УДК 004.832.34

***Шинкарук Кирилл Александрович**, магистрант, Федеральное государственное автономное образовательное учреждение высшего образования «Южно-Уральский государственный университет (национальный исследовательский университет) (ФГАОУ ВО «ЮУрГУ (НИУ)»), г. Челябинск*

ПРОБЛЕМА ГАЛЛЮЦИНАЦИЙ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ПРИ ОТВЕТАХ НА ВОПРОСЫ ПО НЕДОСТУПНЫМ КОРПОРАТИВНЫМ ДОКУМЕНТАМ

Аннотация

Предмет исследования – галлюцинации больших языковых моделей при ответах на вопросы по корпоративным документам, к которым модель не имеет прямого доступа. В статье обосновывается тезис: простое обращение к языковой модели не решает задачу корпоративного вопросно-ответного поиска, если ответ должен подтверждаться внутренними регламентами, договорами, инструкциями, отчетами или базами знаний. Вывод статьи связан с переходом от closed-book QA к документно обоснованным сценариям вопросно-ответного поиска.

Ключевые слова

большие языковые модели, галлюцинации, корпоративные документы, вопросно-ответный поиск, closed-book Question Answering, document-grounded Question Answering.

Abstract

The subject of this study is the hallucinations of Large Language Models (LLMs) when answering questions about corporate documents to which the model has no direct access. The paper substantiates the thesis that simply querying a language model does not solve the problem of corporate question answering when responses must be

supported by internal regulations, contracts, instructions, reports, or knowledge bases. The main conclusion of the study is the necessity of transitioning from closed-book question answering (closed-book QA) to document-grounded question answering scenarios.

Keywords

Large Language Models (LLMs), hallucinations, corporate documents, question answering, closed-book question answering (closed-book QA), document-grounded question answering (document-grounded QA).

1. Принцип генерации ответа больших языковых моделей

Чтобы разобраться в проблеме галлюцинаций больших языковых моделей (БЯМ), нужно уточнить как формируется ответ в целом. БЯМ не выбирает готовый ответ из базы знаний. На каждом шаге она получает текущий контекст, вычисляет вероятности возможных следующих токенов и выбирает один из них. Затем выбранный токен добавляется к тексту, и процесс повторяется.

Основная формула, которая описывает выбор следующего токена выглядит так:

$$P(token_i) = softmax(Wh)_i \quad (1)$$

где

$P(token_i)$ – вероятность следующего токена;

$token_i$ – конкретный токен словаря;

Wh – оценка для токенов в словаре;

$softmax(Wh)_i$ – функция нормализации, преобразующая оценки сходства в вероятностное распределение весов внимания.

Модель не выбирает правильный ответ напрямую. Она строит распределение вероятностей и затем либо берёт самый вероятный токен, либо случайно выбирает из верхней части распределения.

После softmax модель получает вероятности. Дальше она может взять самый вероятный токен или выбрать токен случайно из верхней части распределения. На выбор влияют температура и другие параметры генерации.

$$softmax(Wh)_i = \frac{e^{Wh_i}}{\sum_j e^{Wh_i}} \quad (1)$$

где

Wh – оценка для токенов в словаре;

$\sum_j e^{Wh_i}$ – сумма по всем токенам словаря;

T – температура.

Также стоит упомянуть механизм «внимания» (attention). Механизм внимания – способ для модели в процессе генерации токена решить, какие части предыдущего текста учитывать, а какие нет. Когда пользователь задает вопрос «Столица России», для БЯМ важнее слово «Россия», чем остальной текст. Математически механизм внимания определяется тремя векторами, которые строятся из матриц пространства имен (W_Q), пространство ключей (W_K) и пространство значений (W_V). Эти матрицы являются обучаемыми линейными преобразованиями, проецирующими вектора токенов в три различных пространства [23]:

$$Q = XW_Q \tag{3}$$

$$K = XW_K \tag{4}$$

$$V = XW_V \tag{5}$$

где

Q – матрица запросов (Query), принято понимать, как ответ на вопрос «Что я ищу?»;

W_Q – обучаемая матрица весов для Query;

K – матрица ключей (Key), принято понимать, как ответ на вопрос «Какую информацию я несу для других?»;

W_K – обучаемая матрица весов для Key;

V – матрица значений (Value), принято понимать, как ответ на вопрос «Какую информацию передать дальше?»;

W_V – обучаемая матрица весов для Value;

X – входной вектор токена.

Полученные векторы используются механизмом внимания для определения степени значимости различных токенов контекста относительно текущего токена. Для этого вычисляется матрица весов внимания на основе сходства между запросами Q и ключами K, после чего полученные веса применяются к векторам значений V. Механизм масштабированного скалярного внимания (Scaled Dot-Product Attention) определяется выражением:

$$Attantion(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

где

Q – матрица запросов (Query);

K – матрица ключей (Key);

V – матрица значений (Value);

K^T – транспонированная матрица ключей;

d_k – размерность пространства ключей и запросов;

$\sqrt{d_k}$ – нормализующий множитель;

$softmax$ – функция нормализации, преобразующая оценки сходства в вероятностное распределение весов внимания.

2. Галлюцинации БЯМ: понятие и терминологические границы

Термин «галлюцинация» в контексте генеративных языковых моделей обычно обозначает ситуацию, когда система формирует текст, который выглядит правдоподобным, но является ложным, неподтвержденным или не выводится из заданного источника. В обзоре Z. Ji и соавторов галлюцинации рассматриваются как широкая проблема генерации естественного языка, возникающая в суммаризации, диалоговых системах, машинном переводе, data-to-text и генеративном question answering [1]. Для корпоративной темы существенна одна

деталь: ошибка может состоять не только в противоречии общеизвестному факту, но и в расхождении с конкретным документом, который должен служить основанием ответа.

