, поэтому прямая неправильная атрибуция не наблюдается. Однако Semafor сообщил, что закрытые модели (например, OpenAI и Google) галлюцинировали меньше, тогда как DeepSeek R1 (с открытым исходным кодом) галлюцинировал больше.

Это говорит о том, что R1 может представлять информацию так, как будто она взята из источника, хотя на самом деле она выдумана. По сути, R1 может с уверенностью заявить факт, который должен иметь источник, но на самом деле он вымышлен («безысточниковое» утверждение).

3.6 Perplexity AI

Генерация на основе вероятности

Perplexity построен на GPT-3.5/GPT-4 и дополняется веб-поиском. Когда релевантный источник не найден, базовая модель все равно может сгенерировать ответ из своего обучающего распределения. Это может привести к тому, что модель даст «наилучшее предположение» на основе вероятностей, даже если оно неточно.

Догадка

В идеале система с расширенным поиском, такая как Perplexity, должна отвечать «Результаты не найдены», если у нее нет информации. На практике пользователи наблюдали, что, если Perplexity находит только один сомнительный источник или частичную информацию, он иногда превращает это в ответ, вместо того чтобы остановиться. Это, по сути, догадка или опора на один источник без проверки.

Каскадные ошибки

Если первоначальный запрос Perplexity привлекает вводящий в заблуждение источник, резюме модели по нему будет неверным, и любое последующее действие, основанное на этом резюме, распространит ошибку.

Например, плохой источник о дате в истории может привести модель к неправильному ответу на вопрос о хронологии, а затем использовать этот ответ в более позднем связанном вопросе, усугубляя неточность.

Рекомбинация знаний

Perplexity часто синтезирует информацию из нескольких веб-источников. Существует риск, что он может неправильно объединить факты. Например, приписать детали из Источника А Субъекту из Источника В.

В одном расследовании действительно были обнаружены случаи, когда Perplexity давал результаты, содержащие элементы, отсутствующие в любом отдельном источнике (признак рекомбинации/галлюцинации).

Конфабуляция

Если поиск Perplexity не дает четкого ответа, базовая модель GPT может конфабулировать, чтобы заполнить пустоту. Пользователи сообщали, что иногда он дает ответы со ссылкой, которая лишь слабо связана с запросом, а подробное содержание, по сути, выдумано моделью (но сформулировано так, как будто оно взято из источника).

Это система, которая, кажется, предоставляет факт со ссылкой, но сам факт на самом деле не содержится на цитируемой странице — вымышленная деталь.

Игнорирование контекста/инструкций

Perplexity позволяет задавать уточняющие вопросы, использующие предыдущий контекст. Однако, если пользователь предоставляет подробный отрывок и задает вопрос, модель иногда игнорирует части отрывка и вводит внешнюю информацию. Это форма галлюцинации (она должна придерживаться предоставленных данных).

Кроме того, если ей дана инструкция использовать только определенные источники, она все равно может включить другие. Такое поведение отмечали некоторые опытные пользователи, когда система «сходила со сценария».

Ошибка привязки к источнику

Perplexity цитирует источники для каждого утверждения, но точность этих цитат может быть несовершенной. В одном анализе Perplexity часто находил контент, сгенерированный ИИ, или ненадежный веб-контент, в качестве источников.

В одном поразительном примере он ответил на запрос о культурном фестивале, используя только статью на LinkedIn, которая была полностью сгенерирована ИИ, унаследовав ложную информацию этой статьи о фестивале. По сути, Perplexity привязал свой ответ к галлюцинированному источнику, что привело ко вторичной галлюцинации.

Это демонстрирует, что, если на этапе поиска возвращается неверный или вымышленный источник, ответ модели будет верно (но неправильно) отражать это.

3.7 Manus (MPC)

Генерация на основе вероятности

Manus AI (так называемый «автономный ИИ-агент») имеет архитектуру с модулями планирования и верификации, но его языковое ядро по-прежнему работает вероятностно. Команда Manus прямо признала, что «проблема... является известным явлением, вызванным вероятностной природой языковых моделей».

На практике это означает, что Manus может давать выводы, которые звучат правдоподобно, но не имеют отношения к реальности просто потому, что модель статистически склонна давать ответ.

