

СБЕР Про x TADVISER

Новая архитектура рынка генеративного ИИ:

от моделей к агентным мультимодальным системам

как

?

компании применяют
генеративный ИИ

насколько

?

зрелыми становятся
новые решения

что

?

мешает их
масштабированию

среде: **агентный и мультимодальный ИИ**. Мы рассматриваем основные сценарии их использования, накопленный опыт внедрения, существующие ограничения и проблемы, а также оценки заказчиков и экспертов рынка. Задача обзора — показать, какие сценарии уже доказали свою применимость, где технологии достигли уровня, достаточного для практического использования, а где сохраняются существенные ограничения и риски.

О методологии

Основу обзора составляют три независимо проведенных исследования.

Первое — данные открытых источников: аналитика международных исследовательских компаний, публичные кейсы внедрений, отраслевая пресса.

Второе — глубинные интервью с ИТ-лидерами крупных российских компаний: производственных, торговых, девелоперских, фармацевтических, агропромышленных, розничных. Мы спрашивали их о том, как генеративный ИИ применяется в их организациях на самом деле: какие сценарии дошли до продуктива, какие застряли в пилоте и почему, как считается экономика проектов, каково отношение компаний к тому, что их сотрудники используют публичные модели. Отдельно свои комментарии дали представители поставщиков технологий: специалисты «Сбера» и фонда «Сколково», отвечающие за внедрение ИИ-технологий у крупных заказчиков, а также руководитель практики Data & AI компании Axenix и руководители подразделений в «Яндексе».

Третье — количественный опрос. Параллельно с интервью был проведён онлайн-опрос более 100 руководителей, отвечающих за ИИ-повестку в российских компаниях из различных секторов: ИТ-директоров, директоров по цифровой трансформации, директоров по ИИ, руководителей бизнес-функций, участвующих в решениях о старте ИИ-проектов и закупке решений. Опрос позволил оценить распространённость практик или проблем среди компаний, уже занимающихся генеративным ИИ.

В тексте обзора выводы этих трех исследований сопоставляются и иногда противопоставляются, что позволяет получить многомерную картину состояния развития и использования генеративного ИИ в России.

Ключевые выводы исследования

Российская специфика развития ГенИИ имеет ряд структурных отличий от мировой. Среди них:

- особая роль государства как заказчика, регулятора и соинвестора;
- концентрация рынка вокруг трёх-четырёх экосистем вместо стартапов;

- приоритет on-premise-развёртывания;
- открытый код как основной способ доступа к передовым возможностям;
- фактически свёрнутый венчурный контур;
- высокая стоимость капитала, смещающая выбор в пользу проектов с горизонтом 12-18 месяцев.

Это устойчивые характеристики рынка, а не временные обстоятельства. В результате задачи, которые формально выглядят одинаково, могут заметно различаться по трудоемкости: там, где западный заказчик использует готовый коннектор из библиотеки, российскому приходится реализовывать отдельный интеграционный проект.

Рынок расслоился по зрелости. Лидеры — банки, телеком, ИТ и e-commerce — имеют продуктивные масштабированные внедрения с измеримой экономикой. Крупная промышленность, ритейл, страхование, логистика и энергетика находятся на этапе масштабирования. Часть промышленности, строительство, транспорт, фарма и госсектор составляют осторожное большинство. **Банковский сектор — главный двигатель агентного ИИ.** Около половины всех задокументированных агентных кейсов российской выборки так или иначе связаны с банками, причём технология освоена сразу по нескольким направлениям: клиентский сервис, разработка ПО, комплаенс, поддержка сотрудников. Банки выступают одновременно потребителями и разработчиками платформ, поэтому отработанные ими подходы тиражируются в другие отрасли через тех же интеграторов и те же платформы.

Реальный набор сценариев внедрения шире публичной витрины. Среди наиболее часто упоминаемых участниками исследования направлений — разработка ПО (47%), аналитика данных и подготовка отчётности (47%), создание маркетингового контента (45%). При этом клиентский сервис — главный витринный сценарий публичного поля — назвали лишь 22% опрошенных. Сценарии, не создающие поводов для пресс-релиза, вроде работы с договорами, распространены заметно шире, чем принято думать.

Экономика стала входным условием для проектов, а заявленный эффект систематически не подтверждается. С 2025 года финансовые директора требуют оценивать экономический эффект до начала проекта. При полной оценке стоимости владения эффект оказывается существенно меньше обещанного. Помимо приукрашивания публичных цифр есть и более глубокая причина: в отрасли еще нет общепринятой методологии измерения. Даже в разработке ПО, сценарии с наиболее очевидным эффектом, нет согласия о том, что именно необходимо измерять.

Корневые барьеры внедрения — зрелость данных и процессов и компетенции в ИИ. Компаниям не хватает не только компетенций в ИИ, но и понимания собственных процессов, а также качества собственных данных. Формализованное описание процессов покрывает, по оценке интеграторов, лишь 10-20% того, что происходит в действительности. Несмотря на то, что среди причин недоведения проектов до продуктивной стадии респонденты чаще всего называют

несовершенство моделей (52%), интервью показывают, что в большинстве разобранных случаев за этой формулировкой стоит не качество модели как таковой, а неописанные данные и процессы на входе. Соседние позиции того же списка - низкое качество данных (48%) и сложности интеграции (46%) - описывают ту же ситуацию более прямо.

Следом идут неготовность инфраструктуры (43%) и внутренние ИБ-политики (42%); оба фактора называются чаще, чем высокая стоимость решений (33%) и отсутствие явных эффектов (31%). Требования безопасности, в свою очередь, часто предполагают работу в закрытом контуре, а значит, предъявляют серьезные дополнительные требования к инфраструктуре.

Конкуренция сместилась на качество данных, инфраструктуру интеграции и экспертизу команды внедрения, а предметом инженерной работы стала обвязка вокруг модели — маршрутизация запросов между моделями разной стоимости, трассировка, контроль качества ответов и управление стоимостью вычислений.

Высокий отсев пилотов не сводится к незрелости рынка. Почти две трети опрошенных наблюдали в своих компаниях случаи, когда пилотные ИИ-проекты не доходили до устойчивого продуктивного состояния. Однако в зрелых организациях отказ от части пилотов изначально заложен в процесс: одна из задач пилотирования — на раннем этапе выявить проекты, дальнейшее развитие которых экономически нецелесообразно. Наряду с этим действуют две другие причины: отсутствие управленческого решения о промышленном применении и неготовность владельца бизнес-функции принять новый порядок работы.

Зрелость управления отстаёт от внедрений. Распространение агентов уже привело в мире к росту числа «нечеловеческих» учётных записей в корпоративных системах на 76%, а средний уровень зрелости управления ответственным ИИ McKinsey оценивает в 2,3 балла из 5 — порог в 3 балла преодолела лишь треть организаций. В российских компаниях проблема выглядит приземлённее: появляется набор разрозненных решений без общей архитектуры, а агент с доступом к системам оказывается новым типом «сотрудника», для которого нет готовых моделей прав, аудита и ответственности. Мешает и терминология: пока «агентом» называют и сценарный чат-бот, и автономную многошаговую систему — объекты управления просто не определены. У отставания есть и преимущество: компания с тремя-четырьмя агентами выстроит единый контур существенно дешевле, чем та, которой придётся наводить порядок в десятках внедрений.

Роли меняются раньше, чем кадровые процедуры, а обучения не хватает. Первыми меняются требования к навыкам, должностные обязанности перестраиваются значительно медленнее. Между ними лежит зона, где сотрудник работает по-новому, а описан по-старому. При этом более трети компаний пока вообще не ведут обучение навыкам ИИ.

Автономность агентов ограничена серединой процесса, и это снижает эффект. Агент берёт на себя середину цепочки, на входе и выходе остаётся человек, а

автономность дозируется по цене ошибки. Пока ИИ ассистирует, не меняя устройства процесса, сквозной выигрыш составляет около 10%. Нелинейный эффект появляется на следующей ступени зрелости, где человек перемещается из цикла на позицию над циклом, — там же сосредоточены и основные барьеры. Готовых фреймворков, аналогичных тем, которые задавали ориентир в прошлую волну цифровизации, для этого перехода пока не существует.

Российская мультимодальность держится на двух опорных парах: речи и документах. В продуктиве работают связки «речь плюс текст» и «документ плюс текст». Вторая распространена не меньше первой, но собирается внутренними командами и реже попадает в пресс-релизы. Свободная визуальная мультимодальность остаётся в пилотной зоне сразу по нескольким причинам: качество, стоимость обработки, интерпретируемость, дефицит размеченных русскоязычных корпусов.

Теневой ИИ — норма, а не исключение. Использование личных аккаунтов в публичных моделях для рабочих задач — очень распространённая практика, с которой чаще пытаются не бороться, а обратить её на пользу либо предоставить сопоставимую по удобству легальную альтернативу.

