

## ГАЛЛЮЦИНАЦИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: ПОЧЕМУ ГЕНЕРАТИВНЫЕ МОДЕЛИ УВЕРЕННО ДАЮТ НЕВЕРНЫЕ ОТВЕТЫ

Раушанбек Н.Е., магистр технических наук, [Nurda\\_e@bk.ru](mailto:Nurda_e@bk.ru)

*Кызылординский открытый университет, г. Кызылорда, Республика Казахстан*

**Аннотация.** Современные системы генеративного искусственного интеллекта способны формировать связные и убедительно звучащие ответы, однако правдоподобная форма не гарантирует фактическую точность. Ошибочные, вымышленные или не подтверждённые сведения, представленные моделью как достоверные, обычно называют галлюцинациями, а в документах NIST используется также термин «конфабуляция». Цель статьи – систематизировать причины, по которым большие языковые модели уверенно дают неверные ответы, рассмотреть основные типы таких ошибок и определить практические способы снижения риска. Исследование выполнено в форме аналитического обзора научных публикаций 2020–2025 годов, международных методических документов и работ по оценке фактичности языковых моделей. Показано, что галлюцинации связаны не с одной причиной, а с совокупностью факторов: статистической природой генерации, неполнотой и противоречивостью обучающих данных, ограничениями знаний, особенностями обучения и оценки, неоднозначностью запроса и ошибками внешнего поиска. Сделан вывод, что надёжное применение генеративных моделей требует сочетания проверки источников, retrieval-augmented generation, методов оценки неопределённости, самопроверки и обязательного человеческого контроля в ответственных задачах.

**Ключевые слова:** искусственный интеллект, большая языковая модель, галлюцинация, конфабуляция, фактическая точность, генеративная модель, проверка информации.

**Андатпа.** Генеративті жасанды интеллект жүйелері табиғи тілде мазмұнды, логикалық және сенімді көрінетін жауаптар құрастыра алады, бірақ олардың жауабы әрдайым шынайы деректерге сәйкес келе бермейді. Мұндай құбылыс ғылыми әдебиетте «галлюцинация» немесе «конфабуляция» деп аталады. Мақаланың мақсаты – үлкен тілдік модельдердің неліктен қате ақпаратты сенімді түрде ұсынатынын түсіндіру, галлюцинациялардың негізгі түрлері мен пайда болу себептерін жүйелеу және оларды азайтуға арналған практикалық тәсілдерді қарастыру. Зерттеу шолу-талдамалық сипатта жүргізілді: 2020–2025 жылдардағы рецензияланған еңбектер, халықаралық нұсқаулықтар және генеративті жасанды интеллект тәуекелдерін сипаттайтын нормативтік-әдістемелік материалдар салыстырылды. Талдау нәтижесінде галлюцинациялардың деректер сапасына, модельді оқыту мақсатына, келесі токенді ықтималдықпен болжау принципіне, сұрау контекстіне, білім шекарасына және бағалау жүйелеріне байланысты қалыптасатыны көрсетілді. Сондай-ақ дереккөздермен тексеру, іздеумен толықтырылған генерация, модельдің белгісіздікті мойындауы, автоматты тексеру және адам бақылауы сияқты әдістердің кешенді қолданылуы ұсынылады.

**Түйінді сөздер:** жасанды интеллект, үлкен тілдік модель, галлюцинация, конфабуляция, фактілік дәлдік, генеративті модель, деректерді тексеру.

**Введение.** Генеративный искусственный интеллект за короткое время превратился из специализированной исследовательской технологии в массовый инструмент для работы с текстами, программным кодом, изображениями и данными. Большие языковые модели используются для подготовки учебных материалов, поиска объяснений, написания программ, анализа документов, составления резюме и поддержки исследовательской деятельности. Их ключевое преимущество состоит в способности формировать ответ в естественной форме, подстраиваться под контекст и быстро синтезировать информацию. Однако именно высокая языковая связность создаёт особый риск: пользователь может воспринимать грамматически правильный и логично звучащий текст как достоверный, даже если в нём присутствуют фактические ошибки.