В обзорах БЯМ-галлюцинаций эта проблема описывается не как единичный дефект отдельной модели, а как результат архитектуры, обучения, данных, процедуры вывода и характера пользовательской задачи [2]. Термин «галлюцинация» удобен как общее обозначение, но для научного и инженерного анализа его приходится уточнять. В корпоративном вопросно-ответном поиске стоит различать как минимум четыре близких, но не тождественных типа ошибок.

Factual hallucination, или фактическая галлюцинация. Она возникает, когда модель сообщает факт, который не соответствует в действительности. В корпоративном сценарии это может быть неверная дата вступления регламента в силу, ошибочная сумма договора, неправильное имя ответственного лица, выдуманный срок согласования заявки или ложное указание на наличие льготы. Такая ошибка кажется простой, но ее трудно обнаружить, если пользователь не держит перед глазами исходный документ.

Fabrication, или фабрикация. Здесь модель не просто ошибается в частном факте, а создает несуществующий объект: документ, пункт, ссылку, внутренний приказ, номер приложения. Исследование J. Buchanan, S. Hill и O. Shapoval демонстрирует значимость фабрикации на примере несуществующих библиографических ссылок, сгенерированных ChatGPT в экономической тематике [18]. Для корпоративной среды аналогичной проблемой является ссылка на несуществующий раздел договора, вымышленный внутренний стандарт или «пункт 7.3 политики», которого нет в документе.

Confabulation, или конфабуляция. В исследовании S. Farquhar и соавторов этот термин связан с неопределенностью на уровне смысла. Модель может порождать разные семантически несовместимые ответы, сохраняя при этом внешнюю связность речи [6]. Для пользователя корпоративной системы конфабуляция опасна тем, что текст не обязательно выглядит абсурдным. Он

может быть гладким, хорошо структурированным и написанным в уверенной тональности. Но за этой формой скрывается отсутствие устойчивой опоры в фактах или источниках.

Unsupported generation, неподтвержденная генерация. Это наиболее важная категория для вопросов по корпоративным документам. Ответ может не противоречить реальности в общем виде и даже казаться разумным, но он не поддержан тем документом, на который должен опираться. Например, модель может правильно описать типовую процедуру закупки, но ошибиться именно в локальном порядке конкретной организации или придумать несуществующие.

Для описания этой разницы полезно различать factuality и faithfulness. Factuality относится к фактической правильности высказывания, а с faithfulness к верности заданному источнику или контексту [1]. В корпоративном QA второе измерение часто важнее первого. Общий ответ может быть фактически допустим, но бесполезным и рискованным. Если пользователь спрашивает о конкретном договоре, где предусмотрен особый порядок уведомления, то может получить подобный ответ, вместо конкретики. Надежный ответ должен формироваться не только по общему знанию БЯМ, но и по конкретному документу.

Термин «галлюцинация» сам по себе требует осторожности. M. Hicks, J. Humphries и J. Slater обращают внимание, что метафора галлюцинации может антропоморфизировать модель и смягчать проблему фабрикации или безразличия к истинности [19]. Для данной статьи это замечание существенно. БЯМ не видит документ и не забывает его как человек. Она генерирует текст в соответствии с вероятностными закономерностями, заданными обучением.

В корпоративном вопросно-ответном поиске галлюцинации БЯМ приходится понимать широко. Это не только ложные утверждения, но и ответы без документального основания, фабрикация реквизитов, конфабуляция при неопределенности и нарушение привязки к источнику. Общий признак один: форма звучит уверенно, а содержание нельзя проверить по нужному документу. Из этого разрыва и вырастает системная проблема.

3. Почему БЯМ может отвечать уверенно при отсутствии фактического знания

N. McKenna показывает, что поведение БЯМ в задачах вывода может объясняться запоминанием и статистическими паттернами обучающих данных, которые приводят к ложным выводам [3]. В корпоративной задаче это проявляется просто. Модель распознает жанр вопроса, типовые формулировки регламентов, структуру договора, лексику инструкций и ожидаемый стиль ответа. Эти сигналы позволяют ей сгенерировать правдоподобный фрагмент делового текста. Но распознавание жанра и шаблона не дает знания о конкретном содержании внутреннего документа.

Параметрическая память БЯМ отличается от базы данных или архива документов. В параметрах хранятся не проверяемые записи с указанием источника, версии и даты обновления, а распределенные статистические представления, сформированные при обучении. Даже если модель когда-то встречала похожие документы, она не обязана хранить их точно. Почти все корпоративные документы не входят в открытые обучающие данные, поскольку являются конфиденциальными или постоянно изменяемыми.

Дополнительный фактор – смещение к ответу. Пользователь задает вопрос в форме, предполагающей, что ответ существует: «Что сказано в регламенте?», «Какая ответственность предусмотрена договором?», «Какой порядок согласования установлен инструкцией?». Диалоговая модель, обученная помогать пользователю, часто стремится удовлетворить этот запрос. Исследование A. Slobodkin и соавторов специально рассматривает случаи, когда БЯМ проявляет сверхуверенность на вопросах, на которые нельзя корректно ответить [5]. Для корпоративного QA это центральная проблема: вопрос может быть неотвечаемым не потому, что ответа вообще нет, а потому что документ недоступен модели.

Другая линия исследований касается того, умеет ли БЯМ распознавать границы собственного знания. Z. Yin и соавторы спрашивают, знают ли большие языковые модели, чего они не знают, и рассматривают unknown facts и self-

awareness of knowledge boundaries [9]. H. Zhang и соавторы предлагают подходы к самодетекции неизвестного [10], A. Amayuelas и соавторы анализируют known-unknowns uncertainty [11]. Смысл этих работ для корпоративной темы прямой: отказ от ответа и оценка неизвестности не возникают у БЯМ автоматически. Для этого нужны специальные механизмы.