Повестка государственного сектора совпадает с корпоративной. Пилотный эксперимент по внедрению ИИ в Аппарате Правительства Российской Федерации подтвердил применимость технологии для текстовых, аналитических, правовых и коммуникационных задач и показал, что условиями масштабирования являются защищённая инфраструктура, интеграция с действующими информационными системами и подготовка сотрудников, а не возможности самих моделей. Содержательные и управленческие решения, как и ответственность за них, остаются за человеком, а требование объективной оценки эффекта выделено отдельно.

Рынок корректирует ожидания, но не сворачивает направление. Расширять области применения генеративного ИИ планируют более трёх четвертей компаний, а отказаться от технологии или приостановить её применение собираются единицы. Разочарование отдельными проектами не превращается в разочарование технологией.

Генеративный, агентный и мультимодальный ИИ: о чём идёт речь

Раздел рассчитан на тех, кто только начинает осваивать тематику. Уверенно чувствующие себя в этой сфере могут перейти к следующему.

Генеративный ИИ — класс ИИ-систем, которые обучаются на больших массивах данных, выявляют их смысловые закономерности и **создают новый контент** или новые варианты решений, сохраняющие характер обучающих данных, но не сводящиеся к простому воспроизведению исходных примеров.

Агентные системы — ИИ-системы, ориентированные не только на генерацию ответа, но и на достижение цели через **цикл определённых действий**: восприятие контекста → рассуждение → планирование → выбор действия → исполнение → корректировка результата. Такие системы способны использовать внешние данные, инструменты и системы, действовать с частичной или высокой автономией и выполнять многошаговые задачи от имени пользователя или организации.

Степень автономии агента варьируется от «человека в цикле» (Human-in-the-Loop), когда согласовывается каждый значимый шаг, до высокой автономии агента в принятии решений. Российская практика пока сосредоточена ближе к первому полюсу.

Мультимодальные системы — ИИ-системы, которые могут одновременно воспринимать, сопоставлять и интегрировать несколько типов данных (текст, изображения, аудио, видео, код), чтобы формировать более полное представление о ситуации и на этой основе давать более точные выводы, решения или генерацию результата в одной или нескольких модальностях.

Мировой рынок: контекст

Российский рынок имеет свою специфику, однако развивается в русле глобальных технологических тенденций: общими остаются базовые модели, протоколы, архитектурные подходы и основные направления применения технологий. Поэтому международный контекст важен для анализа российского рынка: он позволяет понять, какие тенденции носят общий характер, а какие определяются местными условиями.

Масштаб рынка

Рынок агентного ИИ в западных странах оценивается примерно в 7 млрд долларов в 2025 году с перспективой роста до 45+ млрд к 2030 году. Рынок мультимодального ИИ пока меньше — около 2,5 млрд с выходом на 42 млрд к 2034 году. По оценке Gartner, к концу 2026 года 40% корпоративных приложений будут включать задачно-ориентированных ИИ-агентов против менее чем 5% в 2025 году.

Параллельно западному сформировался второй центр — Китай, где рынок облачных AI-сервисов за 2025 год удвоился, а доля китайских моделей в мировом использовании генеративного ИИ выросла примерно с 1% до 15%. Экосистемы двух центров расходятся на уровне чипов и регулирования, но переплетены на уровне открытого кода: серия Qwen от Alibaba стала самой скачиваемой на Hugging Face, модели DeepSeek доступны через AWS Bedrock и Azure AI Foundry.

Как устроена экосистема поставщиков

Рынок выстроен как многоуровневая цепочка: производители чипов — разработчики фундаментальных моделей (OpenAI, Anthropic, Google DeepMind, Meta, Mistral; DeepSeek, Alibaba, ByteDance) — облачные гиперскейлеры — фреймворки разработки агентов — корпоративные бизнес-платформы — вертикальные отраслевые решения. Параллельно всем уровням работают системные интеграторы и консультанты.



На уровне карты поставщиков различие между агентным и мультимодальным ИИ практически стёрлось: ведущие модели по умолчанию и мультимодальны, и агентны.

Заказчику важнее различать три способа получить агентную функциональность:

- **корпоративные платформы** с преднастроенными под бизнес-процессы агентами (Agentforce у Salesforce, AI Agents у ServiceNow, Joule у SAP);
- **сборку на фреймворках** (LangChain / LangGraph, CrewAI, LlamaIndex, Microsoft Agent Framework), дающую гибкость ценой инженерных затрат;
- **вертикальные отраслевые решения** — слой самый многочисленный, где капитализация значит меньше предметной экспертизы.

Стандартами де-факто стали **открытые протоколы MCP (Anthropic) и A2A (Google)**: первый описывает подключение агентов к внешним инструментам, второй — взаимодействие агентов разных производителей. Ожидается, что вокруг них сформируется агентная экономика — библиотеки готовых агентов, корпоративные «магазины» решений, вендоры компонентов. Процесс сопоставим со становлением рынка веб-API в 2010-х, где стандарт закреплялся быстро, а экосистема вызревала за два-три года. Порог входа при этом снижается, но на сложные корпоративные внедрения облегчение почти не распространяется.

Изменился и слой **системных интеграторов**: крупные игроки строят собственные платформы (AI Refinery у Accenture, watsonx Orchestrate у IBM, Zora AI у Deloitte, Agent OS у PwC) и продают в первую очередь управляемость.

Что меняется в самих внедрениях

Модель перестала быть источником преимущества, а предметом инженерной работы стала архитектура. По Stanford AI Index 2026, в марте 2026 лидирующая

закрытая модель опережала лучшую открытую на 3,3% (против 0,5% в августе 2024)». Разрыв между лучшими открытыми и проприетарными моделями сократился до менее чем 3%, и результаты гонки качества всё меньше влияют на исход конкретного внедрения: на первый план выходят качество данных, зрелость инфраструктуры и экспертиза команды. Одновременно усложняется конструкция решений: на смену пилотам вокруг одного агента с одной функцией приходит связка «оркестратор плюс команда специализированных исполнителей», которую Gartner относит к пяти ключевым трендам 2026 года. Проектировать приходится уже не только качество отдельного шага, но и распределение задач, передачу контекста и разбор ситуаций, где каждый агент принял локально верное решение, а совокупный результат оказался нежелательным.

Приближается порог производственной надёжности. Сегодня агенты справляются примерно с двумя задачами из трёх на ключевых компьютерных бенчмарках. Преодоление планки 85–90% сделает автономную работу без постоянного контроля экономически оправданной и сместит логику проектирования от «агента под надзором» к «агенту, отвечающему за результат».

Мультимодальность становится свойством корпоративного ПО по умолчанию, а отраслевые сдвиги видны уже сейчас. Флагманские модели обрабатывают текст, изображения, аудио и видео в едином контуре, и эта способность распространяется вниз по стеку: по прогнозу Gartner, к 2030 году 80% корпоративного ПО будет мультимодальным против менее 10% в 2024-м. В клиентском сервисе идёт переход к схеме agent-first, в разработке ПО роль инженера смещается к оркестратору команды агентов, в кибербезопасности начинается «гонка агентов» по обе стороны. На горизонте нескольких лет ожидается распространение персональных наборов агентов под каждую роль, автономное планирование в цепочках поставок и ускорение R&D. Так, например, цикл идентификации лекарственных мишеней сократится с 60–80 дней до 4–8.

Мировой рынок переходит из фазы, где обсуждались возможности моделей, в фазу, где обсуждаются управляемость, ответственность и организационная готовность: технологический выбор упрощается, а управленческая часть задачи усложняется — именно в ней сосредоточен основной риск неудачного внедрения. Эта проблематика в значительной степени распространяется и на Россию, что можно будет увидеть в следующих частях обзора.

Российский рынок: в чем его специфика

Структурные особенности

Российский рынок отличается от мирового не столько темпом, сколько принципами устройства. Это важное различие: отставание со временем сокращается само, а структурные особенности задают логику рынка надолго и требуют не догоняющих, а

собственных решений. Ниже описаны те характеристики, которые определяют, как принимаются решения о внедрении, у кого покупают технологию и почему одни мировые сценарии воспроизводятся, а другие остаются нерелевантными.

Нужно отметить, что для агентных и мультимодальных систем эти особенности значат больше, чем для генеративного ИИ в целом. Чат-бот, работающий на корпоративной базе знаний, относительно нечувствителен к тому, где развёрнута модель и есть ли вокруг неё экосистема готовых компонентов. Агент, которому нужен доступ к десятку корпоративных систем, и мультимодальная система, обрабатывающая потоки аудио и видео, оказываются чувствительны и к архитектуре развёртывания, и к наличию коннекторов, и к стоимости вычислений. Именно поэтому структурные отличия российского рынка сейчас, когда агентный и мультимодальный сценарии внедрения ИИ получают все большее распространение, должны быть проанализированы и учтены при принятии решений.