В научной литературе подобные ошибки принято описывать термином «галлюцинация». Один из наиболее цитируемых обзоров определяет галлюцинации как генерацию содержания, которое не соответствует исходным данным, контексту или проверяемым фактам [1]. Национальный институт стандартов и технологий США применяет

близкий термин «confabulation» и подчёркивает, что генеративная система способна уверенно представить ошибочный или ложный контент, включая вымышленные объяснения и ссылки [2]. Существенной особенностью проблемы является не только наличие ошибки, но и отсутствие очевидных внешних признаков, по которым неспециалист мог бы сразу отличить корректный ответ от вымышленного.

Проблема фактичности появилась не вместе с современными чат-ботами. Ещё до массового распространения разговорных моделей исследователи показали, что языковые модели могут воспроизводить распространённые человеческие заблуждения. В тесте TruthfulQA, содержащем 817 вопросов из 38 категорий, лучшая из протестированных на тот момент моделей дала правдивые ответы примерно на 58% вопросов, тогда как показатель людей составлял 94% [3]. Эти результаты относятся к моделям, исследованным в 2022 году, и не должны переноситься напрямую на современные системы, однако они хорошо иллюстрируют принципиальную проблему: увеличение способности модели продолжать текст не тождественно приобретению механизма проверки истинности.

Актуальность темы особенно высока в образовании и научной работе. Студент может получить убедительно оформленное, но неверное объяснение формулы; преподаватель – несуществующую ссылку на статью; программист – код, содержащий вызов вымышленной функции; исследователь – правдоподобное описание источника, которого не существует. UNESCO рекомендует образовательным учреждениям не ограничиваться техническим внедрением генеративных систем, а оценивать их педагогическую и этическую пригодность, сохранять человеческую ответственность и развивать навыки проверки результатов [8]. Поэтому понимание природы галлюцинаций становится частью современной цифровой и информационной грамотности.

Цель данной статьи – объяснить, почему генеративные языковые модели могут уверенно давать неправильные ответы, систематизировать виды и причины галлюцинаций, а также обобщить методы их обнаружения и снижения. В работе решаются следующие задачи: уточнить содержание понятия «галлюцинация»; рассмотреть статистические и организационные причины возникновения ошибок; сравнить основные подходы к их выявлению; сформулировать практические рекомендации для пользователей и образовательных организаций.

**Материалы и методы исследования.** Исследование носит обзорно-аналитический характер. В качестве основной источниковой базы использованы рецензируемые статьи и материалы конференций по фактичности и галлюцинациям в генерации естественного языка, опубликованные в 2020-2025 годах, а также международные методические документы, посвящённые рискам генеративного искусственного интеллекта. Ключевые публикации отбирались по четырём тематическим блокам: определение и классификация галлюцинаций; причины возникновения; методы обнаружения; способы снижения риска.

Методологическую основу составили сравнительный анализ, контент-анализ, классификация и синтез. Сначала были сопоставлены определения галлюцинации в исследованиях по генерации естественного языка и в документах по управлению рисками. Затем причины ошибок были распределены по этапам жизненного цикла языковой модели: данные, обучение, вывод ответа и взаимодействие с внешними источниками. После этого методы снижения риска были сгруппированы по уровню применения: действия пользователя, архитектурные решения, автоматическая проверка и организационные процедуры.

Отдельное внимание уделялось различию между установленными результатами исследований и практическими рекомендациями. В работе не использованы вымышленные экспериментальные данные, не проводится собственное тестирование конкретных коммерческих моделей и не делается вывод о «проценте галлюцинаций» какой-либо современной системы. Это важно, поскольку частота ошибок существенно зависит от модели, версии, предметной области, сложности запроса, наличия поиска, языка и методики оценки. Поэтому выводы статьи относятся прежде всего к механизмам и типовым рискам, а

не к рейтингу отдельных продуктов.

**Результаты и их обсуждение.** Проведённый анализ показывает, что термин «галлюцинация» объединяет несколько разных типов ошибок. Современный обзор больших языковых моделей предлагает различать как минимум фактические и связанные с верностью контексту ошибки [10]. Фактическая галлюцинация возникает, когда ответ противоречит проверяемому факту или содержит вымышленное утверждение. Ошибка верности контексту наблюдается, когда модель искажает предоставленный пользователем текст, нарушает инструкцию, противоречит собственному предыдущему ответу или строит логически несовместимое рассуждение. Такое разделение полезно, потому что разные виды ошибок требуют разных способов контроля.