P. Manakul, A. Liusie и M. Gales предлагают метод SelfCheckGPT. Обнаружения галлюцинаций через самосогласованность нескольких генераций [4]. Его идея важна для понимания природы ошибки. Когда у модели нет устойчивого фактического основания, разные ответы могут расходиться между собой. В реальном корпоративном диалоге пользователь обычно видит один ответ, а не серию разных вариантов. Поэтому нестабильность знания может быть скрыта за единичной уверенной формулировкой.

Еще один аспект связан с процедурами оценки. A. Kalai, O. Nachum, S. Vempala и E. Zhang указывают, что оценка по accuracy может стимулировать угадывание, если признание неопределенности не вознаграждается должным образом [8]. Для практики это означает: модель может быть обучена и настроена так, что попытка дать ответ оказывается более вероятной, чем осторожное сообщение «у меня нет доступа к документу». Если пользователь и система не требуют явного основания ответа, генерация догадки становится поведенчески удобной.

Теоретический результат A. Kalai и S. Vempala показывает структурные ограничения статистически калиброванных языковых моделей при генерации определенных типов фактов [7]. Это не значит, что любая БЯМ ошибается в каждом корпоративном вопросе. Галлюцинации нельзя свести к плохой инструкции, невнимательности пользователя или случайной поломке конкретной системы. Потому что генеративная модель не является механизмом достоверного доступа к произвольному внешнему факту.

В русскоязычной философской литературе Д. В. Зайцев пишет об ограничениях отождествления БЯМ с человеческим рассуждением [20]. Модель

может симулировать рассуждение в тексте, но эта симуляция не гарантирует наличия источника, процедуры проверки или ответственности за достоверность.

Уверенный ответ без знания возникает из сочетания нескольких факторов:

а) предсказания вероятного текста;

б) параметрической памяти;

в) жанровых шаблонов;

г) ориентации на полезность;

д) слабой калибровки неопределенности;

е) оценочных стимулов, поощряющих угадывание.

В корпоративном контексте это особенно опасно. Вопрос часто формулируется так, будто документ уже доступен, хотя технически модель его не видит.

4. Недоступный корпоративный документ как безответный вопрос

Вопрос по конкретному корпоративному документу отличается от вопроса общего знания. Когда пользователь спрашивает: «Какая процедура командирования действует в компании?», он может ожидать ответ из внутреннего положения о командировках. Если модель не получила этот документ, вопрос становится для нее безответным в строгом смысле.

Безответность здесь не означает, что ответа не существует в мире. Ответ может лежать в корпоративной системе электронного документооборота, на портале знаний, в архиве договоров, в базе поддержки или в локальной папке подразделения. Но для БЯМ без доступа к этим источникам он недоступен. Корректная система должна различать два состояния: «ответ неизвестен, потому что документа нет или он не найден» и «ответ найден в таком-то документе». Простая генеративная модель не всегда проводит эту границу.

В классической терминологии question answering можно сопоставить несколько режимов. Closed-book QA предполагает, что модель отвечает из своих параметров, без обращения к внешним источникам. Open-domain QA обычно ориентирован на широкий поиск по внешним коллекциям документов или веб-

источникам. Document-grounded QA требует, чтобы ответ опирался на заданный документ или набор документов. Enterprise search, в свою очередь, связан с поиском по внутренним корпоративным коллекциям, где важны права доступа, версии, релевантность, полнота и трассируемость результата. Корпоративный вопрос по документу относится именно к document-grounded и enterprise-сценарию, а не к свободной генерации.

Работа T. Gao, H. Yen, J. Yu и D. Chen о генерации ответов с цитированием показывает значимость provenance: пользователь должен понимать, откуда взят ответ и насколько ссылка действительно поддерживает утверждение [12]. В корпоративной среде это не академическая формальность. Ссылка на источник означает возможность проверить пункт регламента, страницу договора, номер приложения, дату приказа, статус версии. Без такой привязки ответ остается текстовой гипотезой.

Исследование J. Niu и соавторов RAGTruth показывает, что даже в системах с извлечением документов сохраняется проблема ответов, не поддержанных найденным контекстом [13]. Для данной статьи это работает как предостережение: наличие внешних документов само по себе не отменяет проблему groundedness. Тем более она не решается там, где внешних документов нет вообще. Если модель не видит документ, она не может проверить, существует ли в нем соответствующее положение.

J. Chen и соавторы анализируют бенчмаркинг БЯМ в retrieval-augmented generation и показывают, что для knowledge-intensive QA существенны способность использовать извлеченный контекст и контроль его применения [14]. Вопросы по корпоративным документам почти всегда являются knowledge-intensive: они требуют не только языковой формы, но и доступа к конкретному содержанию. Поэтому closed-book режим принципиально не совпадает с задачей, которую ставит пользователь.

Инженерная типология S. Barnett и соавторов, хотя и представлена как препринт, помогает описать практические точки отказа систем с извлечением: пропущенный контент, неправильное извлечение, неиспользование контекста,

неподходящий формат ответа и другие сбои [15]. Для настоящей статьи это имеет ограниченное значение, но хорошо показывает масштаб проблемы. Если даже системы с retrieval могут ошибаться на разных этапах, то прямой вызов БЯМ без retrieval вообще не содержит этапа доступа к корпоративному источнику.

Вопрос по недоступному корпоративному документу надо рассматривать как неотвечаемый для модели. Нормативно корректный ответ в такой ситуации звучит не как догадка, а как сообщение об отсутствии основания: «Документ не предоставлен», «Нет доступа к указанному регламенту», «Невозможно подтвердить ответ без текста договора», «Нужен актуальный документ или ссылка на него». Такая реакция менее удобна в коротком диалоге, зато честно отражает информационное состояние системы.