Особая роль государства. Государство на российском рынке ИИ выступает одновременно крупнейшим заказчиком, регулятором и соинвестором. Госкорпорации и компании с государственным участием — Росатом, Ростелеком, Газпром-Нефть ([Forbes](#), [rt.ru](#), [Ведомости](#)) — являются крупными потребителями ИИ-решений, а государство формирует спрос через обязательную цифровизацию госсектора. Существуют налоговые льготы для покупателей российских решений с ИИ. Есть и более радикальные инструменты: Федеральный закон №233-ФЗ позволяет Минцифры напрямую требовать от операторов персональных данных предоставления обезличенных данных для нужд государственного ИИ. Государство также содействует развитию вычислительной инфраструктуры, поддерживая строительство энергоэффективных ЦОДов в Сибири и на Дальнем Востоке — в непосредственной близости к центрам генерации электроэнергии.

Концентрация рынка вокруг экосистем, а не стартапов. Глобальный рынок определяется большим числом стартапов и широким венчурным финансированием. Российский устроен принципиально иначе: его движущими силами являются три–четыре крупные корпоративные экосистемы: Сбер (GigaChat, Kandinsky, GigaCowork, GigaAgent), Яндекс (YandexGPT, Yandex Cloud, AI Studio), МТС (Cotype, MWS GPT, DevX) и, в меньшей степени, VK (ruGPT) и Т-Банк (T-Pro). Стартапы существуют — OSMI AI, Nura AI, Just AI, «Наносемантика», однако они не определяют рынок, а встраиваются в экосистемы или занимают узкие ниши.

Там, где на мировом рынке заказчик выбирает из десятков поставщиков и может использовать конкуренцию как рычаг, российский покупатель выбирает из трёх–четырёх экосистем, каждая из которых предлагает связанный набор моделей, облака и инструментов. Вторая сторона той же медали — внутри выбранной экосистемы интеграция действительно проще, поскольку компоненты изначально рассчитаны друг на друга.

Приоритет on-premise-архитектуры. Требования безопасности — особенно для госсектора, оборонных предприятий и критической инфраструктуры — предписывают изолированное развёртывание значительно чаще, чем на Западе. Это формирует специфический спрос на «закрытые» решения: программно-аппаратные комплексы (ПАК-AI, «Скала^p МИИ», «Лира»), локальные MaaS-платформы (MWS GPT on-premise), изолированные агентные системы.

«По их [регулятора] рекомендации те данные, которые потенциально несут высокую ценность и риск утечки, обрабатываются внутри. Поэтому мы отчасти пошли в собственную инфраструктуру, в собственный пилот, используем свои мощности, свой контур, не отдавая это никуда», — представитель фармацевтической компании

«Если есть локальная модель на внутреннем сервере в локальной сети — мы полностью независимы от внешнего периметра, работает стабильно даже если интернет отключится», — представитель торгово-производственной компании

Экономические последствия этого выбора сильно влияют на логику проектов. Облачная модель позволяет начать с минимальных вложений и наращивать потребление по мере подтверждения эффекта. On-premise требует закупить и настроить всю необходимую вычислительную инфраструктуру до того, как получен первый результат. Это удлиняет цикл запуска, повышает порог входа и, что важнее, превращает пилот из недорогого эксперимента в инвестиционный проект, требующий отдельного обоснования. Именно поэтому в российской практике решение о технологической архитектуре часто предопределяет судьбу пилота ещё до его старта. При этом облачное предложение на рынке существует — все крупные провайдеры (Cloud.ru, Yandex Cloud, MWS) предоставляют доступ к моделям и ИИ-сервисам ([Cloud.ru](#), [Хабp](#), [MWS AI](#)), но заказчики выбирают его чаще для быстрых пилотов и некритичных данных.

У этого выбора есть и качественное следствие, о котором говорят консультанты: закрытый контур сужает не только способ развёртывания, но и множество доступных решений.

«Ведущие [модели] — облачные, а для многих компаний, являющихся объектами критической инфраструктуры, да и просто, когда речь идёт о конфиденциальной, коммерческой информации, никто за периметр выпускать не хочет — все хотят иметь инфраструктуру on-premise. А если on-premise, значит, должно быть развёрнуто в собственных ЦОДах, а это только open source, либо что-то проприетарное от отечественных поставщиков, чьи модели всё-таки отстают по своим возможностям от лучших облачных вариантов», — Лариса Малькова, Axenix

Иными словами, требование изолированного развёртывания автоматически выводит за скобки лучшие облачные решения — независимо от того, доступны они формально или нет. К этому добавляется удорожание самой инфраструктуры: по наблюдению того же эксперта, поставки оборудования и видеокарт стали дороже и логистически сложнее, что поднимает порог входа ещё выше.

Открытый код как суверенный ресурс. В России открытые модели — DeepSeek, Llama 4, Qwen, ruGPT — играют иную роль, чем на Западе. Если там open-source — это вопрос философии разработки и снижения затрат, то в России открытые модели стали основным способом получить доступ к передовым возможностям ИИ в условиях, когда проприетарные западные модели недоступны. DeepSeek-R1 стал в России лидером по популярности именно как открытая высококачественная модель, которую можно развернуть в собственном контуре. При этом GigaChat как открытая российская суверенная модель поддерживает задачу независимости страны от зарубежных решений.

«Qwen понравилась — подошла на наш GPU, перевели с CPU на GPU, скорость увеличилась в 20 раз, получился нормально работающий чатик, и бесплатно», — представитель торгово-производственной компании

Эта короткая реплика описывает целую модель поведения: заказчик самостоятельно подбирает модель под имеющееся железо, добивается приемлемой производительности инженерными средствами и получает работающий инструмент без лицензионных платежей. Такой путь предполагает наличие внутренней технической команды, но снимает зависимость и от вендора, и от валютных расчётов.

Формирование уникальной экспертизы «экономичных моделей». Прямое следствие предыдущего пункта — специфическая культура оптимизации. Российские разработчики вынуждены добиваться большего при меньших ресурсах, что ведёт к широкому использованию квантизации, дистилляции и дообучения малых моделей вместо развёртывания больших. По сути, в России формируется экспертиза в области экономичного инференса, которая с учётом массового роста спроса на ИИ-вычисления может оказаться востребованной и за её пределами.

Отсутствие экосистемы вокруг агентных протоколов. MCP и A2A возникли в западной экосистеме и распространяются через западные open-source-решения. Российские разработчики технически могут использовать эти протоколы — например, Yandex AI Studio декларирует поддержку MCP ([Yandex Cloud](#)), — однако экосистемная поддержка (библиотеки готовых агентов, магазины компонентов, сертифицированные интеграции) развита существенно меньше.

«Настраиваем MCP-коннекторы — это внутренняя IT-трансформация», — представитель торгово-производственной компании

Формулировка респондента заслуживает внимания. В мировом контексте подключение агента к корпоративным системам через MCP описывается как техническая деталь, снижающая порог входа: достаточно выбрать подходящий коннектор из готовой библиотеки и подключить его. В российских условиях того же результата приходится добиваться собственными силами, и задача из технической детали превращается в самостоятельный ИТ-проект со своим бюджетом, сроками и рисками. Это одно из существенных различий между мировой и российской практикой: одинаковые по названию работы имеют здесь принципиально разную трудоёмкость.

Высокая ключевая ставка как фактор предпочтения быстрых побед. В 2025 году ставка снизилась с 21% до 16%; в 2026 году остаётся повышенной (14% на октябрь 2026)». При такой стоимости капитала компании переориентируются на ИИ-проекты с быстрым и предсказуемым возвратом на горизонте 12–18 месяцев, откладывая долгосрочные трансформационные инвестиции.