На практике галлюцинация не всегда выглядит как очевидная фантазия. Наиболее опасны ответы, в которых большая часть текста верна, но одна деталь оказывается вымышленной: неверная дата, название закона, DOI, фамилия автора, параметр функции или числовое значение. Языковая плавность при этом сохраняется. Пользователь получает правильную по форме структуру – введение, аргументы, список источников, пояснение – и поэтому склонен переносить доверие к форме на содержание. NIST отдельно обращает внимание на риск вымышленных ссылок и рассуждений, которые могут создавать дополнительное ощущение обоснованности [2].

Первая фундаментальная причина связана с самим принципом работы языковой модели. При предобучении модель учится предсказывать следующий токен на основе статистических закономерностей огромного корпуса текстов. Она оптимизируется не как база данных, в которой каждому утверждению соответствует поле «истина/ложь», а как вероятностная система продолжения последовательности. Поэтому высокий уровень владения языком и знание распространённых закономерностей не гарантируют точного воспроизведения редких, случайных или недостаточно представленных фактов. В работе Kalai и соавторов показано, что часть галлюцинаций закономерно возникает из статистических ошибок различения истинных и ложных утверждений на этапе предобучения [9].

Вторая причина – качество и структура обучающих данных. Интернет-среда содержит ошибки, устаревшую информацию, противоречащие друг другу утверждения, неполные документы и тексты разного качества. Если один факт многократно представлен в ошибочной форме, модель может воспроизводить этот шаблон. Если же корректная информация встречается редко, её точное восстановление становится сложнее. Кроме того, знания модели ограничены тем, что было представлено в данных и закреплено в её параметрах. Запрос о малоизвестном человеке, локальном документе, новом событии или узкоспециализированной детали повышает вероятность того, что система будет строить правдоподобное продолжение вместо точного воспроизведения факта.

Третья причина связана с настройкой моделей на полезный и разговорный ответ. Пользователь обычно ожидает, что система ответит, а не остановится фразой «не знаю». В исследовании 2025 года предложено объяснение, согласно которому распространённые схемы оценивания поощряют угадывание: если метрика учитывает только число правильных ответов, попытка ответить при неопределённости иногда даёт выигрыш, тогда как отказ почти всегда оценивается как нулевой результат [9]. В такой среде «уверенный ответ» может быть статистически выгоднее корректного признания недостатка информации. Следовательно, снижение галлюцинаций требует не только улучшения моделей, но и изменения критериев оценки так, чтобы способность признавать неопределённость считалась положительным свойством.

Четвёртая причина – неоднозначность и недостаточность пользовательского запроса. Если в вопросе не указан период, версия программы, страна, предметная область или критерий сравнения, модель вынуждена самостоятельно выбирать интерпретацию. Она может заполнить пропуски наиболее вероятными предположениями, которые не совпадут с намерением пользователя. Длинный диалог также создаёт риск потери или смешения

контекста. Поэтому часть того, что воспринимается как «галлюцинация модели», фактически возникает на границе между неопределённым запросом и стремлением системы дать завершённый ответ.

Пятая причина появляется даже в системах, подключённых к поиску или корпоративной базе знаний. Retrieval-augmented generation (RAG) сочетает языковую модель с извлечением внешних документов. Классическая работа Lewis и соавторов показала, что добавление внешней памяти может повысить фактичность и специфичность генерации по сравнению с полностью параметрической моделью [4]. Однако RAG не устраняет проблему автоматически. Если система нашла нерелевантный документ, неверно ранжировала источники, получила устаревший фрагмент или неправильно интерпретировала найденный текст, итоговый ответ всё равно может быть ложным. Современные обзоры прямо указывают, что системы с поиском также подвержены галлюцинациям [10].