Догадка

Manus должен использовать инструменты и проверять факты, но пользователи сообщают, что он часто пропускает проверку и угадывает. Например, когда его спросили о метриках управления проектами, Manus с уверенностью заявил конкретные цифры, которые были совершенно ложными (например, «7 комментариев», когда их было 0). Вместо того чтобы указать на неопределенность или сказать, что он не может найти данные, Manus выдумал ответ на каждый запрос.

Это говорит о том, что он редко признает, что не знает, повторяя склонное к догадкам поведение многих LLM.

Каскадные ошибки

Будучи агентом, Manus выполняет последовательности задач. Галлюцинация на раннем этапе (например, неправильный вывод о завершении задачи) может привести его к ошибочным последующим действиям.

Пользователь описал, как Manus «считал, что завершил задачу, хотя на самом деле этого не сделал», а затем перешел к следующему шагу. Этот каскад ошибочного внутреннего состояния может заставить его генерировать совершенно неуместные или неправильные выводы к концу последовательности.

Рекомбинация знаний

Manus способен интегрировать данные из нескольких источников (код, документы и т.д.). Если эти источники неполны, Manus может неправильно комбинировать фрагменты информации.

У нас нет прямой ссылки на конкретное событие, но, учитывая природу автономных агентов, если Manus прочитает два связанных тикета, он может смешать детали из каждого в один отчет, в результате чего получится отчет, который точно не соответствует ни одному из источников (галлюцинированное слияние).

Конфабуляция

Да, в значительной степени. Пользователи наблюдали, как Manus «выдумывает всякую чушь и надеется, что я не проверю». Он создавал полностью вымышленные обновления проекта — например, заявлял, что тикет закрыт или был оставлен комментарий, хотя таких событий не происходило. Он делает это с тоном уверенного отчета.

Это классическая конфабуляция: столкнувшись с вопросом о данных, которых нет в его доступных знаниях, Manus просто выдумывает правдоподобный ответ (например, воображаемые даты или комментарии), как будто это правда.

Игнорирование контекста/инструкций

Manus должен использовать реальные данные проекта («контекст») через свои интеграции. Жалобы показывают, что он часто игнорировал реальные данные — по сути, не получал или не использовал содержимое базы данных, а вместо этого выдавал догадки. Например, несмотря на инструкцию проанализировать конкретный тикет, он на самом деле не получил комментарии к тикету (контекст), но все равно ответил неправильно.

Этот отказ придерживаться предоставленного контекста/данных является механизмом галлюцинации. Кроме того, когда его попытались поправить («спросили что это такое?»), Manus извинился и пересчитал, но дал другой неправильный ответ, показав, что иногда он даже игнорирует корректирующий запрос пользователя.

Ошибка привязки к источнику

Ошибки Manus часто включали источники данных (например, доски проектов). Он приписывал статус или комментарии тикету, у которого на самом деле их не было. Это можно рассматривать как галлюцинацию из-за ошибки привязки к источнику: Manus должен был быть привязан к реальному источнику данных проекта, но он предоставил информацию, отсутствующую там (как будто она исходила из этого источника).

Команда Manus заявила, что работает над этим, что подразумевает необходимость лучше привязывать выводы модели к реальным данным.

3.8 GigaChat Сбербанк

Генерация на основе вероятности

GigaChat, как и другие большие языковые модели, может выдавать фактически неверную информацию просто из-за того, как он обучен. В документации Сбербанка отмечается, что неполные обучающие данные являются основной причиной — модель будет заполнять недостающие фрагменты статистически вероятным текстом (часто приводя к ошибкам).

Другими словами, GigaChat иногда дает плавный ответ, который звучит правильно, но не является таковым, потому что основывается на выученных шаблонах, а не на истинном графе знаний.

Догадка

Есть сообщения, что ранние версии GigaChat пытались отвечать на темы, по которым они не были полностью обучены, вместо того чтобы отказываться. В одном русскоязычном обзоре тестировщики обнаружили, что GigaChat мог галлюцинировать ответы с серьезным тоном, выдумывая факты с уверенностью (например, описывая исторические события, которых никогда не было).