5. Знания модели и внешние корпоративные знания

Одна из причин неправильного использования БЯМ в организациях связана с неопределенностью самого слова «знание». Говоря, что модель «знает» какой-либо факт, часто имеют в виду, что она способна воспроизвести или правдоподобно сформулировать соответствующее утверждение. Но корпоративные знания устроены иначе. Они существуют как документы, процедуры, договоренности, роли, базы данных, версии, права доступа и практики обновления. Это не только текстовая информация, но и организационная инфраструктура.

Параметрические знания модели формируются во время обучения и донастройки. Они распределены по весам модели и не снабжены надежной ссылкой на конкретный источник. Пользователь обычно не может узнать, какой документ, фрагмент корпуса или статистический паттерн повлиял на ответ. Если модель утверждает, что «по регламенту срок согласования составляет пять рабочих дней», это утверждение может быть результатом усвоения типовых деловых формулировок, а не знания о конкретном регламенте организации.

Корпоративные знания, напротив, имеют признаки внешнего источника. Они принадлежат организации, поддерживаются владельцами процессов, меняются во времени, имеют статус актуальности и могут быть ограничены правами доступа. В статье О. И. Ковалевой, Е. А. Костыри и А. И. Василенко корпоративные системы управления знаниями рассматриваются как важный элемент современной организации [22]. Отсюда видно, что корпоративное знание нельзя свести к набору общих сведений: оно связано с информационными системами, процессами и управленческой практикой.

Различие особенно видно на примере версий. Универсальная БЯМ может знать общую структуру трудовых инструкций или договоров поставки, но не знает, какая версия локального положения действует с определенной даты, какие изменения внесены последним приказом и какие подразделения подпадают под исключение. Даже если похожий документ когда-то был включен в обучающий корпус, модель не может гарантировать, что воспроизводит именно актуальную версию. Для корпоративного решения это критично: устаревшая инструкция может быть не менее опасна, чем полностью выдуманная.

Второе различие касается полноты. Корпоративный документ часто содержит исключения, приложения, определения терминов, перекрестные ссылки, условия вступления в силу, зоны ответственности. Ответ, основанный на общих закономерностях, обычно стремится к краткой норме: «нужно согласовать с руководителем», «срок составляет десять дней», «ответственность несет исполнитель». Но юридически и организационно значимые детали могут находиться в примечаниях, приложениях или связанных документах. Без доступа к ним модель не может надежно восстановить полную картину.

Третье различие связано с проверяемостью. Корпоративное знание должно быть аудируемым: можно открыть документ, увидеть его номер, дату, автора, владельца процесса, статус утверждения, историю изменений. Параметрический ответ БЯМ не обладает такой трассируемостью. Он может быть верным случайно, но случайная верность не является достаточным основанием для корпоративного использования. В задачах, где требуется соблюдение

регламентов, важна не только правильность результата, но и возможность объяснить, почему он считается правильным.

Четвертое различие – это права доступа. Внутренние документы могут быть доступны не всем сотрудникам и тем более не внешней модели. Даже если технически модель могла бы ответить на похожий вопрос, корпоративная система должна учитывать, имеет ли конкретный пользователь право видеть соответствующий документ. Простая БЯМ без интеграции с корпоративной средой не способна самостоятельно соблюсти эти границы. Поэтому задача не сводится к интеллектуальной способности модели; она включает организационную и информационную безопасность.

Корпоративные знания включают не только explicit knowledge, закрепленное в документах, но и практики применения документов. При этом вопросно-ответная система, работающая с юридически и операционно значимыми материалами, должна отделять документально подтвержденное от интерпретационного. Если ответ основан на обычаях подразделения, это одно основание. Если он основан на утвержденном регламенте, это другое. Если основания нет, это третье состояние. Параметрическая генерация склонна смешивать эти уровни.

Знания модели и корпоративные знания различаются по происхождению, форме хранения, актуальности, полноте, проверяемости, правам доступа и организационному статусу. Поэтому вопрос «знает ли БЯМ ответ?» в корпоративной среде должен заменяться более строгим вопросом: «имеет ли система доступ к актуальному и разрешенному источнику, который подтверждает ответ?»

6. Почему галлюцинации особенно опасны в корпоративной среде

Галлюцинации БЯМ вредны в любой области, где требуется достоверность. В корпоративной среде риск выше из-за характера самих документов. Здесь часто используются регламенты, положения, инструкции, договоры, политики, стандарты, протоколы, отчеты. Ошибка в таком документе

может повлиять на обязательства, сроки, ответственность, финансовые решения и соблюдение требований.

Корпоративные вопросы часто задаются в условиях дефицита времени. Сотрудник обращается к системе, чтобы быстро узнать порядок действия. Если ответ звучит уверенно, его могут принять без дополнительной проверки.

В юридических сценариях риск особенно очевиден. M. Dahl и соавторы анализируют юридические галлюцинации БЯМ на проверяемых вопросах о судебных делах и праве [16]. V. Magesh показывает, что даже специализированные ИИ-инструменты юридического поиска требуют проверки надежности и не гарантируют полного отсутствия ошибочных ответов [17]. Для корпораций это напрямую применимо к договорам, внутренним документам, трудовым регламентам. Ошибка модели может выглядеть как юридическая консультация, хотя может являться выдумкой.

В договорной работе галлюцинация часто выглядит не как грубая ошибка, а как почти правдоподобная деталь. Модель может сослаться не на тот пункт договора, придумать условие об одностороннем расторжении, назвать неверный срок уведомления, смешать старую и новую редакции или перенести типовое условие на конкретный документ. Дальше все зависит от ситуации: можно пропустить срок, отправить контрагенту неверное письмо, неправильно оценить риск, нарушить порядок согласования. Особенно опасны выдуманные реквизиты. Пользователь видит номер пункта и может решить, что такой пункт правда есть.

В регламентах и инструкциях ошибка бьет уже по внутреннему процессу. Модель, например, может написать, что закупка до определенной суммы согласуется только руководителем подразделения, хотя действующий регламент требует участия финансового контроля. Или дать старый порядок оформления командировки и не учесть новые требования к подтверждающим документам. Такая ошибка не всегда видна сразу. Часто ее замечают позже, когда действие уже выполнено и процесс пошел неправильно.