«Мы ограничиваем себя в размере пилота по деньгам — есть порог затрат, выше которого не идём. По срокам — все пилоты не больше двух месяцев», — представитель девелоперской компании

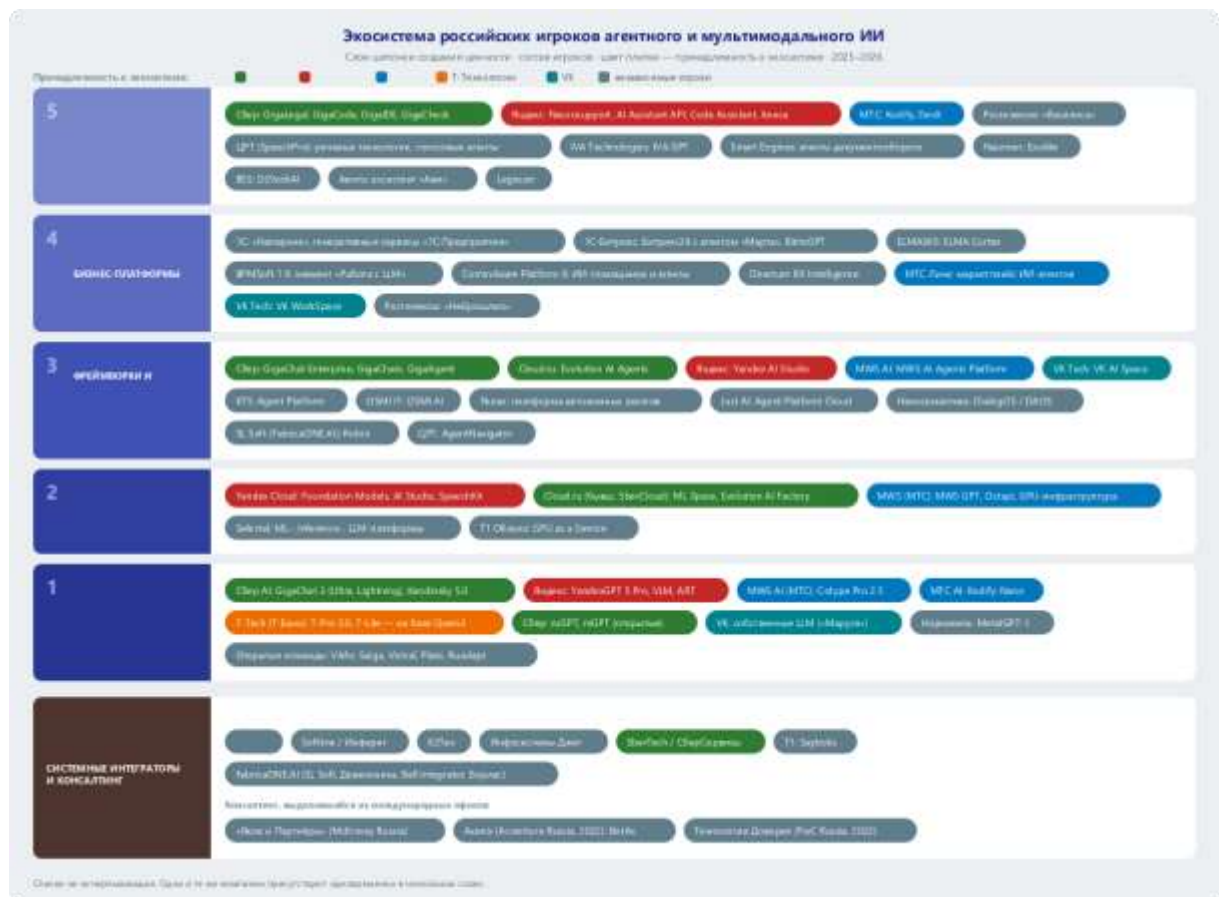
Такой подход позволяет на раннем этапе отказаться от заведомо неэффективных проектов, однако у него есть и обратная сторона. Жёсткие ограничения по бюджету и срокам могут привести к отказу от перспективных сценариев, которым для раскрытия потенциала требуется предварительная подготовка.

Стимулы для развития локальной проектной экспертизы. Уход с российского рынка международных компаний, сочетавших глубокую ИИ-экспертизу, отраслевой опыт и масштаб для трансформационных проектов, таких как McKinsey, Accenture, IBM, PwC, Deloitte, создал разрыв, который в основном закрывают российские интеграторы и консалтинговые компании, получая при этом стимулы для собственного развития. Дополнительное преимущество отечественных команд — понимание специфики российской экономики и меньшая склонность к ошибкам, характерным для международных интеграторов при работе на локальных рынках.

Экосистема российских игроков

Российская экосистема повторяет мировую логику цепочки создания ценности — от фундаментальных моделей через облака и инструменты разработки к бизнес-платформам и отраслевым решениям.

Отличается не устройство, а плотность: игроков на каждом уровне меньше, и одни и те же компании занимают сразу несколько уровней.



Слой моделей держат подразделения нескольких корпораций — Сбера, Яндекса, МТС, VK и «Т-Технологий» ([TAdviser](#), [VC.ru](#)): модели создаются там, где есть капитал, собственные данные и внутренний заказчик. Сбер выкладывает веса всех своих моделей, в том числе единственной российской модели с рассуждениями GigaChat 3.5 Reasoning под открытой лицензией MIT, «Норникель» — отраслевую MetalGPT-1. Открытый код здесь не идеология, а способ набрать аудиторию разработчиков на замкнутом рынке.

Облачный слой устроен иначе, чем западный: гиперскейлеров нет, а ведущие провайдеры — Cloud.ru, Yandex Cloud, MWS — входят в те же корпоративные экосистемы, что и разработчики моделей. Независимых игроков единицы, и выбор облака нередко предопределяет и выбор модели.

Инструменты разработки агентов, в основном, создают те же корпорации: GigaChat Enterprise у Сбера, AI Studio у Яндекса, MWS AI Agents Platform у МТС, VK AI Space у VK Tech. Рядом — около десятка независимых команд, от ЦРТ и Just AI до Nurax и OSMI IT. Совместимость с MCP и A2A декларируется всё чаще, но экосистемы готовых компонентов уровня LangGraph или CrewAI здесь пока нет.

Слой корпоративных бизнес-платформ заполняется очень быстро. Ещё в 2024 году вендоры ограничивались отдельными генеративными функциями. За 2025–2026 годы картина изменилась. ELMA встроила платформу агентов Cortex в процессное ядро ELMA365, BPMSoft — элемент работы с LLM в конструктор процессов, Comindware вывела агентов в Platform 6, «МТС Линк» открыл маркетплейс готовых агентов ([МТС Линк](#)), VK Tech связала VK WorkSpace с VK AI Space ([VK Tech](#)). Аналога Salesforce Agentforce по охвату пока нет, но разрыв здесь сокращается быстрее, чем на других уровнях.

Вертикальный слой самый населённый. Конкурентоспособность на этом уровне определяется знанием отрасли и качеством данных, а не капитализацией, — и по этим параметрам отечественные игроки сопоставимы с западными. Но устроен он иначе: решения создаются внутри корпораций, либо силами давно работающих в нише компаний — ЦРТ, Naumen, BSS, Smart Engines. Венчурного контура, питающего этот слой на Западе, здесь нет.

Слой интеграторов и консультантов пересобран после 2022 года: международных игроков не осталось, часть их роли взяли команды из бывших российских офисов — «Яков и Партнёры», Ахених, «Технологии Доверия» ([ComNews](#), [Интерфакс](#)), а часть перешла к ИТ-группам: [Softline](#), [T1](#), [K2Tex](#), [Ланит](#). Сместился и спрос: нужна не поставка продукта, а способность довести изменение до результата.

Что из этого следует практически. Российский рынок концентрируется вокруг нескольких корпоративных экосистем. Такой подход имеет и преимущества, и недостатки: внутри экосистемы интеграция проще и дешевле, но конкуренция слабее, а переход к другому вендору дороже. Выбирая модель, компания нередко выбирает заодно облако, инструменты и интегратора, — решение о поставщике принимается на уровне архитектуры, а не отдельной закупки.

Где генеративный ИИ реально работает: российская практика

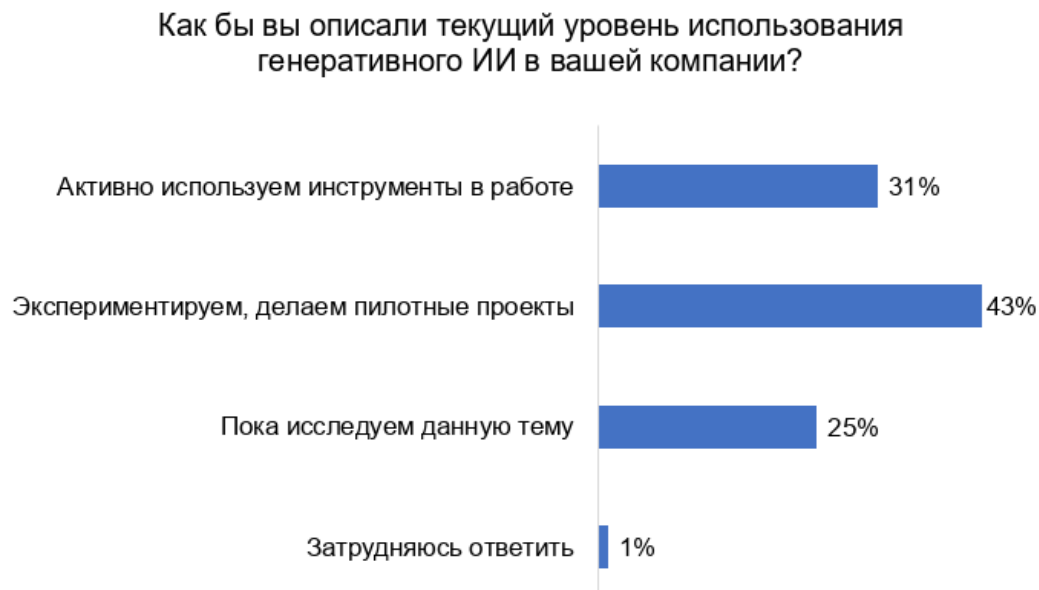
Общая картина

Россия вписывается в мировую картину внедрений лишь частично. По ряду сценариев — прежде всего в клиентском сервисе и автоматизации разработки ПО — отечественные лидеры демонстрируют результаты, сопоставимые с мировыми. Однако ширина отраслевого охвата заметно уступает: ряд сегментов с высокой глобальной зрелостью в российской выборке практически отсутствует или обозначился только первыми пионерскими проектами.