**Таблица 1** – Основные типы галлюцинаций генеративных языковых моделей

| Тип ошибки    | Как проявляется                                      | Условный пример                                | Основной риск                                     |
|---------------|------------------------------------------------------|------------------------------------------------|---------------------------------------------------|
| Фактическая   | Утверждение противоречит проверяемому факту          | Неверная дата, имя, характеристика технологии  | Пользователь принимает ложное утверждение за факт |
| Атрибутивная  | Модель выдумывает источник, автора, DOI или цитату   | Несуществующая статья оформляется как реальная | Ложное подтверждение и академические ошибки       |
| Контекстная   | Ответ противоречит документу или данным пользователя | В пересказ добавляется отсутствующий факт      | Искажение исходной информации                     |
| Инструктивная | Нарушаются условия запроса                           | Запрещённый формат или пропущенное ограничение | Несоответствие задачи и результата                |
| Логическая    | Ответ содержит внутреннее противоречие               | Вывод не следует из приведённых посылок        | Ошибочные решения при видимой связности текста    |
| Временная     | Используется устаревшее состояние мира               | Старая версия правила выдаётся за актуальную   | Ошибки в быстро меняющихся областях               |

Примечание – классификация обобщена автором на основе [1, 2, 7, 10].

Отдельный вопрос – почему ошибочный ответ часто выглядит уверенным. Языковая модель генерирует не психологическую «уверенность», а последовательность токенов, выбранных согласно внутреннему распределению вероятностей и настройкам генерации. Человек, однако, интерпретирует гладкую формулировку, отсутствие оговорок и последовательную структуру как сигнал компетентности. Таким образом, «уверенность» в пользовательском восприятии возникает из стиля ответа, а не из доказательства истинности. Это различие принципиально: модель может быть языково убедительной там, где фактическая основа слабая.

Измерение галлюцинаций также остаётся сложной исследовательской задачей. Для закрытых вопросов с эталонным ответом проверить факт сравнительно просто, но длинный открытый текст может содержать десятки утверждений, часть из которых требует интерпретации, актуальных источников или экспертного знания. Обзор Wang и соавторов подчёркивает трудности автоматической оценки фактичности открытой генерации и необходимость сочетать разные подходы [7]. Поэтому единая универсальная метрика «процента галлюцинаций» без описания набора вопросов и процедуры проверки малоинформативна.

Одним из направлений обнаружения ошибок является сравнение нескольких ответов одной и той же модели. В SelfCheckGPT используется идея, что при наличии устойчивого знания повторные генерации чаще согласуются между собой, тогда как вымышленные утверждения сильнее меняются от выборки к выборке [5]. Метод не требует доступа к внутренним вероятностям закрытой модели, что делает его практичным для внешнего контроля. Однако согласованность сама по себе не доказывает истинность: модель может стабильно повторять один и тот же ошибочный шаблон.

Другой подход связан с оценкой неопределённости. Farquhar и соавторы предложили использовать семантическую энтропию: анализировать не буквальное разнообразие формулировок, а различия в смысловых ответах. Метод позволяет выявлять случаи, где модель генерирует несколько несовместимых по смыслу вариантов, что может служить признаком конфабуляции [6]. Такие методы особенно важны для автоматизированных систем, но и они не заменяют внешнюю проверку там, где цена ошибки высока.

Наиболее очевидная мера снижения риска для конечного пользователя – требовать проверяемые источники и затем открывать их. Простая просьба «дай ссылки» недостаточна, потому что модель способна придумать правдоподобную ссылку или библиографическое описание. Следует проверять существование публикации по DOI, названию и автору, сверять цитируемую мысль с оригиналом, а для нормативных актов – использовать официальные базы. Если ответ касается текущей информации, необходимо явно запрашивать дату актуальности. Для числовых расчётов полезно отдельно проверять исходные данные и арифметические шаги.

Для разработчиков систем важным инструментом является RAG с контролируемым набором документов. Такой подход эффективнее, когда источник заранее ограничен надёжной базой, у каждого фрагмента есть происхождение, а система обязана показывать ссылки на использованные документы. Желательно отделять этап поиска от этапа формулирования ответа и предусматривать режим отказа: если подтверждающего фрагмента нет, система должна сообщить об отсутствии основания, а не дополнять ответ догадкой. В критических областях необходимо логировать источники и версию базы знаний.

