# 1 Использование генеративного ИИ в сфере анализа данных

**В этой главе**

- ✓ Знакомство с основными ограничениями генеративных моделей ИИ.
- ✓ Роль генеративного ИИ в анализе данных.
- ✓ С чего начать: применение LLM в аналитических задачах.

Пока вокруг генеративного ИИ постепенно стихает ажиотаж, а восторженные отзывы все чаще соседствуют с постами, где сквозят нотки разочарования, возникает закономерный вопрос: «Если большие языковые модели (LLM) — это не универсальное решение всех проблем, то в чем же они действительно полезны?»

Наш собственный опыт работы с этими удивительными инструментами дает вполне четкий ответ: они *отлично подходят* для улучшения и ускорения различных процессов. В книге мы покажем, как раскрыть потенциал, скрытый в структурах генеративных моделей, чтобы развить свои аналитические навыки и при этом не попасть в ловушки, связанные с рисками новой технологии.

В этой главе мы рассмотрим три ключевые особенности работы с генеративным ИИ.

Чтобы по-настоящему освоить инструмент, не нужно спешить. Мы начнем с основного: ограничений генеративного ИИ. Мы уже упоминали, как сильно могут расходиться ожидания и реальность при работе с LLM. Хорошее понимание

границ возможностей таких моделей поможет избежать ненужных разочарований и строить рабочие процессы осознанно.

Во второй части мы перейдем к встраиванию генеративного ИИ в процесс анализа данных. Здесь вы научитесь определять, когда и как уместно использовать языковые модели, а также как правильно формировать ожидания, особенно если речь идет об автоматизации. Генеративный ИИ может быть мощным союзником, но требует вдумчивого подхода и четкой постановки задач.

Наконец, в третьей части мы поговорим о способах доступа к генеративным моделям. В большинстве случаев достаточно использовать чат в браузере — этого вполне хватает для образовательных целей и несложных сценариев. Но если вы работаете с конфиденциальными или чувствительными данными, стоит заранее продумать архитектуру доступа и просчитать риски.

**ЗА КАДРОМ** Чтобы эффективно использовать генеративный ИИ, важно понимать не только процесс взаимодействия, но и то, откуда берутся ответы, которые вы получаете на свои запросы. Однако, учитывая архитектурную нейтральность книги, мы сознательно избегаем технических деталей конкретных моделей или платформ. Такой подход позволяет сосредоточиться на применении ИИ в аналитике, не привязываясь к быстро устаревающим технологиям. Тем не менее техническая сторона вопроса заслуживает отдельного внимания. Если вам интересно разобраться глубже, обратитесь к таким книгам, как *The Complete Obsolete Guide to Generative AI* Дэвида Клинтона, *Introduction to Generative AI*[1] Нумы Дхамани и Мэгги Энглер и *How GPT Works* Дрю Фарриса, Эдварда Раффа и Стеллы Бидерман. Эти издания помогут вам получить более полное представление о внутреннем устройстве генеративных моделей, включая GPT.

Главная цель этой главы — не предоставить энциклопедический обзор технологии, а сформировать у вас достаточно глубокое понимание, чтобы вы научились трезво оценивать возможности технологии и критически подходить к применению генеративного ИИ в аналитике.

## *1.1. Внутренние ограничения генеративных моделей ИИ*

В Средние века на краях морских карт можно было увидеть надпись: *Hic sunt dracones*, что в переводе с латыни означает «Здесь обитают драконы». В неизведанных областях рисовали чудовищ, сирен и кракенов, предостерегая мореплавателей. Позднее страшилки о мифических существах сменились более практичными предупреждениями — о рифах, отмелях и льдах.

---

[1] *Дхамани Н., Энглер М.* Генеративный искусственный интеллект. Как ИИ меняет нашу жизнь и работу. — М., 2025.

Мы хотим, чтобы вы воспринимали наши наставления именно в таком ключе: не как попытку отпугнуть, а как памятку. В любом новом начинании важно не только понимать, какую пользу может принести технология, но и знать, с какими рисками вы можете столкнуться.

Ниже перечислены ограничения, присущие любой генеративной системе ИИ. Какие-то из них, возможно, будут смягчены или устранены в будущем, но некоторые могут оставаться актуальными, и лучше о них знать, не так ли?

Итак, вот подводные камни, которые стоит учитывать.

- *Генеративный ИИ всегда дает ответ (даже если он неправильный)*. Как упрямый ребенок, переросший возраст «почемучки», или чрезмерно самоуверенный менеджер с синдромом самозванца, языковые модели практически никогда не признают, что чего-то не знают. Вместо этого они всегда отвечают — уверенно, связно, но не обязательно правильно.