Среди черт российского рынка агентного и мультимодального ИИ, характеризующих его развитие, можно выделить следующие:

Основная масса внедрений продолжает находиться в пилотной стадии.

Здесь данные опросов и интервью имеют расхождение в цифрах, однако сходятся в принципиальном соотношении. Так, данные опросов говорят о 43% доле проектов на стадии экспериментов и пилотов и о 31% активного использования. Оценка же, данная в интервью со стороны интегратора, звучит иначе: около 10–15% холдингов вообще не пытаются проводить пилоты, порядка 60–65% пилотируют несистемно, и лишь 10–15% имеют продуктивные внедрения.



Данные опроса «Как бы вы описали текущий уровень использования генеративного ИИ в вашей компании?»

Это расхождение можно объяснить тем, что респонденты количественного опроса оценивают стадию использования не так строго, как интеграторы: для первых «активное использование» может означать в том числе и т.н. «теневое использование», когда сотрудники начинают пользоваться инструментами генеративного ИИ по собственной инициативе, в то время как вторые имеют дело с крупным корпоративным сектором, в котором критерии перехода от пилотного использования к продуктивному значительно более жесткие.

Суверенная технологическая инфраструктура. Сложившийся в России путь стимулировал форсированное развитие собственных LLM-платформ: GigaChat и SaluteSpeech Сбера, YandexGPT, AlfaGen Альфа-Банка, Cotype МТС. Это одновременно ограничение (меньший выбор базовых моделей, меньше международной экспертизы) и точка концентрации инвестиций: крупнейшие компании вкладываются в собственные платформы, а не только в настройку готовых решений.

Российская мультимодальность держится на **двух опорных парах модальностей**: «речь плюс текст» и «документ плюс текст». Если глобально ведущими модальностями для бизнеса являются речь и компьютерное зрение, то в российских продуктивных внедрениях доминируют распознавание и синтез речи, с одной стороны, и работа со сканами, формами и первичными документами, с другой. Свободная визуальная мультимодальность — распознавание неструктурированных изображений и видео — пока преимущественно на стадии пилотов.

Коммодитизация моделей: российские особенности. Разрыв между российскими проприетарными и доступными открытыми моделями для большинства бизнес-задач невелик. Массовые задачи уходят в эффективные модели среднего класса, конкуренция идет за цену токена и op-rgem; это требует собственной линейки моделей, а не отменяет ее, в том числе поднимая ценность малых моделей (SLM). Но добавляется ценовое измерение — оно актуально и в мире, однако благодаря интервью мы имеем возможность получить обратную связь от российского заказчика:

«...в сравнении с китайскими альтернативами, если бы условный Яндекс или Гигачат был бы сильно дешевле, мог бы быть конкурентоспособным... Когда ты считаешь это на массовых запросах, ты удивляешься и думаешь: мне проще купить GPU и посадить китайскую модель, мне дешевле выйдет», —
представитель ритейлера спортивных товаров

Логика здесь сугубо экономическая: при сопоставимом качестве выбор определяется стоимостью токена на массовых нагрузках, и в этом сравнении покупка собственного железа под открытую модель нередко выигрывает у платного API. Регуляторные соображения вступают в силу позже — там, где речь идёт о чувствительных данных.

Переход от RAG-чат-ботов к агентным системам российский рынок проходит по той же траектории, что и мировой, с отставанием на один-два года. Большинство компаний сейчас движется от базовых RAG-решений к первым агентным архитектурам. Мультиагентные системы — следующий этап, ещё не ставший стандартом.



Данные опроса «Какие из указанных типов ИИ-агентов уже есть или находятся в процессе внедрения у вас в компании?»

«Уже есть 7 ботов в продуктиве, настроен мониторинг, логирование ошибок — это энтерпрайз-решение... Сейчас уже группа сотрудников разрабатывает и внедряет в свои рабочие процессы», — представитель торгово-производственной компании

Сопровождается этот переход определенной терминологической неразберихой, которую стоит иметь в виду и в дальнейшем:

«У нас немножко размытое по компании понимание, что такое агент. Каждое подразделение, особенно бизнесовое, понимает это по-своему, потому что приходят подрядчики-вендоры и начинают словом "агент" даже чат-бота называть», — представитель ритейлера спортивных товаров

Банковский сектор — главный двигатель агентного ИИ

Около половины всех задокументированных агентных кейсов российской выборки, продуктивных и пилотных вместе, так или иначе связаны с банками. Это закономерно: финансовый сектор располагает крупнейшими в стране ИТ-бюджетами, исторически высокими расходами на персонал контакт-центров и аналитических подразделений, а также огромными массивами текстовых данных: транзакционных, клиентских, нормативных. Но важнее другое обстоятельство: в

отличие от большинства отраслей, где агентные внедрения концентрируются в одной точке приложения, банки освоили технологию сразу по нескольким направлениям: в клиентском сервисе, разработке ПО, комплаенсе, поддержке сотрудников. Именно широта фронта, а не глубина отдельного кейса, отличает финансовый сектор от остальных.

Четыре крупнейших банка страны — **Сбербанк, Альфа-Банк ([Альфа-Банк](#)), ВТБ ([АСРОС](#)) и Совкомбанк** — в 2025-2026 году перевели агентные решения в промышленную эксплуатацию. Сбербанк создал систему сервисных AI-агентов для дистанционного обслуживания бизнес-клиентов по наиболее востребованным продуктам: расчётам, платежам, зарплатным проектам, эквайрингу; финансовый эффект от GenAI для бизнеса банка в 2025 году оценивается не менее чем в 50 млрд рублей ([CNews](#), [Mail.ru](#)). Совкомбанк пошёл по пути максимального охвата: ИИ-помощник используется восемь из десяти сотрудников ежедневно, совокупная экономия составляет порядка 310 тыс. рабочих часов в год, в разработке — более трёх десятков проектов автономных агентов ([РИА Новости](#), [Generation AI](#)).

Пилоты расширяют эту картину в сторону регулируемых процессов.

Альфа-Банк совместно с GlowByte Consulting завершил в начале 2026 года пилот агента валютного контроля: обработка одного внешнеторгового контракта сократилась с двух часов до одной минуты при пропускной способности порядка 2 700 контрактов в день и точности около 80% ([ComNews](#), [Ведомости](#)). Это первый случай в России, когда агентный ИИ автоматизировал функцию, ранее выполнявшуюся исключительно людьми в рамках жёстко регулируемого банковского процесса — и именно поэтому кейс важен как прецедент, а не только как экономический результат.

Отраслевой пилот Ассоциации ФинТех с участием **Альфа-Банка, ВТБ, Газпромбанка** и других организаций (всего более 20 банков, страховых и ИТ-компаний) позволил создать более десяти прототипов агентов на трёх отечественных LLM от Яндекса, Сбера и MWS AI, — охватывающих анализ обращений в колл-центре, извлечение данных из контрактов, помощь разработчикам и формирование персонализированных депозитных продуктов ([Компьютерра](#)).

Условием такого прогресса стал тот факт, что банки в России выступают не только потребителями агентного ИИ, но и его разработчиками, создавая собственные платформы — GigaChat у Сбера, AlfaGen у Альфа-Банка, агентная среда в платформе «Сфера» у ВТБ ([АСРОС](#)) — и делясь наработками внутри отраслевых консорциумов. Это превращает финансовый сектор в центр компетенций, определяющий траекторию технологии в российском корпоративном контексте: подходы, отработанные в банках, впоследствии тиражируются в другие отрасли через тех же интеграторов и те же платформы.