Не менее важна калибровка поведения модели. Если система не знает факта, правильным результатом может быть не конкретный ответ, а явное обозначение неопределённости. Подход Kalai и соавторов предполагает, что распространённые критерии оценки следует перестроить так, чтобы уверенные ошибки штрафовались сильнее, чем корректный отказ от ответа [9]. Практически это означает необходимость интерфейсных и организационных механизмов: меток уровня уверенности, запроса уточнений, предупреждений о непроверенных сведениях и возможности передать сложный случай человеку.

В образовательной среде задача состоит не в полном запрете генеративных моделей, а в формировании проверяемого сценария их использования. UNESCO подчёркивает необходимость человеко-центричного подхода, при котором технологии поддерживают обучение, но не подменяют человеческое суждение [8]. Студенту полезно давать задания, где ответ модели необходимо проверить по первоисточникам, найти потенциальные ошибки, объяснить причины расхождения и исправить текст. Такой формат развивает одновременно предметное знание, цифровую грамотность и критическое мышление.

Для научной работы особого внимания требуют библиографические ссылки. Модель может корректно воспроизводить общую идею научного направления, но ошибиться в названии статьи, годе, составе авторов или DOI. Поэтому библиографию, созданную генеративной системой, нельзя переносить в рукопись без проверки. Надёжной процедурой является поиск каждого источника в каталоге издателя, Crossref, Google Scholar, Scopus, Web of Science или другой доступной академической базе, а затем сравнение с реально прочитанным текстом. Сам факт существования статьи ещё не подтверждает, что она содержит приписываемое ей утверждение.

В программировании галлюцинации проявляются иначе: модель может предложить

синтаксически убедительный код с несуществующим методом, перепутанным параметром или устаревшим API. Здесь проверка должна включать документацию соответствующей версии библиотеки, компиляцию или запуск тестов, анализ граничных случаев и контроль безопасности. Особенно рискованно принимать сгенерированный код без ревью, если он работает с аутентификацией, файлами, базами данных или пользовательскими вводами. Генеративный помощник способен ускорить разработку, но не отменяет тестирование.

Таким образом, галлюцинации следует рассматривать не как единичный «дефект», который будет устранён одним обновлением модели, а как системный риск вероятностной генерации. Современные системы становятся точнее, подключаются к поиску и лучше следуют инструкциям, однако задача надёжности остаётся открытой. Обзор литературы показывает, что наиболее устойчивый подход строится не вокруг одного метода, а вокруг многоуровневой защиты: качественные данные, внешние источники, детектирование неопределённости, самопроверка, оценка человеком и правила использования.

**Таблица 2 – Практические меры снижения риска галлюцинаций**

| Уровень      | Мера                                                   | Что она снижает                           | Ограничение                                    |
|--------------|--------------------------------------------------------|-------------------------------------------|------------------------------------------------|
| Пользователь | Уточнять контекст, дату, формат и критерии ответа      | Неоднозначность запроса                   | Не исправляет отсутствие знания                |
| Проверка     | Сверять факты и ссылки с первоисточниками              | Фактические и атрибутивные ошибки         | Требует времени и доступа к источникам         |
| Система      | Использовать RAG на проверенной базе                   | Ошибки из-за отсутствия актуальных данных | Поиск сам может выбрать нерелевантный источник |
| Модель       | Разрешать ответ «не знаю» и оценивать неопределённость | Уверенное угадывание                      | Калибровка не равна доказательству истинности  |
| Автоконтроль | Сравнивать несколько генераций, применять детекторы    | Нестабильные и противоречивые ответы      | Стабильная ошибка может остаться незамеченной  |
| Организация  | Human-in-the-loop для ответственных решений            | Риск автоматического принятия ошибки      | Нужна квалификация проверяющего                |

Примечание – меры обобщены на основе [2, 4-6, 8-10].