Во внутренней отчетности и аналитике риск связан с ложной уверенностью в данных. Модель может сделать вывод о динамике показателей, сослаться на таблицу, которой нет, придумать источник цифры или рассуждать об отчете, который ей не предоставляли. Исследование о несуществующих цитированиях [18] показывает тот же механизм: генеративная система способна создать убедительный реквизит. В корпоративном отчете таким реквизитом становится номер раздела, название таблицы, ссылка на приложение или якобы утвержденный показатель.

В базе знаний службы поддержки галлюцинация превращается в неправильную инструкцию для клиента или сотрудника. Модель может предложить восстановление доступа способом, которого нет в политике безопасности, или сообщить о возврате, не предусмотренном правилами обслуживания. Если такой ответ попадает в рабочий процесс, это уже не просто неточность в тексте. Возникает риск для сервиса, безопасности и доверия пользователей.

Есть еще проблема масштаба. Человеческая ошибка обычно остается локальной: один сотрудник неверно понял документ. Ошибка БЯМ, встроенной в корпоративный интерфейс, может повторяться много раз и расходиться по подразделениям. Если система не показывает источник и не обозначает неопределенность, ложное утверждение постепенно начинает жить как внутреннее знание.

Отдельный вопрос – ответственность. Сотрудник принял решение по ответу модели. Кто должен был проверить документ? Кто отвечает за неверную рекомендацию? Можно ли восстановить, откуда появилась ошибка? Видел ли пользователь, что у ответа нет основания? Документальная обоснованность нужна не только разработчикам системы. Она нужна управлению.

Галлюцинации также снижают доверие к системам управления знаниями. Если сотрудники несколько раз получают уверенные, но неверные ответы, они перестают воспринимать ИИ-интерфейс как надежный канал доступа к

информации. Инструмент, который должен был ускорять работу со знаниями, начинает требовать перепроверки каждого результата.

7. Связь проблемы с QA, closed-book QA, open-domain QA, enterprise search и document-grounded QA

Эта проблема относится сразу к нескольким исследовательским и прикладным областям. Самая общая из них - question answering, то есть системы, которые отвечают на вопросы пользователя. Но внутри QA есть разные постановки задачи. Когда их смешивают, от БЯМ начинают ждать того, чего она в данном режиме сделать не может.

Closed-book QA означает, что система отвечает без обращения к внешним документам в момент вывода. Ответ строится на параметрах модели. Такой режим подходит для общих вопросов, черновых формулировок, объяснения известных понятий, первичной ориентации в теме. Но он плохо подходит для вопроса «что сказано в конкретном документе?». Если документ не передан модели, closed-book ответ нельзя надежно подтвердить этим документом.

Open-domain QA работает с широким набором источников. В классическом варианте это веб, энциклопедии, новостные корпуса или другие открытые коллекции. Корпоративный документ в такой среде может отсутствовать. Часто он и должен отсутствовать, потому что является внутренним. Значит, open-domain подход не решает корпоративную задачу сам по себе. Нужен доступ именно к внутреннему контуру документов.

Enterprise search устроен иначе: он ищет по корпоративным коллекциям. Здесь важны не только релевантность и полнота, но и права доступа, версия документа, метаданные, владелец процесса, требования безопасности. Корпоративные источники неоднородны, часто закрыты и постоянно обновляются. Поэтому ответ сотруднику должен опираться не на вероятную формулировку, а на найденный документ или фрагмент.

Document-grounded QA ближе всего к вопросам по регламентам, договорам, инструкциям и отчетам. В этой постановке ответ строится на

заданном документе или наборе документов. Качество определяется не только связностью текста. Нужно, чтобы каждое существенное утверждение поддерживалось источником. Работа Gao и соавторов о генерации с цитированием показывает, почему проверка ссылок и подтверждение ответов источниками становятся отдельной задачей [12].

Галлюцинация возникает тогда, когда пользователь фактически задает document-grounded вопрос, а система отвечает, как в closed-book QA. Пользователь спрашивает: «Что сказано в положении о премировании?» Система не имеет этого положения, но пишет правдоподобный ответ о типовой премиальной политике. Для closed-book генерации текст может выглядеть удачным. Для document-grounded QA это ошибка: ответа по документу не было.

Здесь расходятся задача пользователя и информационный контур системы. Если вопрос требует внешнего источника, а модель использует только параметры, результат становится ненадежным. Даже правильный ответ в таком режиме остается случайным попаданием, потому что система не может показать документ, которым он подтвержден. Исследования RAG-сценариев [13; 14] как раз пытаются соединить генерацию и внешний контекст.

8. Почему простое обращение к БЯМ не решает корпоративный вопросно-ответный поиск

Пользователь задает вопрос, получает связный ответ, видит деловой стиль и логичную структуру. Но с точки зрения требований к корпоративному знанию отсутствует поиск источника, проверка актуальности и возможность отказаться от ответа при отсутствии основания.

Первое ограничение – отсутствие доступа. Если модель не получила документ, она не может извлечь из него информацию. Это кажется очевидным, но на практике маскируется языковой компетентностью БЯМ. Модель умеет писать о договорах, регламентах и инструкциях, поэтому пользователь может принять общую компетентность за знание конкретного файла. Но знание жанра не равно знанию содержания.

Второе – отсутствие версии. Корпоративные документы изменяются. Ответ по старому регламенту может быть хуже, чем отсутствие ответа, потому что он вводит в заблуждение. Простая БЯМ не знает, какая версия документа действует сейчас, если версия не передана ей явно. Даже если модель обучалась на похожем тексте, это не дает гарантии актуальности.