Помимо клиентских и разработческих сценариев, банки применяют агентный ИИ во внутренних аналитических функциях — и здесь уже появляются полноценные мультиагентные конструкции:

«...это агенты по подготовке, например, ко встрече с клиентом. То есть мы можем объединять внутрикорпоративные данные, внешние источники информации... У тебя есть некие агенты-аналитики: специалист по GR, специалист по IT-рынку, специалист по финансовым рынкам, специалист по корпоративной отчётности — как некая мультиагентность. А с другой стороны, ты можешь создать агента-психолога, который анализирует портрет топ-менеджера и предлагает варианты общения с ним», — Андрей Дмитриев, Сбер

Важно обратить внимание на устройство сценария: каждый агент отвечает за свою предметную область, а ценность возникает из сведения их результатов воедино. Это ровно та мультиагентная схема, которую мировая аналитика относит к ближайшему будущему. В российской практике она уже работает, но пока во внутренних, а не в клиентских процессах, где цена ошибки выше.

Клиентский сервис — самый зрелый сценарий, пронизывающий несколько отраслей

Из всех типов задач, охваченных российской выборкой, клиентский сервис демонстрирует наивысшую зрелость: большинство кейсов уже в промышленной эксплуатации, а показатели автономности обработки обращений сопоставимы с лучшими мировыми примерами. Закономерность не ограничена одной отраслью — она охватывает банкинг, телеком, фармацевтику и ряд промышленных кейсов.

Причина такой зрелости объясняет и то, почему другие сценарии отстают. Клиентский сервис обладает редким сочетанием свойств: обращений много, они однотипны, результат каждого проверяем, а цена отдельной ошибки невысока и обратима — неудачный ответ можно исправить эскалацией на оператора. К этому добавляется готовая метрика эффекта: доля обращений, закрытых без человека, переводится в экономию напрямую, без методологических споров. Ни один другой корпоративный процесс не даёт такой комбинации.

Наиболее измеримый результат — у **ВТБ**: благодаря использованию LLM сегодня около 80% обращений в чат-канале разрешается автоматически, а чат-бот охватывает более 1 800 тематик при клиентской базе «ВТБ Онлайн» около 27 млн пользователей ([CNews](#), [Workspace](#), [РИА Новости](#)).

Схожие результаты в масштабе крупных контакт-центров показывают «**Ростелеком**»: половина обращений через чат обрабатывается автоматически, свыше 23 млн запросов обработано с применением речевой аналитики за 2025 год, показатель первичного решения впервые превысил 80% ([Ростелеком](#), [CNews](#)), в «**СберМобайл**» мультиагентная система из трёх агентов на GigaChat закрывает шесть

из десяти обращений без оператора при удовлетворённости около 80% ([TelecomDaily](#)).

У **КНАУФ** RAG-агент поддержки умеет отвечать на 3000 вопросов по продуктам компании, демонстрирует 89%-ю точность ответов и, по оценке KNAUF, позволяет экономить до 50% рабочего времени сотрудников. ([Generation AI](#)).

Интересный архитектурный паттерн демонстрирует фармацевтика. **Сбер Еаптека** запустила цифрового фармацевта на базе GigaChat: агент за считанные секунды формирует ответы о схемах дозирования, показаниях и противопоказаниях, но отправляет их только после верификации оператором-фармацевтом ([CNews](#)). GigaChat прошёл терапевтическую, а затем кардиологическую аккредитацию, сдав соответствующие экзамены ([Фармзнание](#)). Ценность кейса не в скорости, а в схеме: в регулируемых секторах клиентский сервис реализуется по модели human-in-the-loop, где агент генерирует, а специалист отвечает, и эта схема, вероятно, станет типовой для всех отраслей с профессиональной ответственностью.

Реальный набор сценариев использования — шире охвата публичными кейсами

Публичные кейсы формируют представление о том, где именно генеративный ИИ применяется в России, но это представление заметно расходится с тем, что компании говорят о собственной практике. Чаще всего они называют разработку ПО (47%), аналитику данных и подготовку отчётности (47%) и создание маркетингового контента (45%), тогда как клиентский сервис — главный витринный сценарий публичного поля — упомянут лишь в 22% случаев.



Данные опроса «В каких процессах вашей компании уже применяются технологии генеративного ИИ?»

Сценарий	Применяют	Из них в продуктиве
Разработка ПО, DevOps	47%	64%
Аналитика данных и подготовка отчётности (BI)	47%	42%

Маркетинговые коммуникации, создание контента	45%	74%
Интеллектуальный поиск (RAG) по корпоративным данным	39%	57%
HR, обучение и адаптация сотрудников	30%	57%
Управление проектами	27%	52%
Юридическая функция, работа с договорами	27%	50%

Отдельного комментария заслуживает юридическая функция. Публичных российских кейсов применения генеративного ИИ в работе с договорами практически нет, и это легко принять за отсутствие практики. Опрос показывает иное: почти треть компаний уже применяет здесь генеративный ИИ, и половина из них — в промышленной эксплуатации. Объяснение простое: работа с договорами не создаёт поводов для пресс-релиза, а внутренний эффект при этом вполне ощутим.

Схожая история с маркетинговым контентом: по доле продуктива (74%) сценарий входит в число лидеров. Причина та же, что и у клиентского сервиса, — быстрая обратная связь и низкая цена ошибки. Но публичного внимания он не получает, поскольку выглядит слишком обыденно.

Чем отличаются компании, активно использующие генеративный ИИ

Количественный опрос позволил сравнить две группы респондентов: компании, которые уже активно используют инструменты генеративного ИИ в работе, и компании, которые пока экспериментируют и ведут пилотные проекты без выхода в продуктив. Различия между ними показывают, что меняется в компании по мере накопления практики.

Шире набор сценариев. У активно использующих компаний практически все основные сценарии имеют более высокий охват: в среднем генеративный ИИ применяется у них в 5,4 процесса против 3,2 у экспериментирующих. Особенно заметна разница в следующих процессах:

- управление проектами — 55% против 12%;
- документооборот — 41% против 12%;
- продажи — 38% против 15%;
- клиентский сервис — 38% против 15%;
- маркетинг — 69% против 37%;
- аналитика — 69% против 39%;
- разработка ПО и DevOps — 66% против 39%.

Опора на собственную команду. Среди активно использующих генеративный ИИ 79% выполняют работы по внедрению и настройке решений преимущественно силами внутренней команды, среди пилотирующих — только 50%; на стадии пилотов заметно чаще встречаются смешанные команды и внешние интеграторы. В целом по выборке внутренняя команда выполняет эти работы в 62% случаев, смешанная модель используется в 31%, а полностью на внешнюю команду полагаются лишь 7% компаний. По мере накопления практического опыта компании формируют собственную экспертизу и всё чаще используют её для новых внедрений, снижая зависимость от внешних специалистов. Высказывания поставщиков о растущей роли интеграторов (см. раздел «Мнения экспертов») стоит читать с учётом этой картины: к внешним командам чаще обращаются на входе, а по мере взросления практики от интегратора ждут уже не столько выполнения работ, сколько роли архитектора изменений и долгосрочного технологического партнёра.

Другие барьеры. Ограничения в распространении и масштабировании генеративного ИИ зависят от достигнутого уровня его использования. На стадии пилотов сильнее ощущаются базовые ограничения, связанные с готовностью компании к внедрению технологии, и по мере перехода к активному использованию часть из них снимается. Наиболее заметна разница в оценке готовности инфраструктуры: её неготовность называют ограничением 51% компаний на стадии пилотов и лишь 21% активно использующих генеративный ИИ. Высокая стоимость также чаще становится барьером на стадии пилотов — 41% против 29%.

Заметнее изменения в работе людей. Влияние генеративного ИИ на роли сотрудников отмечают 84% респондентов из компаний, активно использующих генеративный ИИ, и 44% — из компаний на стадии пилотов (о характере этих изменений — в разделе «Влияние генеративного ИИ на роли и компетенции сотрудников»). Вместе с этим растёт и потребность в обучении: его ведут 72% активно использующих компаний против 56% пилотирующих. Осваивая технологию, компании сталкиваются с необходимостью адаптировать навыки сотрудников к новым требованиям и поддерживают эти изменения обучением.

Автоматизация разработки ПО — нишевый лидер

Среди всех сценариев автоматизация разработки выделяется концентрацией зрелых внедрений: три независимых проекта — от **Альфа-Банка** ([Хабр](#)), **Сбербанка** ([CNews](#)) и отраслевого консорциума под эгидой АФТ ([CNews](#)) — вышли на уровень промышленной эксплуатации или завершённого пилота, причём каждый с собственной платформой и архитектурой.

Наиболее развёрнутый кейс — платформа **AlfaGen** Альфа-Банка. Четыре группы агентов по автотестированию помогают шести десяткам платформенных команд мобильного банка: автотесты создаются на 70% быстрее, а задачи, прежде занимавшие месяцы, решаются за дни ([Habr](#)). Агенты действуют автономно: разрабатывают план решения, пишут код, проверяют результат. Система выросла из эксперимента в полноценную мультиагентную платформу с главным агентом и

специализированными субагентами, охватывающими автотестирование, документацию, разработку SDK, код-ревью и суммаризацию видеоконференций ([Forbes](#)).

