Just AI

[13 ограничений]

генеративных моделей и пути их преодоления

Апрель 2023

Доступ к персональным данным, требования к ресурсам и другие сложности применения GenAI в бизнесе

Генеративные модели громко заявили о себе всего пару лет назад, а сегодня — это уже неотъемлемая часть современного технологического ландшафта. AI привносит в наш мир инновации, творчество и дарит удивительные возможности. Написать научную статью, создать видео по текстовому описанию, сделать саммари многостраничного отчета или придумать код программы — это лишь часть его возможностей. Потенциал его огромен. По оценке Mckinsey, генеративный AI может принести ежегодную экономическую выгоду в размере 2,6–4,4 трлн долл США в 63 вариантах использования.

Но вместе с тем у таких моделей есть ряд ограничений, которые нужно учитывать. Особенно бизнесу. Здесь мы рассмотрим как технические, так и этические аспекты использования AI. Цель этого гайда — рассказать об ограничениях, которые есть у генеративных моделей, а также о том, как преодолевать эти риски при их обучении, внедрении и использовании.

Генеративный AI — это тип искусственного интеллекта, обученного на больших объемах данных, который может по запросу человека создавать уникальный контент, включая тексты, изображения, видео, музыку и код. Но при этом генеративный AI не может заменить человека полностью. Ему не хватает креативности, он плохо понимает юмор и иронию, не всегда улавливает контекст и испытывает трудности в анализе.

Прямые ограничения для бизнеса

[1] Большинство моделей доступны из облаков их разработчиков

Проблема:

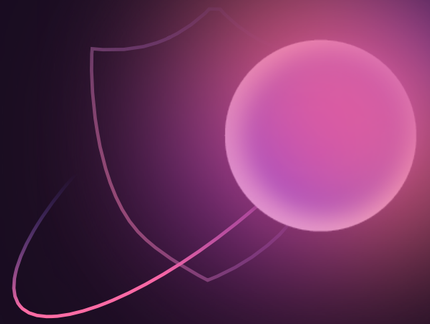
Это снижает применимость их в корпорациях, работающих с чувствительными данными. Например, финтехе. Даже пилотные проекты банки вынуждены запускать в контуре, чтобы обезопасить себя от потери персональных данных. Это сопряжено с высокими расходами и длинным периодом реализации. Облачные решения проще и быстрее, но не всегда обеспечивают необходимый уровень безопасности.

Решение:

Компании должны четко отслеживать, что их персональная информация не будет встроена в генеративную модель. Необходимо тренировать модели, доступные для развертывания в контуре, либо использовать шлюз для выявления и замены/анонимизации части данных.

Шлюз-платформа Jay Guard

Решает проблему фильтрации и маскирования данных при работе с LLM. Вся чувствительная информация заменяется на вымышленную, при этом сохраняется семантическая связность и контекст сессий.



Преимущества:

- [Возможность использовать самые мощные генеративные модели (GPT-4 Turbo, GigaChat, YaGPT 2 и подобные)
- [Экономия ресурсов компании и отсутствие затрат на хостинг и эксплуатацию собственных LLM
- [Сокрытие чувствительных данных без потери контекста при работе с LLM

Прямые ограничения для бизнеса

[2] Персонализация = доступ к личным данным

Проблема:

ИИ-модели активно используют в создании персонализированного контента, например, в клиентском обслуживании, но это сопряжено с доступом к персональным данным.

Решение:

Бизнес-сообщество берет на себя инициативу и разрабатывает собственные политики в области генеративного AI, чтобы защитить себя и своих клиентов. Кроме того, многие компании уже сейчас при работе с ИИ-моделями используют методы шифрования данных, анонимизации, используют цифровые двойники.

Прямые ограничения для бизнеса

[3] Быстродействие

Проблема:

Они не подходят для решения real-time задач, например, общение по телефону клиента и сотрудника, которые говорят на разных языках.

Решение:

Необходимо обучать более легкие и быстрые модели, оптимизировать их производительность на базе своих мощностей или с использованием сторонних платформ.

Caila –

платформа агреграции, хостинга, эксплуатации и дообучения ML-сервисов и нейромоделей.

Уже содержит каталог готовых генеративных моделей с возможностью их использования через API и SDK. Но можно загружать свои модели, адаптировать их или создавать новые.