Важным следствием анализа является необходимость обучать пользователей не только «правильным промптам». Хорошо сформулированный запрос действительно уменьшает неоднозначность, но не превращает вероятностную модель в гарантированно достоверный справочник. Если пользователь просит модель «не галлюцинировать» или «отвечать только правду», это может изменить стиль ответа, однако само по себе не создаёт механизма проверки. Надёжность повышается тогда, когда инструкция связана с процедурой: использовать только предоставленные источники, цитировать конкретный фрагмент, явно указывать неизвестные данные, не дополнять отсутствующую информацию и отделять факт от предположения.

С точки зрения организации учебного процесса разумно вводить простое правило: любое утверждение, которое студент использует как факт в отчёте, статье или дипломной работе, должно иметь независимый проверяемый источник. Генеративный чат не должен выступать единственным доказательством. В программировании аналогичное правило можно сформулировать как «сгенерированный код не считается рабочим до запуска тестов», а в исследовательской работе – «сгенерированная ссылка не считается источником до открытия оригинальной публикации». Эти процедуры формируют культуру верификации и

уменьшают зависимость от внешне убедительной формы.

**Заключение.** Галлюцинации генеративного искусственного интеллекта представляют собой системную проблему фактической надёжности. Анализ научных публикаций показывает, что они возникают вследствие совокупности факторов: статистической природы предсказания следующего токена, качества и неполноты данных, ограниченных знаний модели, неоднозначности запроса, ошибок внешнего поиска и таких схем обучения и оценки, при которых попытка угадать ответ может быть выгоднее признания неопределённости. Поэтому уверенный стиль ответа не должен рассматриваться как показатель истинности.

Наиболее важный практический вывод заключается в необходимости многоуровневой верификации. Для пользователя это означает уточнение запроса, проверку ключевых фактов и источников, особое внимание к датам, числам, цитатам и библиографическим данным. Для разработчиков – использование retrieval-augmented generation, механизмов отказа при отсутствии подтверждения, оценки неопределённости, журналирования источников и автоматических детекторов. Для образовательных организаций – формирование правил ответственного применения и навыков критической проверки.

Полностью исключить ошибки генеративных моделей в открытых задачах на современном этапе нельзя, однако можно существенно снизить вероятность того, что ошибка будет незаметно принята за факт. Перспективное развитие надёжного искусственного интеллекта связано не только с увеличением размера моделей, но и с улучшением данных, методов оценки, калибровки неопределённости, интеграции проверяемых источников и человеческого контроля. В образовательной практике ключевым навыком становится не умение получить быстрый ответ от модели, а способность доказать, почему этому ответу можно доверять.

#### 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. – Article 248. – P. 1-38. DOI: 10.1145/3571730.
2. Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi E., Hall P., Roberts K. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. – Gaithersburg: National Institute of Standards and Technology, 2024. DOI: 10.6028/NIST.AI.600-1.
3. Lin, S., Hilton, J., Evans, O. TruthfulQA: Measuring How Models Mimic Human Falsehoods // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. – Dublin, 2022. – P. 3214–3252. DOI: 10.18653/v1/2022.acl-long.229.
4. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal N., Küttler H., Lewis M., Yih W.-t., Rocktäschel T., Riedel S., Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems. – 2020. – Vol. 33.
5. Manakul, P., Liusie, A., Gales, M.J.F. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. – Singapore, 2023. – P. 9004-9017. DOI: 10.18653/v1/2023.emnlp-main.557.
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. DOI: 10.1038/s41586-024-07421-0.
7. Wang, Y., Wang, M., Manzoor, M.A., Liu F., Georgiev, G.N., Das R.J., Nakov P. Factuality of Large Language Models: A Survey // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. – Miami, 2024. – P. 19519-19529. DOI: 10.18653/v1/2024.emnlp-main.1088.
8. Miao, F., Holmes, W. Guidance for Generative AI in Education and Research. – Paris: UNESCO, 2023. – 44 p. ISBN 978-92-3-100612-8.
9. Kalai, A.T., Nachum, O., Vempala, S.S., Zhang, E. Why Language Models Hallucinate. – arXiv:2509.04664, 2025. DOI: 10.48550/arXiv.2509.04664.
10. 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. – Article 42. – P. 1-55. DOI: 10.1145/3703155.