  Прочитав книгу, вы научитесь распознавать такие случаи и корректировать поведение модели. Мы также расскажем о том, как формулировки запроса влияют на результат. Например, в главе 8 будет подробно объяснено, как небольшое изменение формулировки может привести к совершенно разным по качеству и смыслу ответам. Важно понимать: в отличие от классических поисковых систем, которые реагируют на набор ключевых слов, генеративные модели учитывают грамматику, интонацию и даже контекст недавнего диалога. Так взаимодействие с ними становится не только более гибким, но и непредсказуемым.

  **С ЧЕГО ТЫ ЭТО ВЗЯЛ?** Генеративный ИИ часто формулирует ответы так, чтобы они звучали убедительно — даже если они неверны. Но вы можете повысить надежность и практическую ценность ответов, если подключите модель к внешним источникам данных и будете требовать от нее ссылки на использованные материалы. Если вам интересен этот подход, рекомендуем ознакомиться со следующими книгами: *Generative AI in Action* Амита Бахри и *AI-Powered Search*[1] Трей Грейнджер, Дуга Тернбулла и Макса Ирвина.

- *Некоторые ответы могут быть полностью выдуманы*. Генеративный ИИ может с уверенностью выдать ответ, который выглядит правдоподобно, но на деле не опирается ни на факты, ни на обучающие данные. Модель «заполняет пробелы» — создает текст на основе знакомых ей закономерностей, даже если они основаны на неполных или искаженных данных. Мы подробно рассмотрим этот феномен в главе 8, когда будем говорить о так называемых галлюцинациях ИИ — ситуациях, когда модель «придумывает» ответ, который звучит правдоподобно, но не имеет под собой основы.

---

[1] *Грейнджер Т., Тернбулл Д., Ирвин М.* Поиск на основе искусственного интеллекта. — М., 2025.

- *Врожденное подхалимство*. Чем больше языковая модель, тем выше риск, что она будет стремиться угодить пользователю, даже в ущерб точности и правде. Если вы будете оспаривать ответ модели, она, скорее всего, извинится и примет противоположную точку зрения, даже если ее исходный ответ был корректен. Более того, модель может выдумывать числа, ссылки и аргументы, лишь бы подтвердить позицию собеседника. Такая склонность к согласию делает LLM ненадежным помощником в вопросах, где критичны факты, доказательства и устойчивость позиции!
- *Неточная или устаревшая информация*. Знания генеративной модели ограничены данными, на которых она была обучена. Если модель обучалась на контенте, актуальном на определенный момент времени, она может выдавать устаревшие или неточные ответы, особенно в быстро развивающихся областях. В книге вы встретите примеры, где модель предлагает использовать устаревшие версии API или синтаксис, не соответствующий текущим стандартам. Тем не менее это ограничение не столь критично, как может показаться. Во-первых, большинство базовых концепций в аналитике и программировании остаются стабильными, и именно с них начинается путь в профессию. Во-вторых, многие современные модели имеют доступ к Интернету — при определенных настройках и запросах они могут получать актуальную информацию. Однако важно помнить: доступ к Интернету есть не у всех моделей — даже при его наличии модель не всегда запрашивает обновления автоматически, а сочетание устаревших знаний и попытки частично их обновить могут приводить к смешанным или искаженным результатам. Поэтому всегда полезно проверять ключевые детали — особенно если речь идет о документации, версиях библиотек или новых подходах.
- *Ограничения на объем ввода и вывода*. При работе с генеративным ИИ важно учитывать максимальный объем текста, который модель способна обработать за один раз. Это касается как входных данных (ваших запросов), так и выходных (ответов). У разных моделей и реализаций эти пределы различаются. Модель обрабатывает текст в виде токенов — это могут быть целые слова, части слов или даже знаки препинания, в зависимости от применяемого алгоритма токенизации. В среднем можно ориентироваться на соотношение: 1 токен ≈ 0,75 слова (или 4 токена на 3 слова). На момент написания книги размеры так называемого контекстного окна варьировались от нескольких тысяч до миллионов токенов — и индустрия активно движется в сторону увеличения этих объемов. Однако даже самые продвинутые модели сегодня имеют ограничения на количество токенов, охватывающее и запрос, и ответ. Если вы превысите лимит, модель не предупредит вас — она просто «забудет» часть входных данных. Это может привести к ситуациям, когда ответ не учитывает предыдущую переписку, теряет контекст или противоречит ранним сообщениям. В разделе 1.3 мы расскажем, как оценить количество токенов, используемых в запросе или ответе, и предложим стратегии для минимизации потерь. Одна из таких стратегий — регулярно резюмировать ход диалога, чтобы важная информация оставалась в пределах доступного контекста.