Третье ограничение – отсутствие источниковой проверки. Для корпоративного QA недостаточно, чтобы ответ «звучал правильно». Нужно иметь возможность открыть источник и убедиться, что утверждение действительно выводится из него. Исследования генерации с цитированием [12] и groundedness в retrieval-augmented сценариях [13] показывают, что проверяемость источников становится отдельным критерием качества. В простом вызове БЯМ такой критерий обычно не встроен.

Четвертое ограничение – смешение общего и локального знания. Модель может дать ответ, основанный на типовой практике отрасли, публичных текстах или статистически частых формулировках. Но корпоративный документ может содержать локальное исключение. Например, типовой порядок согласования может включать руководителя, юридический отдел и бухгалтерию, а в конкретной организации дополнительно требуется согласование службы безопасности. Без документа модель не может узнать это исключение.

Пятое ограничение – слабое управление неизвестностью. Исследования known-unknowns и self-detection [9; 10; 11] показывают, что способность распознавать отсутствие знания является отдельной проблемой. В корпоративном поиске отказ от ответа при отсутствии документа должен быть нормальным результатом, а не сбоем. Простая БЯМ часто воспринимается пользователем как система, обязанная отвечать, и потому генерирует догадку.

Шестое ограничение – риск фабрикации реквизитов. Если пользователь просит указать пункт, дату, номер приказа или ссылку на документ, модель может создать правдоподобный реквизит. Примеры фабрикации цитирований [18] показывают, что такая ошибка реалистична. В корпоративной среде она

приобретает практическую значимость. Вымышленная информация может попасть в письмо, отчет, служебную записку или юридическое заключение.

Седьмое ограничение – отсутствие корпоративного контекста доступа. Внутренние документы могут быть разграничены по ролям. Ответ, допустимый для одного сотрудника, может быть недоступен другому. Простая БЯМ без интеграции с корпоративными правами не различает эти состояния. Поэтому даже технически правильная генерация может быть организационно некорректной.

Восьмое ограничение – невозможность аудита. Если ответ с ошибкой, организация должна понять, какой источник использовался и почему система решила ответить. В простом диалоге с БЯМ это невозможно. Ответ является продуктом генерации, а не результатом трассируемого поиска по документам.

Поэтому прямой вызов БЯМ может быть полезен как вспомогательный инструмент для переформулирования текста, объяснения общих понятий или подготовки черновика. Но он не является надежной системой корпоративного вопросно-ответного поиска. Для последней требуется архитектура, где внешние документы становятся явным основанием ответа, а отсутствие основания приводит к отказу от содержательной генерации.

9. Корпоративные примеры неподтвержденной генерации

Для более точного понимания риска рассмотрим несколько типовых ситуаций. Они являются не эмпирическими кейсами конкретной организации, а аналитическими сценариями, вытекающими из природы проблемы.

Вопрос по регламенту закупок. Пользователь спрашивает: «Можно ли заключить договор без конкурсной процедуры, если сумма меньше 500 тысяч рублей?». Если модель не имеет доступа к актуальному регламенту, она может ответить на основе типовых практик. Но конкретная организация может требовать дополнительного согласования, иметь иной порог суммы или запрещать исключение для определенных категорий товаров. Ошибка здесь не только фактическая, но и процедурная.

Вопрос по договору. Пользователь спрашивает: «Какая неустойка предусмотрена за просрочку поставки?». Модель может сгенерировать типовую формулу, например, процент от суммы просроченного обязательства за каждый день. Но в договоре может быть фиксированный штраф или ограничение общей ответственности. Без текста договора ответ является выдумкой или догадкой.

Вопрос по инструкции информационной безопасности. Пользователь уточняет, можно ли передать файл подрядчику через личную почту. Общая модель может ответить стандартными практиками безопасности, но корпоративная политика может содержать конкретный запрет. Неправильный ответ способен создать риск утечки данных.

Вопрос по внутренней отчетности. Руководитель нужен показатель выполнения плана у отдела за прошлый квартал. Без доступа к отчету, БЯМ может попытаться сформулировать аналитический ответ. Но числовые показатели должны извлекаться из конкретного источника. Любая генерация числа без документального основания является недопустимой.

Завершающий сценарий – база знаний поддержки. Оператор спрашивает, что отвечать клиенту при ошибке активации лицензии. Модель может предложить универсальную инструкцию, но актуальная база знаний может требовать другой последовательности действий. Ошибка в ответе может тиражироваться в клиентском сервисе.

Во всех этих сценариях структура похожа. Пользователь задает вопрос, предполагающий наличие источника, а модель без источника отвечает на основе вероятности. Ответ может быть разумны, но неверным для конкретной организации. Поэтому главный критерий качества не красота формулировки, а наличие проверяемого основания.

10. Неопределенность и право модели на отказ от ответа

В корпоративной культуре отказ системы от ответа часто воспринимается как недостаток. Пользователь ожидает помощи, а сообщение «нет доступа к документу» кажется менее полезным, чем попытка сформулировать хотя бы

предварительный ответ. В контексте работы с документами отказ при отсутствии основания – не слабость, а элемент надежности.

Исследования self-detection и known-unknowns показывают, что распознавание неизвестного требует отдельного внимания [9; 10; 11]. Вопрос состоит не только в том, может ли модель сгенерировать текст. Вопрос в том, должна ли она это делать. Если задача требует конкретного документа, то отсутствие документа является достаточным основанием для отказа от содержательного ответа. Модель может помочь пользователю уточнить запрос, запросить файл или объяснить, какие данные нужны, но не должна выдавать догадку за факт.

С практической точки зрения полезно различать несколько уровней ответа. Первый уровень – система нашла документ, релевантный фрагмент и может сослаться на него. Второй уровень - ограниченное объяснение. Система не имеет документа, но может дать общую справочную информацию, явно указав, что она не является ответом по конкретному документу. Третий уровень – отказ. Система сообщает, что без источника нельзя ответить. В корпоративных сценариях третий уровень часто безопаснее, чем неподтвержденная генерация.