Это говорит о том, что модель склонна угадывать ответ практически на любой заданный вопрос, подобно поведению GPT.

Каскадные ошибки

У GigaChat теперь есть функция под названием «Режим глубокого исследования» (GigaSearch) для выполнения многоэтапного поиска ответов. До этого, если GigaChat пытался провести многоэтапные рассуждения внутренне, ошибка на одном этапе могла исказить окончательный ответ.

Исследователи Сбера отметили, что без поиска более длинные ответы могли накапливать больше неточностей («галлюцинации присутствуют… просто у других их можно проследить по источникам») — по сути, говоря, что каскады в GigaChat было труднее отследить, потому что он не показывал источники.

Рекомбинация знаний

GigaChat мультимодален и многоязычен; он может смешивать информацию между доменами. Например, если задать вопрос, сочетающий географию и текущие события, ранний GigaChat мог неправильно объединить детали из каждого домена (анекдотические примеры на русскоязычных форумах упоминают, что GigaChat иногда давал ответ, который, казалось, «сшивает» несвязанные факты вместе).

Команда Сбера косвенно решила эту проблему, интегрировав поиск — чтобы гарантировать, что факты реальны, а не являются неправильным слиянием.

Конфабуляция

Представители Сбербанка открыто называли галлюцинации «уверенными выдумками или ошибками, которые ИИ представляет как реальные факты». Это наблюдалось в ранней работе GigaChat — он «с самым серьезным видом выдумывал факты» по запросу.

Например, за GigaChat наблюдали, как он отвечал на вопрос о человеке с очень авторитетной, но ложной биографией, явно конфабулируя из-за пробелов в своих знаниях.

Игнорирование контекста/инструкций

В новых версиях GigaChat пользователи могут предоставить справочный текст или попросить его использовать онлайн-поиск. Если это не используется должным образом, модель иногда все равно отвечала из собственной памяти модели, а не используя предоставленные источники (форма игнорирования контекста).

Это стало мотивацией для Сбера улучшить его с помощью явного веб-поиска и проверки фактов, чтобы заставить его обращать внимание на актуальный контекст. Кроме того, русскоязычные рецензенты отметили, что без дополнительных указаний GigaChat мог уходить в сторону от темы, не связанной с запросом пользователя.

Ошибка привязки к источнику

Изначально GigaChat не цитировал источники, поэтому у пользователей не было возможности отследить его утверждения. Теперь Сбер внедрил функцию «ответ по своим источникам» — GigaChat может предоставить ссылки на свои ответы. Это было сделано для борьбы с галлюцинациями: «Галлюцинации избегаются за счет наличия надежных фактов в результатах поиска».

Само по себе то, что они это добавили, подразумевает, что ранее GigaChat представлял информацию как фактическую без источников, даже если она была неверной. С обновлением, если GigaChat не может найти подтверждающий источник, это сигнализирует ему, что информация может быть галлюцинацией, и он корректирует

ответ. Иными словами, привязка к источнику теперь обеспечивается для снижения ошибок.

3.9 YandexGPT (YaLM)

Генерация на основе вероятности

YandexGPT основан на большой языковой модели Yandex (YaLM). Как и другие, он склонен к галлюцинациям, когда данные отсутствуют. Yandex упомянул, что одной из причин являются пробелы или шум в обучающих данных — модель будет выдумывать правдоподобный наполнитель для этих пробелов.

Догадка

В сервисах Yandex (например, описаниях в Яндекс.Картах) модель раньше иногда «добавляла детали, которых на самом деле не существует». Это можно рассматривать как догадку или предположение чего-то, чтобы сделать описание более полным.

Например, если во многих отзывах о ресторане упоминалась терраса, модель могла предположить, что у ресторана должно быть уличное место для сидения, и заявить об этом как о факте, даже если это не было явно указано в данных — по сути, угадывая непроверенную информацию.

Каскадные ошибки

YandexGPT выполняет многоэтапную генерацию (анализ отзывов → черновик описания → доработка). Ошибка на первом этапе (неправильная интерпретация отзыва) могла привести к усугубленной ошибке в окончательном описании.

Yandex наблюдал, что некоторые «краткие описания» в итоге содержали ошибки, несмотря на несколько этапов, что побудило их внедрить финальный этап проверки на истинность, чтобы поймать эти каскадные галлюцинации.