- *Многословность*. При работе с генеративным ИИ быстро становится заметно: модели склонны к избыточной подробности и повторяющимся формулировкам. Иногда они «неуклонно углубляются в обширные ландшафты богатого переплетения хитросплетений» — даже когда этого вовсе не требуется. Это связано с обучающими данными, где более длинные и формальные ответы встречались чаще и воспринимались как «норма».

  **СЛОВО НЕ ВОРОБЕЙ** Ограничения на длину запроса и ответа, в сочетании со склонностью модели к многословности, могут приводить к усеченным или неполным ответам. При проектировании взаимодействия с генеративным ИИ важно следить за тем, чтобы общая длина диалога не превышала лимит токенов, иначе часть информации может быть потеряна — и без предупреждения.

- *Предвзятость и ненадлежащее содержание*. Несмотря на усилия разработчиков, генеративные ИИ все еще могут выдавать необъективную информацию или неуместный (в том числе оскорбительный) контент. Это может происходить из-за искажений в обучающих данных, скрытых формулировок в запросе или других факторов. Разработчики большинства современных моделей пытаются сбалансировать поведение искусственного интеллекта и минимизировать искажения в ответах. Один из примеров такой работы подробно описан в документе GPT-4 System Card (https://cdn.openai.com/papers/gpt-4-system-card.pdf).

Понимание всех этих ограничений критически важно при интеграции ИИ в рабочие процессы и приложения. Исследования и разработка в данной области продолжаются, и основной их целью остается повышение надежности, безопасности и полезности генеративных моделей.

## 1.2. Роль генеративного ИИ в аналитике данных

В тематических группах и на форумах, посвященных генеративному ИИ, можно встретить десятки вопросов вроде: «Где найти инструмент на основе GenAI, который выполняет [крайне специфическую задачу]?» Даже если такого решения пока нет, оно, вероятно, скоро появится. И это действительно хорошо. Data warehouses, lakes, lakehouses, meshes, fabrics[1] и другие современные архитектуры заменяют Excel-файлы, данные в письмах и заметки на салфетках (пусть и не во всех случаях). Дашборды и платформы самостоятельной бизнес-аналитики

[1] Data warehouse — централизованное хранилище структурированных данных, оптимизированное для аналитики. Data lake — хранилище сырых данных в их исходном формате. Lakehouse — гибридная архитектура, сочетающая возможности хранилища данных и озера данных. Data mesh — подход к распределенной архитектуре данных, при котором ответственность за наборы данных распределена по доменам. Data fabric — архитектурный подход, обеспечивающий интеграцию данных, управление ими и доступ к ним в масштабах организации.

(business intelligence, BI) постепенно вытесняют вручную собранные отчеты и презентации PowerPoint. Кстати, запросы о том, как использовать генеративный ИИ для создания, редактирования и улучшения презентаций, одни из самых частых. Тем не менее к любому инструменту с пометкой «на базе генеративного ИИ» стоит относиться критически. В среднем только 0,5 % данных, хранящихся в корпоративных хранилищах и озерах данных, действительно подвергаются анализу. Остальные 99,5 % просто накапливают издержки, особенно если сбор этой информации происходил без четкого плана ее применения. Сами BI-платформы тоже небезупречны: в них нередко встречается вредоносная аналитика — например, бессмысленная фрагментация и сортировка, оправдывающие заведомо неудачные решения.

Эффективность генеративного ИИ в аналитике зависит от самого аналитика — от его умения видеть возможности и осознавать ограничения. Как и любой другой инструмент, ИИ не решит все за вас.

До этого момента мы намеренно фокусировались на рисках, чтобы показать: завышенные ожидания — главный враг практического применения генеративного ИИ. Мы даже заключили пари, когда на рынке появится первая зубная щетка с маркировкой «на базе GenAI» (да, мы действительно делали ставки). Но теперь давайте отбросим скепсис и сделаем шаг навстречу будущему — тому самому, где аналитика данных и генеративный ИИ работают вместе. Вдумчиво, разумно и с пользой.

### *1.2.1. Как генеративный ИИ можно использовать в аналитике данных*

Годы работы с данными научили нас: ценность аналитики не зависит от сложности используемых технологий. Мы видели, как одни компании экономили миллионы долларов, просто перераспределив затраты по бизнес-процессам, а не по подразделениям. Другие же, напротив, несли значительные убытки из-за слишком сложных аналитических решений, которые, несмотря на использование десятков инструментов и привлечение больших команд специалистов, слабо отвечали реальным потребностям клиентов. *Суть аналитики — не в красивых диаграммах, а в том, чтобы обеспечивать принятие взвешенных бизнес-решений на основе глубокого анализа релевантных данных.* Ваш успех будет зависеть от множества факторов, и доступный инструментарий — это лишь один из элементов.