Caila

Преимущества:

- [Платформа соответствует ведущим стандартам по информационной безопасности и защите персональных данных
- [В Caila можно назначать и ограничивать права для пользователей и групп с помощью ролевой модели доступа
- [Решение может быть развернуто как on-premise, так и в приватном облаке для большего контроля безопасности

Прямые ограничения для бизнеса

[4] Ограниченное окно контекста

Проблема:

Контекстное окно — это определенный диапазон, в пределах которого модель работает, ее память. Его размер определяют токенами. Например, у GPT-4 Turbo размер контекстного окна равен 128 000 токенов, что равно примерно 300 страницам книги. А GPT-3.5 имеет только 4096 токенов. Для генеративных моделей каждый токен представляет собой часть текста — слова, их фрагменты или даже знаки препинания. Модели присваивают каждому из них свой уникальный идентификатор и используют их для числового кодирования. Как это влияет на работоспособность моделей? Чем больше у них размер контекстного окна, тем дольше и более связно они могут общаться с пользователем или тем больше данных они могут обрабатывать.

Из-за ограниченного окна контекста модель может терять доступ к ранним частям диалога. Именно поэтому GenAI пока недоступны длительные разговоры, а при решении многоступенчатых задач модели могут совершать ошибки или выдавать неточности.

Решение:

Кажется, что самое простое решение — увеличить окно контекста. Но большие модели требуют колоссальных ресурсов как для обучения, так и работы. Кроме того, это не всегда приводит к повышению производительности и точности модели. Иногда важную информацию можно сгенерировать с меньшим количеством токенов, но тут есть нюансы.

Для небольших моделей помогает более тщательная проработка промтов или использование метода скользящего окна — это когда контекст диалога перемещается вперед, сохраняя только самые последние токены, которые важны для продолжения генерации.

Большим моделям помогает P-Tuning. Здесь внимание модели фокусируется на наиболее важных частях большого контекста, которые являются ключевыми в решении конкретной задачи.

Прямые ограничения для бизнеса

[5] Требования к ресурсам

Проблема:

Для эффективного обучения генеративным моделям требуются значительные вычислительные мощности и много времени. Это может сделать их недоступными для небольших проектов или компаний с ограниченными ресурсами.

Пример:

В России для обучения GigaChat, по данным SberDevices, потребовалось столько ресурсов, что такого количества бы хватило, чтобы четыре месяца обеспечивать электроэнергией стадион «Лужники». А один центр обработки данных в среднем потребляет энергию, эквивалентную отоплению 50 000 домов в год.

Решение:

Развитие небольших и специализированных моделей, которым не требуются для обучения и работы такие мощности и ресурсы. Другой путь — использование агрегаторов моделей от сторонних разработчиков. Они через единый интерфейс помогают пользователям получить доступ сразу к нескольким нейросетям. Хороший пример — Jay Copilot.

Jay Copilot —

агрегатор LLM-моделей с 20+ приложениями для повышения эффективности сотрудников.

Функционал популярных нейросетей в едином окне в виде удобных приложений. Помогает персоналу получать ответы на вопросы, создавать тексты любой сложности, генерировать иллюстрации или делать расшифровку встреч

Преимущества:

- [Под капотом решения «спрятаны» популярные нейросети: ChatGPT, GigaChat, YandexGPT 2, JustGPT
- [Уже содержит десятки готовых шаблонов запросов (промптов): от создания SEO-текстов до анализатора сайтов
- [Высокое качество генерируемых изображений обеспечивают продвинутые нейросети – Stable Diffusion и Dalle-3

Прямые ограничения для бизнеса

[6] Контроль качества генерируемых ответов

Проблема:

Сложно проверить генерируемые данные на достоверность. Большинство моделей допускают ошибки, неточности и «галлюцинации». Например, модель может создать изображение, которое на первый взгляд выглядит реалистичным, но при ближайшем рассмотрении имеет небольшие аномалии.

Кроме того, на большинство запросов нет единственного правильного ответа. И они могут быть очень субъективны. Мы не имеем «золотого стандарта», которому должен соответствовать AI. Модель может сгенерировать что угодно.

Решение:

Пока готового решения тут нет. Но попытки создать систему оценки работы AI или хотя бы научиться определять, где генеративные модели не достоверны — не раз предпринимались. Например, еще в 2021 году команда OpenAI и Оксфордского университета разработали тест оценки под названием TruthfulQA. Он помогает определить, когда ИИ-боты имитируют человеческую ложь.