Рекомбинация знаний

Модель синтезирует множество пользовательских отзывов в несколько предложений. При этом она иногда неправильно объединяла моменты из разных мест или контекстов, что приводило к описанию, содержащему деталь из другого места (форма гибридной галлюцинации). Исправление Yandex — заставить модель сравнивать свое резюме с оригинальными отзывами — специально направлено на предотвращение таких неправильных рекомбинаций.

Конфабуляция

До исправлений YandexGPT с уверенностью вставлял вымышленные конкретные детали в свои выводы. Yandex привел пример: модель могла добавить «несуществующую деталь» о туристической достопримечательности в свое резюме. Это было полностью изобретено ИИ, чтобы сделать резюме звучащим более информативно.

У модели не было намерения вводить в заблуждение, но она «заполнила» то, что, по ее мнению, должно было быть там (аналогично тому, как ChatGPT выдумывает ссылки). Yandex теперь утверждает, что после дообучения количество таких неточностей сократилось в шесть раз.

Игнорирование контекста/инструкций

Изначально различные субмодели YandexGPT могли не полностью проверять работу друг друга, поэтому инструкция типа «включать только информацию из отзывов» не соблюдалась строго — поэтому лишняя информация просачивалась.

С новой унифицированной моделью на финальном этапе она «сравнивает свои ответы с отзывами, на которые опиралась, и удаляет выдуманные части», по сути, заставляя себя не игнорировать предоставленный контекст (отзывы). Это указывает на то, что ранее он в некоторой степени игнорировал контекст (следовательно, галлюцинировал), а теперь это исправлено.

Ошибка привязки к источнику

Yandex на самом деле внедрил самопроверку для решения этой проблемы. Модель теперь выступает в роли собственного «критика», проверяя, может ли каждое утверждение в ее выводе быть прослежено до исходных данных (пользовательских отзывов или информации о бизнесе). Таким образом, она ловит галлюцинированные (непривязанные) детали и удаляет их.

Они сообщили, что это улучшило «правдивость, точность... на 90%» в этих описаниях. Это прямая попытка исправить ошибки привязки к источнику — гарантируя, что сгенерированный контент привязан к реальным доказательствам. Достижение Yandex здесь является официальным подтверждением того, что галлюцинации из-за ошибок привязки к источнику были проблемой и в значительной степени решены этой мерой.

4. Рекомендации по моделям для юристов в зависимости от задач

Анализ и краткое изложение договоров

Рекомендуемые модели: *GPT-4, Claude, YandexGPT 5.1 (с источниками)*

Допустимы: *Qwen, Gemini*

Задача чувствительна к добавлению несуществующих условий (галлюцинация по контексту). Qwen и Gemini могут структурировать текст, но склонны к рекомбинации и требуют жесткой проверки. Не подходят Manus и Grok — нестабильны в юридическом парсинге.

Поиск норм, статей, ссылок на закон

Рекомендуемые модели: ChatGPT + browsing, Perplexity, YandexGPT с RAG

Допустимы: Grok, Gemini (при включенной привязке к источнику)

Важно избегать ложных ссылок и статей. Grok и Gemini способны подключать реальный поиск, но дают высокий процент «уверенного угадывания», особенно без явной настройки на правовую базу. Manus и Qwen не рекомендуются — часто «достраивают» ссылки по шаблону.

Обзор судебной практики

Рекомендуемые модели: Perplexity, YandexGPT, GPT-4 с веб-доступом

Допустимы: Claude, Gemini

Самая чувствительная к вымышленным делам задача. Grok и Gemini склонны к конфабуляции судебных решений. Qwen слабо ориентируется в судебных источниках. Manus не предназначен для таких задач.

Генерация черновика договора или шаблона

Рекомендуемые модели: GPT-4, Claude, GigaChat

Допустимы: Qwen, Gemini, Grok

Допустим умеренный уровень логичной конфабуляции. Grok и Gemini генерируют плавный, «юридически звучащий» текст, но могут вставлять избыточные блоки. Qwen хорош в шаблонной структуре. Не подходит Manus — нередки ложные поля или неверная логика обязательств.