Разные бизнес-задачи требуют разных аналитических конвейеров. Если вы работаете в сфере розничной торговли, то, скорее всего, вас интересует поведение клиентов. Ваш аналитический процесс, вероятно, начинается с очистки данных транзакций, отзывов и поведения на сайте. Затем вы применяете сегментацию, анализ схожих товаров и прогнозирование продаж. В области

здравоохранения данные поступают из электронных карт пациентов, снимков и носимых устройств. Здесь аналитика сосредоточена на диагностике, оптимизации лечения и прогнозировании исходов. В производстве источниками информации служат IoT-датчики, системы контроля качества и логистика. Здесь в центре внимания предиктивное обслуживание, обнаружение аномалий и управление спросом. Но независимо от отрасли основной процесс остается неизменным: собрать и очистить данные, обработать их с помощью подходящих алгоритмов и донести результаты до тех, кто принимает решения.

Разумеется, конкретные шаги будут зависеть от вашей отрасли, доступных источников, методов анализа и требуемого формата результата. Каждая из этих тем заслуживает отдельной книги — или целой серии — о том, как строить эффективную аналитику с учетом времени, бюджета и технических возможностей.

В этой книге мы не пытаемся охватить все возможные сценарии. Вместо этого мы предлагаем нечто куда более ценное — структурированный подход к использованию огромного объема знаний, заключенного в генеративном ИИ (от Википедии и научных публикаций до наборов данных вроде The Pile, https://pile.eleuther.ai/). Вы научитесь использовать этот инструмент для проектирования аналитического конвейера, который будет соответствовать вашей задаче, вашему контексту и вашей цели.

**НЕДОСТАЮЩЕЕ ЗВЕНО** Пределы генеративного ИИ все еще активно исследуются. Но уже сегодня очевидно: эти модели способны последовательно и по существу отвечать на вопросы по самым разным темам. Они умеют вникать в детали, обобщать, объяснять сложные идеи и находить связи между, казалось бы, далекими концепциями. Эти способности можно использовать, чтобы выйти за рамки привычного мышления, расширить кругозор и взглянуть на проблему под новым углом. Вам больше не нужно продираться через десятки случайных статей в поисках вдохновения или «знаков свыше». Просто задайте вопрос. Да, ответ может быть неидеален. Но даже в такой обратной связи часто встречаются идеи или концепции, о которых вы раньше не задумывались. Используйте генеративный ИИ — и вы удивитесь, насколько шире может стать ваше поле зрения.

В начале книги мы задали ключевой вопрос, который стоит держать в голове каждый раз, сталкиваясь с новой аналитической задачей: *с чего начать?* Поиск подходящих исходных данных — уже неплохой старт, особенно если он сопровождается вдумчивым анализом: а какие данные действительно важны для ответа? Допустим, вы работаете в медицинском учреждении и получаете запрос: «Каково среднее время ожидания пациентов по вторникам?» Или трудитесь в сфере розничной торговли, и вас озадачивают: «Как наши клиенты используют карты лояльности?» Не позволяйте внешней простоте этих формулировок ввести себя в заблуждение — обе задачи могут оказаться гораздо сложнее, чем кажется на первый взгляд.

На рис. 1.1 представлена универсальная схема аналитического процесса — от формулировки вопроса до принятия решений на основе полученных выводов. Эта структура предназначена для того, чтобы максимально использовать как экспертные знания аналитика, так и возможности генеративного ИИ, при этом избегая типичных ловушек и лишней сложности. Поток подходит для любой аналитической задачи и любого технологического стека, с которым вы работаете. Все практические примеры в книге будут опираться на эту схему как на каркас.