Прямые ограничения для бизнеса

[7] Персонал

Проблема:

Для компаний внедрение моделей часто сопряжено с проблемой отсутствия ИИ-специалистов, которые могут просчитать все риски и грамотно реализовать проект. В последнем отчете Ai Accelerator Institute говорится, что 11,8% респондентов не используют инструменты генеративного AI по следующим причинам: 46,2% - недостаток знаний, 30,8% - опасения по поводу конфиденциальности данных, 15,4% - опасения по поводу интеллектуальной собственности.

Решение:

Может потребоваться время, чтобы полностью понять возможности генеративного AI, как его применять для решения конкретных отраслевых или бизнес-задач. Еще больше может занять обучение сотрудников, формирование собственных команд по внедрению ИИ-моделей. Но это решаемые вопросы. Начать можно с пополнения собственной базы знаний блоком про работу LLM. Не лишним будет и интеграция модуля GenAI в поисково-справочную систему компании.

Jay Knowledge Hub —

модуль управления корпоративными данными с помощью генеративного AI и RAG технологий.

Формирует развернутый ответ на вопрос сотрудника или клиента, используя данные из всех корпоративных баз и документов.

Преимущества:

- Принцип одного окна — вся информация доступна через единого ИИ-бота, не надо дергать коллег или искать по разным папкам
- Почти мгновенная обработка запроса — пользователь получает полноценный ответ на свой вопрос всего за 10-20 секунд
- Возможность обеспечить быстрый доступ к обновленным данным — достаточно дообучить Jay KnowledgeHub на актуальной информации

Косвенные ограничения для бизнеса

[8] Доступ моделей к материалам с авторским правом

Проблема:

Большинство моделей учится на открытых данных из интернета, часто невольно нарушая авторское право или выдавая откровенный плагиат. Это уже вызвало правовые прецеденты.

Пример:

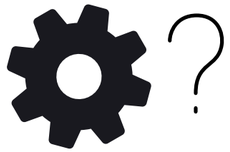
Издательство The New York Times подало в суд на OpenAI и Microsoft за нарушение авторских прав при обучении ИИ-моделей. Журналистов волнует, что материалы ИИ-бота начинают конкурировать с реальными статьями. Причем часть сгенерированного контента изобилует неточностями. Налицо репутационные и финансовые риски.

Решение:

Формирование правового поля в части доступа моделей к авторским материалам. Законы об авторском праве, касающиеся генеративного AI, пока во многом являются серой зоной, но процесс уже запущен. OpenAI недавно представила компенсационную программу Copyright Shield, которая покрывает судебные издержки по искам о нарушении авторских прав для определенных уровней клиентов, а не удаляет защищенные авторским правом материалы из набора обучающих данных ChatGPT.

Косвенные ограничения для бизнеса

[9] Отсутствие интерпретируемости



Проблема:

В отношении большинства генеративных моделей сложно понять, как они принимают решения или почему они выдают конкретные результаты, что может быть проблемой в таких важных областях, как здравоохранение или правосудие. Отсутствие четкого понимания процесса принятия решений мешает доверию и принятию ИИ-моделей.

Судья, вынося решение, руководствуется юридическим кодексом. Но на окончательный приговор влияет и моральный кодекс. Как ИИ, созданный в лаборатории, может это совместить? И нужна ли мораль генеративным моделям? Здесь пока нет правильного ответа. И не появится, пока работа моделей остается для нас черным ящиком.

Решение:

Исследователи активно изучают как можно улучшить интерпретируемость генеративных моделей. Для этого используют методы LRP(Layer-wise Relevance Propagation) и Grad-CAM. Они помогают разработчикам визуализировать области входных данных — изображения, видео или текст. Именно они наиболее существенно влияют на решения, которые принимают модели.

Косвенные ограничения для бизнеса

[10] Зависимость данных

Проблема:

Качество сгенерированного контента во многом зависит от обучающих данных и используемых шаблонов. Если они предвзяты, имеют сомнительное происхождение или устарели, результаты модели тоже будут некорректными. AI уже не раз проявлял расовое и гендерное неравенство.

Пример:

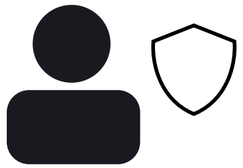
Еще в 2018 году Amazon был вынужден отказаться от ИИ-инструмента подбора персонала после того, как выяснилось, что он непропорционально отдает предпочтение мужчинам. Так как он был обучен на данных о предыдущих сотрудниках, в основном мужчинах, то начал понижать рейтинг любых резюме, содержащих слово «женские».