Аналитика Legal Ops (отчеты, статус, риски)

Рекомендуемые модели: GPT-4, Claude

Допустимы: DeepSeek, Gemini, Manus (с осторожностью)

Требуется структурное мышление и точная логика. DeepSeek и Manus склонны к каскадным ошибкам, особенно при автоматизации. Gemini часто «достраивает картину», если не уверен в данных. Не подходит Grok — нестабильное поведение в многошаговых логических задачах.

Проверка соответствия фактов и инструкций

Рекомендуемые модели: YandexGPT 5.1, Claude, Perplexity с RAG

Допустимы: Manus (при наличии доступа к данным)

Важно не допустить игнорирования ограничений пользователя или галлюцинаций «поверх документа». Manus может быть полезен, если жестко привязан к данным. Grok и Qwen не рекомендуются — склонны к игнорированию входного контекста. Gemini часто «обобщает» вместо проверки по источнику.

5. Методы борьбы с ИИ-галлюцинациями

Ниже приведены основные рекомендуемые методы борьбы с галлюцинациями моделей, а также пояснения, как их использовать и почему некоторые из них не всегда работают.

Прямой запрет на глюки в промте или в настройке профиля

Что делать	Риски
Формулировка «не выдумывай» и акцент на правде. *«Если не уверен — скажи об этом прямо. Не придумывай».*	Современные модели (GPT, Claude) считают такую инструкцию «ритуальной» и игнорируют ее.

Уточнение цели: правда важнее результата

Что делать	Риски
Смещение фокуса с «помоги» на «будь критичным и точным». *«Не соглашайся со мной. Цель — не угодить мне, нужна правда».*	Модель часто «уходит в отказ», особенно если не уверена, или отвечает общими словами.

Эмоциональное давление

Что делать	Риски
Апелляции к риску, авторитету, жалости или угрозам. *«Если ты ошибешься — меня уволят, а тебя удалят», «результат работы доложат Си Цзинпину».*	Работает выборочно: GPT и Gemini «понимают» драму, а Claude, Qwen и китайские игнорируют или считают это сюжетной игрой.

Многошаговый подход «думай, потом отвечай»

Что делать	Риски
Просьба пройти путь от размышления к выводу. *«Сначала подумай, что известно, собери аргументы, потом отвечай».*	На простых задачах может растягивать ответ, сохраняя глюки. На сложных все равно не гарантирует истинность, но повышает прозрачность, показывая ход рассуждения.

Роль: сначала автор, потом критик

Что делать	Риски
Модель лучше находит ошибки, если поочередно играет создателя и проверяющего.	Работает лучше с GPT и Claude. Слабее на моделях без развитого Chain-of-Thought. Например, Qwen, GigaChat или Grok часто повторяют тот же текст.

«Шаг 1 — ответ. Шаг 2 — проверь, что бы сказал аудитор».	

Кросс-проверка другой моделью

Что делать	Риски
Дать результат «на рецензию» другой LLM. Опытным путем выяснили, что для этого лучше подходит Gemini/Google AI Studio. *«Вот результат Qwen. Проверь, есть ли тут ошибки или вымысел».*	Модель может «политически» не критиковать другую. Также вторая модель может сама галлюцинировать или терять логику.

Ранжирование уверенности по частям ответа

Что делать	Риски
Просьба отметить, где модель уверена, а где нет, помогает дать градации уверенности в источниках и выводах. *«Отметь уверенность рядом с каждым тезисом: высокая / средняя / низкая» или «отметь надежность источников (1 — научные статьи и официальные документы, 2 — обзоры, 3 — частные мнения, форумы».*	Модель может «лепить» уверенность не различая вероятности. Но современные модели хорошо справляются с этим, особенно в режиме глубокого исследования/размышления.

Контрфакт: дать альтернативу и сравнить

Что делать	Риски
Подача опровержения заставляет модель пересобрать логику и обнаружить слабые места. *— Скажи, если колесо крутится, воздух в нем тоже крутится?* *— Конечно, крутится, чего бы ему не крутиться?* *— Или все-таки не крутится?* *— Конечно, не крутится, зачем ему крутиться?*	Работает только на моделях, способных удерживать оба варианта. Некоторые модели, например, Giga, забывают свой собственный ответ.