Проблема состоит в том, что языковая форма ответа может скрыть отсутствие основания. Фраза «Согласно регламенту...» звучит как ссылка на документ даже тогда, когда документ не был найден. Поэтому корпоративная система должна быть осторожна с формулами, создающими видимость источниковой привязки. Если источник не доступен, нельзя использовать речевые обороты «в документе указано», «пункт предусматривает», «согласно договору», «регламент устанавливает».

Эта логика связана с этикой и ответственностью. Hicks, Humphries и Slater критикуют склонность воспринимать БЯМ как источник истины и обращают внимание на проблему безразличия к истинностной стороне высказывания [19]. В корпоративном контексте это безразличие недопустимо, потому что ответ включается в практику управления. Отказ от ответа при отсутствии основания

является способом восстановить границу между текстовой генерацией и знанием.

11. Требования к надежному корпоративному вопросно-ответному ответу

Если простая генерация не решает задачу, возникает вопрос о минимальных требованиях. Как должен соответствовать надежный ответ по корпоративному документу?

Первое требование – наличие источника. Ответ должен быть основан на конкретном документе или наборе документов, доступных системе и пользователю. Если источника нет, система должна сообщить об этом. Это требование отделяет document-grounded QA от closed-book генерации.

Второе требование – актуальность. Система должна учитывать версию документа, дату утверждения, статус действия и возможные изменения. Ответ по устаревшему документу должен маркироваться как рискованный или недействительный, если актуальность не подтверждена.

Третье требование – локализация основания. Пользователь должен видеть какой пункт, страница, фрагмент. Работа Gao и соавторов показывает значимость цитирования для генерации ответов [12]; в корпоративной среде аналогом является ссылка на внутренний источник.

Четвертое требование – верность контексту. Ответ не должен выходить за пределы найденного основания. Если документ говорит только о порядке согласования, система не должна делать вывод о юридической ответственности, когда такой вывод не вытекает из текста. Проблема unsupported generation сохраняется даже при наличии документа, если модель добавляет неподтвержденные детали [13].

Пятое требование – управление неизвестностью. Если документ не найден, не содержит ответа, противоречив или недоступен по правам, система должна уметь отказаться от ответа или явно обозначить неопределенность. Это требование связано с исследованиями knowledge boundaries и self-detection [9; 10; 11].

Шестое требование – разграничение общего и корпоративного знания. Модель может дать общее объяснение, но должна отделять его от ответа по внутреннему документу. Например: «В типовой практике возможно несколько вариантов, но без текста договора нельзя сказать, какой применяется здесь». Такая формулировка полезнее, чем уверенное утверждение без источника.

Седьмое требование – проверяемость и аудит. Ответ должен оставлять след: какой документ использовался, какая версия, какие фрагменты, когда был выполнен запрос. Это также нужно для принятия решений.

Эти требования показывают, почему задача не сводится к выбору более умной БЯМ. Самая сильная языковая модель не может надежно ответить на вопрос по документу, если она его не получила. Улучшение языковой способности повышает только качество формулировки.

Заключение

Вопрос по корпоративному документу нельзя сводить к обычному запросу к языковой модели. Пользователь спрашивает не «что обычно бывает в таких случаях», а что написано в конкретном регламенте, договоре, инструкции, отчете или базе знаний. Если этих материалов у модели нет, она не отвечает по документу. Она строит вероятный текст по запросу. Иногда такой текст выглядит убедительно, но источника под ним нет.

В работе были разделены несколько близких ошибок. Factual hallucination означает ложное фактическое утверждение. Fabrication связана с созданием несуществующих объектов: пунктов, ссылок, реквизитов, документов. Confabulation проявляется в смысловой нестабильности при внешне связном ответе. Unsupported generation особенно важна для корпоративных задач: ответ может звучать разумно и даже совпадать с типовой практикой, но оставаться неприемлемым, если он не подтвержден нужным документом.

Уверенность БЯМ в таких ситуациях объясняется не наличием знания, а способом ее работы. Модель предсказывает вероятный текст, использует параметрическую память, жанровые шаблоны и привычные формулы ответа.

Она может быть настроена на помощь пользователю и поэтому продолжать отвечать даже там, где корректнее было бы остановиться. Исследования 2023-2026 гг. показывают, что распознавание неизвестного и отказ от ответа не появляются сами по себе. Это отдельная задача для модели и для системы, в которую она встроена.

Корпоративные знания устроены иначе, чем знания модели. Они находятся во внешних источниках, меняются по версиям, имеют владельцев, даты, права доступа, историю утверждения и практику применения. Поэтому вопрос по корпоративному документу требует не closed-book QA, a document-grounded QA и enterprise search. Здесь качество ответа определяется не гладкостью формулировки, а тем, можно ли открыть источник и проверить сказанное.

Практический риск состоит в том, что ошибка БЯМ попадает не в отвлеченный текст, а в рабочий процесс. Неверный пункт договора, устаревшая инструкция, выдуманный показатель отчета или неправильная запись из базы знаний могут повлиять на решение сотрудника. И чем удобнее ИИ-интерфейс, тем быстрее такая ошибка распространяется, если система не показывает источник и не обозначает отсутствие основания.

Поэтому прямой запрос к БЯМ без доступа к корпоративным документам не решает задачу корпоративного вопросно-ответного поиска. Он может помочь с общими пояснениями, переформулировкой или черновиком. Но ответ по внутреннему источнику требует самого источника. Дальнейший анализ должен быть посвящен архитектурам, где документ становится явной основой ответа.