Сбер пошёл по пути создания публично доступного SWE-агента: GigaCode получил агентный режим, превративший его из инструмента автодополнения в автономного цифрового разработчика: агент открывает файлы, правит код, выполняет консольные команды, запускает тесты, собирает проект и делает коммиты ([CNews](#)). Поддерживается архитектура субагентов, где основной агент вызывает специализированных экспертов для анализа, написания и проверки кода.

Пилот **АФТ** добавляет к картине важный штрих: разработчики оказались одной из трёх приоритетных целевых групп для агентных решений в банковском секторе — наряду с клиентским сервисом и документооборотом ([CNews](#)). То, что несколько крупнейших банков независимо друг от друга пришли к одному выводу, говорит о выраженном рыночном спросе.

За пределами банковского сектора агентная разработка входит в повседневную практику ИТ-подразделений компаний меньшего масштаба, и там она зачастую начинается не с написания кода, а с эксплуатации:

«ИТ-мониторинг: работа с логами, помощь новым сотрудникам, код-ревью — активно используют. Для мониторинга настраивают агентов в графанах, публикуют отчёты. Раньше логи копались в течение дня, сейчас — за минуты, сразу видно проблему», — представитель торгово-производственной компании

«Также активно применяются в разработке агенты, которые делают тесты, агенты, которые код пишут, аналитические агенты, которые изучают данные», — представитель ритейлера спортивных товаров

Обе реплики описывают одну закономерность: агент заходит в разработку через задачи, где результат проверяется быстро и дёшево — разбор логов, генерация тестов, код-ревью — и лишь затем расширяет зону ответственности. Граница автономности проходит там же, где и в других сценариях: агент готовит, человек утверждает.

Разработка ПО предоставляет для агентного ИИ едва ли не идеальные условия: результат верифицируем автоматически (код либо компилируется и проходит тесты, либо нет), задачи стандартизированы, а итеративная проверка встроена в сам процесс. Если добавить к этому острый дефицит ИТ-специалистов, создающий прямой экономический мотив, получится сочетание, которого нет больше нигде. Фактически разработка — единственная область, где обратная связь для агента является машинной, а не человеческой, и это радикально упрощает и обучение, и контроль.

Со стороны поставщиков инструментов разработки картина дополняется данными о фактическом уровне проникновения. В разработке он заметно выше, чем в среднем по компании, но при этом неравномерен:

«Программисты сейчас находятся на фронтире: если среди сотрудников нетехнических профессий пока 64% пользуются ИИ-агентами, то у программистов — это уже 95%. При этом adoption пока неравномерный, и сейчас идёт работа над тем, чтобы он становился более равномерным», — Дмитрий Иванов, «Яндекс»

Речь и документы — две опорные пары модальностей

В российской выборке мультимодальный ИИ в промышленной эксплуатации — это работа с двумя парами данных, а не с одной.

Первая — **речь плюс текст**: распознавание речи на входе, синтез речи или LLM-генерация текста на выходе.

Вторая — **документ плюс текст**: распознавание сканов, форм и первичных документов с последующей обработкой извлечённого содержания.

«OCR – распознавание и извлечение финансовых показателей из первичной финансовой отчётности компаний и банков; speech-to-text – распознавание эфиров передач и репортажей штатных корреспондентов... Извлечение финансовых показателей из первичной отчётности: здесь распознавание изображения (OCR) и последующая текстовая интерпретация работают как единый конвейер», - Мусаев Замир, Директор центра информационных технологий Новостные продукты Группы «Интерфакс»

Публичные кейсы дают больше материала по первой паре, поскольку она сосредоточена в контакт-центрах крупных компаний, охотно рассказывающих о своих результатах. Интервью с заказчиками показывают, что вторая пара распространена не меньше, а по числу компаний, вероятно, шире:

«Наш бухгалтер с искусственным интеллектом — одна модель распознаёт сканы», — Ярослав Черный, «Арнест Юнирусь»

«Там, где нужно распарсить PDF, хорошо справляется open source парсер», — представитель торгово-производственной компании

Различие между двумя парами не только техническое, но и рыночное. Речевые сценарии требуют платформы промышленного класса и потому концентрируются у крупных игроков с большим потоком обращений. Сценарии работы с документами собираются из общедоступных компонентов силами внутренней команды — порог входа заметно ниже, и именно поэтому они встречаются в компаниях любого масштаба, редко попадая в публичное поле.

Ниже — речевая линия, по которой в России накоплено больше всего публично подтверждённых результатов.

«**Ростелеком Контакт-центр**» выстроил наиболее масштабную в российской выборке мультимодальную систему обработки обращений: стек интегрирует аудио на входе, текстовую обработку через LLM и синтез речи на выходе. За год точность распознавания выросла вдвое; более 15% клиентов оценивают качество синтезированной речи как сопоставимое с живым оператором ([Ростелеком](#)). Платформа «Нейрошлюз», объединяющая более 20 ИИ-сервисов, превратила контакт-центр из традиционной операционной функции в технологически насыщенный узел обработки данных. [[IXBT](#)]

SaluteSpeech решает схожую задачу на уровне инфраструктурной платформы: синтез и распознавание речи, создание голосовых агентов, аналитика разговоров ([Sber Developers](#)). Интеграция речевых технологий с моделями позволяет строить гибридные решения, где LLM генерирует содержательную часть ответа, а синтез преобразует её в речь. Эта архитектура стала фактическим стандартом для голосовых агентов в клиентском сервисе.

Речевые технологии проникают и в нетипичные сценарии. **Магнит** в рамках производственного пилота запустил систему «Голос»: операторы сообщают о неисправностях оборудования голосом без ручного ввода, а LLM структурирует и категоризирует обращения ([CNews](#)). **Яндекс Маркет** добавил голосовой режим к торговому ИИ-агенту, причём голосовой запрос обрабатывается совместно с текстовым контекстом диалога ([Яндекс](#)).

Отдельного упоминания заслуживает сценарий, который давно вышел за пределы контакт-центров и стал самой распространённой точкой входа в мультимодальный ИИ для компаний любого масштаба:

«Ещё один кейс — транскрибация и суммаризация переговоров, встреч. Это тоже работает очень хорошо... Практический кейс, который здесь на поверхности, используется очень активно», — представитель фармацевтической компании

«Базовая функция — распознавание, суммаризация совещаний. Кажется, что все», — представитель агропромышленного холдинга

Вторая реплика — «кажется, что все» — заслуживает внимания как индикатор. Транскрибация встреч оказалась тем сценарием, который не требует ни интеграции с корпоративными системами, ни работы с чувствительными данными, ни согласований, и потому распространился практически повсеместно. Одновременно это и предел: дальше начинаются процессы, где нужны данные, интеграции и ответственность, и там продвижение резко замедляется.

Визуальная мультимодальность — активный фронт пилотов

Визуальная мультимодальность, то есть работа с неструктурированными изображениями и видео, а не с документами, чей разбор описан выше как уже сложившаяся практика, переживает в России фазу интенсивного пилотирования. Кейсы разнообразны по отраслям, что говорит о широте интереса, однако большинство остаётся на стадии доказательства концепции или ограниченного запуска.

Наиболее коммерчески зрелые примеры сосредоточены в электронной торговле.

Wildberries реализовал мультимодальный поиск на стеке CLIP + LLM + VLM: пользователь может искать товар, одновременно загрузив фотографию и добавив текстовое уточнение ([Habr](#)); позже запущен поиск по рисункам с возможностью дорисовки и изменения цвета элементов ([CNews](#)), а также виртуальная примерочная, комбинирующая фото товара и фото пользователя ([iXBT](#)).

Яндекс Маркет представил мультиформатного торгового агента, понимающего одновременно текст, изображение и контекст диалога ([Forbes](#)).

За пределами ритейла визуальная мультимодальность реализуется в более специализированных нишах. **Smart Engines** представила мультимодальную модель верификации документов, работающую одновременно в трёх диапазонах спектра — оптическом, ультрафиолетовом и инфракрасном — и анализирующую свыше 600 параметров, включая RFID-метки, штрихкоды, текстовые поля и метаданные; предыдущая версия уже эксплуатируется в ФНС и на автогейтах аэропорта Шереметьево ([ICT.Moscow](#)). **Магнит** тестирует видеоаналитику в теплицах, выявляя аномалии роста растений и ранние стадии заболеваний ([CNews](#)). **МГУ** разрабатывает мультимодальную платформу «ИИ-водитель», интегрирующую видеопотоки, данные навигационных систем и сенсоров ([ТАСС](#)).

Визуальные модели технически сложнее речевых, требуют более крупных обучающих выборок на русскоязычных и специализированных данных и предъявляют более высокие требования к вычислительной инфраструктуре. Тем не менее, динамика пилотов указывает, что визуальная мультимодальность находится сегодня примерно там же, где речевая была два–три года назад: технология работает, первые результаты получены, следующий шаг — промышленное масштабирование.