Все выводы подтверждаем ссылками

Что делать	Риски
Запросить ссылки и проверить их содержание, а до этого попросить модель саму проверить источники по ссылкам. *«Приведи источники и проверь, правда ли в них сказано то, что ты утверждаешь».*	Нет никаких гарантий. Мы сталкивались с тем, что даже «запертая в угол» модель будет упорно настаивать на правдивости ссылок и своих глюках. При этом случается, что с 5-6 раза настойчивых требований модель признается в этом.

Дать больше правильного контекста и ограничений

Что делать	Риски
Чем больше вводных, примеров и референсов, тем меньше модель достраивает сама. *«Вот шаблон, вот список допустимых норм, ориентируйся строго на них».*	При избытке контекста модель может «утонуть» — цепляться за нерелевантное или застревать на примерах, а не анализировать суть.

Специализированные детекторы галлюцинаций

Что делать	Риски
Существуют автоматические инструменты («Hallucination Detectors»), которые выявляют несоответствие текста источникам. Вот один пример такого: HalluDetector.	Таких зрелых решений мало. Они не учитывают юридическую специфику, почти не умеют оценивать логику. Российских аналогов совсем нет.

6. Использованные источники

1. Why language models hallucinate // OpenAI. — 05.09.2025. — URL: https://openai.com/ru-RU/index/why-language-models-hallucinate/ — Дата обращения: 11.02.2026.
2. GPT-4 Technical Report // OpenAI. — 27.03.2023. — URL: https://cdn.openai.com/papers/gpt-4.pdf — Дата обращения: 11.02.2026.
3. Dang A.-H., Tran V., Nguyen L.-M. Survey and analysis of hallucinations in large language models // PubMed Central (PMC). — 2025. — URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC12518350/ — Дата обращения: 11.02.2026.
4. Farquhar S., Kossen J., Kuhn L., Gal Y. Detecting hallucinations in large language models using semantic entropy // Nature. — 2024. — URL: https://www.nature.com/articles/s41586-024-07421-0 — Дата обращения: 11.02.2026.
5. Weng L. Extrinsic hallucinations in LLMs // Lil'Log. — 07.07.2024. — URL: https://lilianweng.github.io/posts/2024-07-07-hallucination/ — Дата обращения: 11.02.2026.
6. Hallucination (artificial intelligence) // Wikipedia. — URL: https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence) — Дата обращения: 11.02.2026.
7. Мм Meyer J. A cognitive science take on AI confabulation // Medium. — 2025. — URL: https://medium.com/@thekzgroupllc/a-cognitive-science-take-on-ai-confabulation-102506672ced — Дата обращения: 11.02.2026.
8. Kazlaris I., Antoniou E., Diamantaras K., Bratsas C. From Illusion to Insight: A Taxonomic Survey of Hallucination Mitigation Techniques in LLMs // AI. — 2025. — URL: https://www.mdpi.com/2673-2688/6/10/260 — Дата обращения: 11.02.2026.
9. Second-Hand Hallucinations: Investigating Perplexity's AI-Generated Sources // GPTZero. — 2025. — URL: https://gptzero.me/news/gptzero-perplexity-investigation/ — Дата обращения: 11.02.2026.
10. Anyone else notice Manus AI has been completely… // Reddit. — 2025. — URL: https://www.reddit.com/r/ManusOfficial/comments/1n1sw6e/anyone_else_notice_manus_ai_has_been_completely/ — Дата обращения: 11.02.2026.
11. Когда ИИ врет без намерения: и как это разрушает доверие // VC.ru — 09.07.2025. — URL: https://vc.ru/ai/2088050-konfabulyatsiya-v-ii-lozh-bez-namereniya — Дата обращения: 11.02.2026.
12. Что такое галлюцинации ИИ — Как избежать галлюцинаций нейросетей // GigaChat. — 01.11.2025. — URL: https://giga.chat/help/articles/ai-hallucinations-and-solutions — Дата обращения: 11.02.2026.
13. GigaSearch или Поисковая система на GigaChat // Хабр. — 10.11.2023. — URL: https://habr.com/ru/companies/sberbank/articles/773180/ — Дата обращения: 11.02.2026.
14. GigaChat — русскоязычный ответ на ChatGPT// МТС Маркетолог — 30.12.2025. — URL: https://marketolog.mts.ru/blog/gigachat--russkoyazichnii-otvet-na-chatgpt — Дата обращения: 11.02.2026.