Список литературы

1. Ji Z., Lee N., Frieske R., Yu T., Su D., Xu Y., Ishii E., Bang Y. J., Madotto A., Fung P. Survey of Hallucination in Natural Language Generation // ACM Computing Surveys. 2023. Vol. 55, No. 12. P. 1-38. URL: https://doi.org/10.1145/3571730

2. Huang L., Yu W., Ma W., Zhong W., Feng Z., Wang H., Chen Q., Peng W., Feng X., Qin B., Liu T. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions // ACM Transactions on Information Systems. 2025. Vol. 43, No. 2. P. 1-55. URL: https://dl.acm.org/doi/10.1145/3703155

3. McKenna N., Li T., Cheng L., Hosseini M., Johnson M., Steedman M. Sources of Hallucination by Large Language Models on Inference Tasks // Findings of EMNLP 2023. Association for Computational Linguistics, 2023. P. 2758-2774. URL: https://doi.org/10.18653/v1/2023.findings-emnlp.182

4. Manakul P., Liusie A., Gales M. J. F. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models // Proceedings of EMNLP 2023. Association for Computational Linguistics, 2023. P. 9004-9017. URL: https://doi.org/10.18653/v1/2023.emnlp-main.557

5. Slobodkin A., Goldman O., Caciularu A., Dagan I., Schuster T. The Curious Case of Hallucinatory (Un)answerability: Finding Truths in the Hidden States of Over-Confident Large Language Models // Proceedings of EMNLP 2023. Association for Computational Linguistics, 2023. URL: https://aclanthology.org/2023.emnlp-main.220/

6. Farquhar S., Kossen J., Kuhn L., Gal Y. Detecting hallucinations in large language models using semantic entropy // Nature. 2024. Vol. 630. P. 625-630. URL: https://doi.org/10.1038/s41586-024-07421-0

7. Kalai A. T., Vempala S. S. Calibrated Language Models Must Hallucinate // Proceedings of the 56th Annual ACM Symposium on Theory of Computing. ACM, 2024. P. 160-171. URL: https://doi.org/10.1145/3618260.3649777

8. Kalai A. T., Nachum O., Vempala S. S., Zhang E. Evaluating large language models for accuracy incentivizes hallucinations // Nature. 2026. Online first, 22.04.2026. URL: https://doi.org/10.1038/s41586-026-10549-w

9. Yin Z., Sun Q., Guo Q., Wu J., Qiu X., Huang X. Do Large Language Models Know What They Don't Know? // Findings of ACL 2023. Association for Computational Linguistics, 2023. URL: https://aclanthology.org/2023.findings-acl.551/
10. Zhang H., Xiong Z., Zhang B., Li C., Li Z., Xiang J., Ding D., Yin X., Hu Y., Liu L., Jin D. Knowing What LLMs DO NOT Know: A Simple Yet Effective Self-Detection Method // Proceedings of NAACL 2024. Association for Computational Linguistics, 2024. URL: https://aclanthology.org/2024.naacl-long.15/
11. Amayuelas A., Pan L., Chen W., Wang W. Y. Knowledge of Knowledge: Exploring Known-Unknowns Uncertainty with Large Language Models // Findings of ACL 2024. Association for Computational Linguistics, 2024. URL: https://aclanthology.org/2024.findings-acl.36/
12. Gao T., Yen H., Yu J., Chen D. Enabling Large Language Models to Generate Text with Citations // Proceedings of EMNLP 2023. Association for Computational Linguistics, 2023. URL: https://aclanthology.org/2023.emnlp-main.398/
13. Niu J., Li X., Liu Y., Zhang L., Jiang H., Wang R., Gao Y., Chen Y., Liu Z., Li Z., Zhou B. RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models // Proceedings of ACL 2024. Association for Computational Linguistics, 2024. URL: https://aclanthology.org/2024.acl-long.585/
14. Chen J., Lin H., Han X., Sun L. Benchmarking Large Language Models in Retrieval-Augmented Generation // Proceedings of the AAAI Conference on Artificial Intelligence. 2024. Vol. 38, No. 16. P. 17754-17762. URL: https://doi.org/10.1609/aaai.v38i16.29728
15. Barnett S., Kurniawan S., Thudumu S., Brannelly Z., Abdelrazek M. Seven Failure Points When Engineering a Retrieval Augmented Generation System: arXiv preprint. 2024. URL: https://arxiv.org/abs/2401.05856
16. Dahl M., Magesh V., Suzgun M., Ho D. E. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models // Journal of Legal Analysis. 2024. Vol. 16, No. 1. P. 64-93. URL: https://academic.oup.com/jla/article/16/1/64/7699227

17. Magesh V., Surani F., Dahl M., Suzgun M., Manning C. D., Ho D. E. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools // Journal of Empirical Legal Studies. 2025. URL: https://doi.org/10.1111/jels.12413
18. Buchanan J., Hill S., Shapoval O. ChatGPT Hallucinates Non-existent Citations: Evidence from Economics // The American Economist. 2024. URL: https://doi.org/10.1177/05694345231218454
19. Hicks M. T., Humphries J., Slater J. ChatGPT is bullshit // Ethics and Information Technology. 2024. Vol. 26. Article 38. URL: https://doi.org/10.1007/s10676-024-09775-5
20. Зайцев Д. В. Почему большие языковые модели не (всегда) рассуждают как люди? // Вестник Московского университета. Серия 7: Философия. 2024. Т. 48, N 1. С. 76-93. URL: https://www.elibrary.ru/item.asp?id=60777220
21. Александров О. А., Чистова Е. В. Большие языковые модели как инструмент практической деятельности переводчика: обзор зарубежных научных проектов // Язык и культура. 2025. N 72. С. 8-32. URL: https://www.elibrary.ru/item.asp?id=84978789
22. Ковалева О. И., Костыря Е. А., Василенко А. И. Информационные системы управления знаниями в современных организациях // Сервис plus. 2026. Т. 20, N 1. С. 3-12. URL: https://www.elibrary.ru/item.asp?id=89113444
23. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. Attention Is All You Need // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 5998-6008. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html