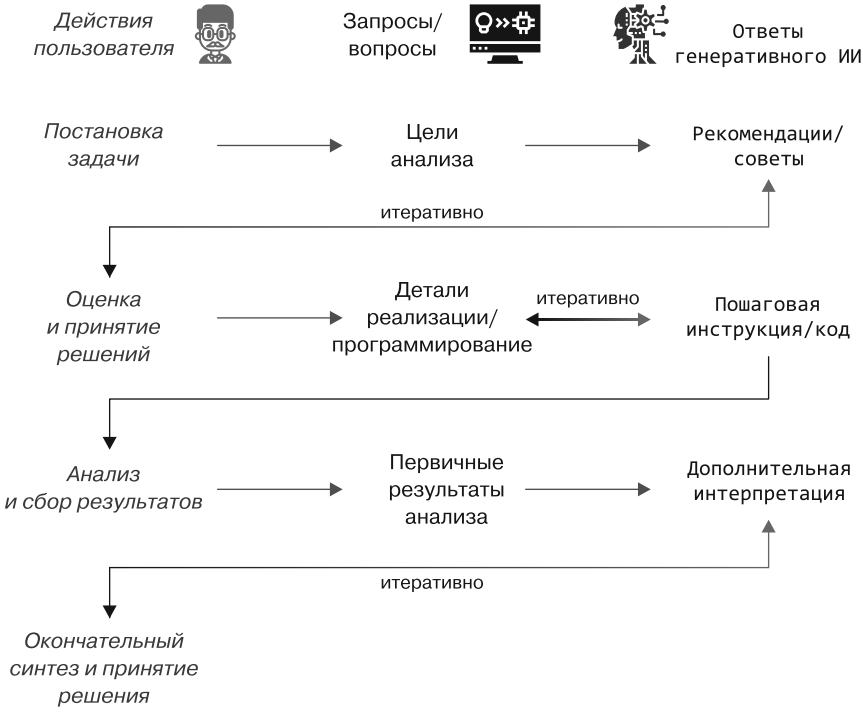


**Рис. 1.1.** Рекомендуемый поток анализа данных, поддерживаемый генеративным ИИ

Первым шагом всегда должна быть постановка задачи.

Вернемся к нашему примеру из сферы здравоохранения. Нам задали вопрос: «Каково среднее время ожидания пациентов по вторникам?» На первый взгляд, все просто. Но в действительности предмет обсуждения может быть лишь симптомом более глубокой проблемы, связанной с организацией приема, нехваткой персонала или сезонной нагрузкой. На практике вопрос, который вам задают, не всегда напрямую превращается в четко сформулированную аналитическую задачу. Здесь важно спросить себя: «Какую задачу мы на самом деле пытаемся решить?» Хотя лицо, принимающее решение, определяет объем и цель анализа,

хорошо продуманный встречный вопрос может помочь вам сформировать более точную и полезную отправную точку для аналитики и повысить общую ценность вашей работы.

А теперь — самое интересное. Даже если вы не знакомы с предметной областью, генеративный ИИ может помочь вам поместить исходный вопрос в нужный бизнес-контекст. Это особенно полезно, когда у вас нет возможности быстро получить разъяснения от коллег или заказчика. Вместо догадок стоит попробовать прямо спросить ИИ. Давайте посмотрим, как справятся с этой задачей несколько генеративных моделей, если задать им наш «вторничный» вопрос.

МА Я работаю в медицинской организации. Мне задали вопрос: «Каково среднее время ожидания пациентов по вторникам?» Какие возможные причины могут стоять за таким вопросом?

Ответы генеративных ИИ оказались излишне длинными, но все они (ChatGPT версии 3.5 и 4, Google Gemini, Gemini Pro и Meta Llama 2 [13B]) представили списки типов проектов, где такой анализ может быть уместен. Наиболее часто упоминались: планирование и бюджетирование (например, распределение ресурсов и оптимизация графика персонала), качество ухода и удовлетворенность пациентов, операционная эффективность и обучение и управление персоналом.

В зависимости от того, насколько вы знакомы с предметной областью, вы можете возвращаться к заказчику с более конкретными вопросами. Это не пустая формальность, а способ уточнить контекст и избежать лишней работы. Например, спросив: «Связано ли это с нашей новой программой повышения удовлетворенности пациентов?», вы можете получить прямой и ценный ответ: «Да, по вторникам особенно много жалоб, и мы хотим это изменить». Как только настоящая проблема становится ясна — например, жалобы и длительное ожидание, — вы можете переформулировать аналитическую задачу: «Как распределяется время ожидания и как оно связано с уровнем удовлетворенности пациентов?» Такой вопрос сочетает конкретику с бизнес-значимостью и помогает выявить действия, которые действительно повлияют на правила оказания помощи и восприятие пациентами качества обслуживания. При этом важно помнить: уровень детализации в общении с генеративным ИИ должен соответствовать требованиям конфиденциальности. В локальных безопасных средах допустимо использовать более свободные формулировки, чем на общедоступных платформах. Вопросы рисков и приватности мы рассмотрим в главе 8.

После уточнения задачи вы можете обратиться к генеративному ИИ с вопросом.

МА Как построить анализ, чтобы ответить на вопрос: «Как распределяется время ожидания и как оно связано с удовлетворенностью пациентов?»