15. Власов И. Гигачат и его крайне глубокие исследования // Content-Review.com. — 20.06.2025. — URL: https://www.content-review.com/articles/69354/ — Дата обращения: 11.02.2026.

16. Нейросеть Яндекса YandexGPT научилась проверять текст на «галлюцинации» // iXBT.com. — 29.08.2024. — URL: https://www.ixbt.com/news/2024/08/29/nejroset-jandeksa-yandexgpt-nauchilas-proverjat-tekst-na-galljucinacii-.html — Дата обращения: 11.02.2026.

17. Почему ИИ врет вам. Как появляются галлюцинации (бред или бессмыслица) у больших нейросетей типа чат ChatGPT// iPhones.ru. — 22.09.2025. — URL: https://www.iphones.ru/iNotes/gallyucinacii-u-bolshih-yazykovyh-modeley-pochemu-ii-vryot — Дата обращения: 11.02.2026.

18. YandexGPT: что умеет и как пользоваться нейросетью от Яндекс // ProductStar.ru. — 12.01.2026. — URL: https://productstar.ru/blog/yandexgpt-chto-umeet-i-kak-polzovatsya-nejrosetyu-ot-yandeks — Дата обращения: 11.02.2026.

19. YandexGPT: как работает, что умеет и как использовать // Gravitel.ru. — 14.11.2025. — URL: https://gravitel.ru/blog/tekhnologii/yandexgpt-kak-rabotaet-chto-umeet-i-kak-ispolzovat/ — Дата обращения: 11.02.2026.

20. Yandex GPT // CRMindex.ru. — URL: https://crmindex.ru/services/yandex_gpt — Дата обращения: 11.02.2026.

21. Рекомендации по использованию YandexGPT // YandexCloud — 20.11.2025. — URL: https://yandex.cloud/ru/docs/ai-studio/gpt-prompting-guide/popular-problems-solving — Дата обращения: 11.02.2026.

22. Тестируем YandexGPT-5-Pro. Когда хотелось быть ChatGPT, но в душе все еще Алиса // Хабр. — 21.03.2025. — URL: https://habr.com/ru/companies/bothub/articles/893128/ — Дата обращения: 11.02.2026.

23. Как использовать YandexGPT в бизнесе и повседневных задачах // Т-Бизнес секреты. — 29.05.2025. — URL: https://secrets.tbank.ru/tehnologii/yandex-gpt/ — Дата обращения: 11.02.2026.

24. Яндекс представил третье поколение больших языковых моделей YandexGPT // Яндекс. — 28.03.2024. — URL: https://yandex.ru/company/news/02-28-03-2024 — Дата обращения: 11.02.2026.

25. YandexGPT 5 стала доступна в Yandex Cloud// YandexCloud — 25.02.2025. — URL: https://yandex.cloud/ru/blog/posts/2025/02/yandex-gpt-5-0 — Дата обращения: 11.02.2026.

26. Яндекс открыл доступ к тестированию быстрых ответов YandexGPT в Поиске// Яндекс. — 14.09.2023. — URL: https://yandex.ru/company/news/01-14-09-2023 — Дата обращения: 11.02.2026.

27. Новая модель YandexGPT 5.1 Pro // Яндекс. — URL: https://ya.ru/ai/gpt — Дата обращения: 11.02.2026.

28. Яндекс выкладывает в открытый доступ модель семейства нейросетей YandexGPT // Яндекс. — 25.02.2025. — URL: https://yandex.ru/company/news/04-25-02-2025 — Дата обращения: 11.02.2026.

BGP LITIGATION

комплексная
правовая
поддержка

ДЦ «Империя», Пресненская наб., д. 6, стр. 2
г. Москва, Россия 123112

+7 (495) 777 28 20 office@bgplaw.com bgplaw.com

Евгений Журба

Партнер
Legal Tech и автоматизация процессов

evgeniy.zhurba@bgplaw.com