Решение:

Тщательная обработка наборов данных для обучения моделей. Это позволило был уменьшить предвзятость AI, снизить количество ошибок и неточностей. Кроме того, разработчики должны соблюдать прозрачность в отношении источников данных для обучения своих ИИ-моделей, что позволит завоевать доверие среди пользователей и компаний, которые используют их продукты.

Косвенные ограничения для бизнеса

[11] Этика



Проблема:

Часть моделей способны генерировать реалистичные данные, которые могут быть использованы злоумышленниками. Например, для создания дипфейков или контрафактного контента. Защита генеративных моделей от использования злоумышленниками требует постоянных исследований и разработки надежных мер безопасности.

Даже в бизнес-коммуникациях AI может сослужить плохую службу. Ситуация — модель создает электронное письмо от лица компании, где непреднамеренно содержатся оскорбительные выражения или некорректные указания для сотрудников.

Решение:

Необходимость создания нормативно-правовой базы для AI — тема для дискуссий во всем мире. Многие страны предпринимают попытки принять законы об искусственном интеллекте. Свой закон о регулировании AI готовят ЕС, Китай, Россия и другие страны. А Италия даже на короткое время запрещала ChatGPT, пока OpenAI не улучшила свои стандарты защиты.

Косвенные ограничения для бизнеса

[12] Невозможность заменить человека полностью

Проблема:

Какими бы совершенными ни были модели, они не обладают такими качествами, как креативность, эмоциональный интеллект или воображение. Генеративные модели создают новый контент на основе наборов данных. Пойти дальше — придумать принципиально новое и оригинальное им сложно.

Плюс генеративный AI по-прежнему испытывает трудности в тех ситуациях, где надо проявить сочувствие, сострадание или уловить настроение. К примеру, модели могут определить, что человек плачет, но вот от расстройства или радости — нет. Даже специализированные ИИ-боты пока с трудом формируют близкое взаимодействие с пользователями.

Решение:

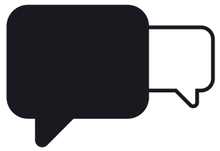
Пока люди и не ждут, что AI станет новым Пикассо или заменит профессионального психолога. Но уже активно используют генеративные модели, ища у них вдохновение или привлекая их как ассистента.

Пример:

В Гонконге часть школ использует продукт Find Solutions AI. Нейросеть измеряет микродвижения мышц на лицах учеников и определяет их настроение. Учителям эта информация помогает контролировать эмоциональное состояние школьников, их сосредоточенность. Если ученик теряет интерес к предмету — пора принимать меры. Но пока они не 100% эффективны. AI сводит выражения лица и интонации к определенной эмоции, не принимая во внимание социальный и культурный контекст человека и ситуации.

Косвенные ограничения для бизнеса

[13] Понимание контекста



Проблема:

Разработчики активно работают над способностью больших языковых моделей улавливать и генерировать ответы на основе контекста. Но пока не все тут гладко. Некоторые LLM не могут работать с документами и данными большого объема, а также «забывают» контекст длинных диалогов.

Решение:

Генеративные модели прошли большой путь, но для лучшего понимания людей им еще придется много учиться. Контекстное обучение поступательно поможет моделям LLM генерировать более релевантные и точные ответы, учитывая контекст, в котором задается вопрос.

Заключение

Сегодня генеративные модели стали для людей большими помощниками. Но вместе с тем они содержат в себе много рисков и ограничений. Да, они в состоянии решить сложные языковые задачи, но пока еще далеки от от рассуждений на человеческом уровне. Внедрение генеративного AI — это масштабная работа по управлению изменениями. Компании должны грамотно использовать новые технологии, а также постоянно адаптироваться перед нововведениями — GenAI постоянно расширяет свои границы.

У генеративного AI сегодня существует множество возможностей и перспектив для развития. Однако в этом гайде мы видим, что существует достаточно ограничений, сдерживающих его внедрение. Компании еще не готовы целиком полагаться на него. Использовать или не использовать GenAI? Бизнес должен взвесить все «за» и «против» для его внедрения. Технология получает большую защиту, появляется правовое регулирование — вполне вероятно это поможет компаниям лучше понять, как GenAI применим в их отрасли.