Интервью подтверждают и интерес, и характерные ограничения:

«Есть пилотное решение по агроскаутингу: дроны облетают поля, делают фотографии, фотографии потом сверяют. Здесь у вас сорняки, здесь у вас такая-то фаза роста растений, здесь у вас такая-то скученность и прочее. Дают соответствующую аналитику», — представитель агропромышленного холдинга

Судя по формулировкам респондента, результат пилота получен, но ещё не встроено в процесс принятия решений. Ровно на этой границе застревает большинство визуальных пилотов — не потому, что модель плоха, а потому, что не определено, кто и как использует её вывод.

Промышленность и отрасли за пределами финансов и телекома

Если финансовый сектор и телеком освоили агентный ИИ вглубь, то большинство остальных отраслей представлены единичными кейсами — как правило, в одном конкретном процессе, а не как системное освоение технологии. При этом диапазон отраслей широк: от металлургии до здравоохранения, что само по себе свидетельствует о растущем интересе за пределами традиционных лидеров.

Промышленность — та область, где российские компании добиваются наиболее документированных результатов, и это объяснимо: концентрация крупных активов, значительные вычислительные ресурсы на производственных площадках и традиционно высокая инженерная культура позволяют выполнять сложные мультидисциплинарные проекты. Однако важна оговорка о характере внедрений. «Северсталь» запустила платформу создания ИИ-агентов без навыков программирования и провела корпоративный конкурс ИИ-решений для ежедневных задач сотрудников, собрав более четырёхсот заявок ([CNews](#)), но применяет генеративный ИИ к офисным и аналитическим задачам, а не к производственным операциям как таковым. Это типичная для отрасли картина: технология входит в контур поддержки, а не в контур управления производством.

В ритейле и производстве продуктов питания **Магнит** ведёт несколько параллельных пилотов: видеоаналитика в теплицах, голосовой ввод данных об отказах оборудования, анализ тендерной документации поставщиков ([CNews](#)). Результаты первых пилотов предполагалось тиражировать с десятка торговых точек на несколько сотен. Это один из немногих российских кейсов, где генеративный ИИ пробует войти непосредственно в процессы производства и цепочек поставок.

Здравоохранение представлено слабо и характерным образом. **К-Скай (Webiomed)** запустила сервис автоматической проверки медицинских документов на гибридном стеке NLP и LLM: документ и детализированный отчёт о его качестве формируются за считанные секунды; реализовано два региональных проекта и более десяти пилотов с другими субъектами РФ ([CNews](#)). Показательно, что российский генеративный ИИ в здравоохранении начинает не с клинической диагностики, а с документооборота — сценария, не требующего клинической валидации и регуляторного одобрения, а значит, с существенно более низким барьером входа. Транспорт пока остаётся в зоне академических разработок.

Фармацевтика представлена двумя принципиально разными по характеру кейсами. Первый — **Сбер Еаптека** — это адаптация зрелого клиентского сценария под требования фармрегулирования ([Фармвестник](#)). Второй иной по природе: совместный проект **Р-Фарм** и Сбер ИИ-лаборатории **LIREGEN** применяет

генеративный молекулярный ИИ непосредственно к задаче drug discovery в иммуноонкологии, сократив стадию поиска лекарственного кандидата примерно втрое: с трёх лет по традиционному маршруту до порядка одного года ([Фармвестник, ComNews](#)). Это первый публично заявленный российский кейс применения генеративного ИИ в разработке онкологических препаратов — прямой аналог того, что глобально делают Pfizer, Genentech и AstraZeneca. Разрыв в зрелости с мировым авангардом остаётся значительным, но сам факт появления такого проекта означает, что фармацевтика перестала быть белым пятном на российской карте.

Отсутствие в выборке кейсов из других секторов не обязательно означает, что внедрений там нет: далеко не все компании готовы заявлять о проектах до получения осязаемого результата. Подтверждение этому нашлось и в интервью. Девелоперская компания рассказала о непубличном пилоте со Сбером:

«Есть пилот со Сбером, на партнерских некоммерческих условиях. Касается классификации элементов BIM-моделей... Мы имеем возможность обрабатывать плюс-минус законченную, но еще не почищенную модель, чтобы модель сделала предсказания на основании данных, на которых научилась... Они используют для этого внутренние модели, которые обучают, и это работает хорошо», — представитель девелоперской компании

Ещё два непубличных примера из практики Сбера как интегратора расширяют отраслевой охват. Первый — горнодобывающая отрасль, где речь идёт не о точечной автоматизации, а о полной перестройке процесса:

«У нас есть такой клиент — больше года полноценно работает система по управлению карьером, по сути агентная система. Мы посчитали, сколько сократится издержек, насколько повысится выработка, — и эффект в разы больше, чем от выявленных потерь при обычной оптимизации процесса. Например, самосвал большегрузный — 2,5 миллиона долларов — постоянно останавливается и трогаются, это износ техники; когда переводишь процесс управления карьером на агентов и коренным образом трансформируешь его, эти потери просто уходят вместе со старым процессом... Там 200 человек, у которых есть план, но никто до деталей не описывает, сколько реально человек может сделать сверх этого плана. Эффективность растёт радикально — не на проценты, как при тюнинге рутинных кусков моделями, а кратно», — Дмитрий Трофимов, Сбер

Этот пример стоит выделить особо, поскольку он описывает единственный в российской выборке случай эффекта иного порядка. Разница между «процентами» и «кратно» здесь объясняется не качеством модели, а глубиной вмешательства: агенты не ускоряют существующий процесс, а заменяют его другим, в котором прежние потери отсутствуют как класс. Плата за такой эффект — необходимость трансформировать процесс целиком, в чему большинство организаций пока не готовы.

Второй пример — кибербезопасность, где агентный ИИ помогает не клиентам напрямую, а самим специалистам по защите информации ([«Лаборатория Касперского»](#)):

«У нас был кейс с Касперским — публичный, реализовывали в виде чат-агента поддержку: агент помогает в процессе реакции на инциденты. Система администрирует средства защиты, случается инцидент — вредоносный файл разворачивает протокол, который выражается в виде сложной длинной ссылки. Раньше инженер вручную разбирал, куда он перемещается, что запускает, откуда выгружается. Агент помогает разобрать это за секунды — дальше понятно, что делать: локализовать, отключить, удалить, переустановить», — Дмитрий Трофимов, Сбер

Здесь стоит обратить внимание на распределение ролей: агент выполняет разбор, а решение о действии остаётся за инженером. Это та же схема human-in-the-loop, что и в фармацевтическом клиентском сервисе. И она устойчиво воспроизводится всюду, где цена ошибки высока.

Государственное управление: пилот в Аппарате Правительства Российской Федерации

Государственный сектор традиционно достаточно консервативен в части применения новых технологий, а публичных сведений о внедрениях здесь заметно меньше, чем в коммерческих отраслях. Тем ценнее позиция, изложенная для настоящего исследования пресс-службой Правительства Российской Федерации:

«Технологии генеративного искусственного интеллекта постепенно переходят от экспериментального применения к использованию в реальных процессах в бизнесе и в рамках государственного управления. Наиболее востребованные сценарии: обработка и анализ информации, подготовка документов, поиск сведений, формирование аналитических материалов и автоматизация рутинных операций.

Опыт проведения пилотного эксперимента по внедрению ИИ в Аппарате Правительства Российской Федерации подтвердил практическую применимость таких технологий для текстовых, аналитических, правовых и коммуникационных задач. Одновременно пилот показал, что для масштабирования важны не только возможности самих моделей, но и защищённая инфраструктура, интеграция с действующими информационными системами и подготовка сотрудников.

ИИ рассматривается прежде всего как инструмент поддержки специалиста, позволяющий сократить объём рутинной работы и повысить эффективность работы с информацией. При этом содержательные и управленческие решения, а также ответственность за них остаются за человеком.

Перспективными направлениями являются развитие ИИ-агентов и мультимодальных систем, способных работать с различными типами данных и

выполнять более сложные последовательные операций. По мере масштабирования будет также возрастать значение объективной оценки эффекта, развития компетенций сотрудников и роли интеграторов, обеспечивающих встраивание ИИ в существующие процессы и информационные системы.

Основной тенденцией является переход к системному применению искусственного интеллекта в отраслях экономики, социальной сферы и государственном управлении с обязательными требованиями к практической эффективности внедряемых решений.»

Обратим внимание, что приведённые формулировки совпадают с наблюдениями, полученными в интервью с корпоративными заказчиками, по нескольким пунктам сразу: набор наиболее востребованных сценариев тот же, что и в бизнесе; условием масштабирования названы инфраструктура, интеграция с действующими системами и подготовка сотрудников, а не качество моделей; ответственность за решения сохраняется за человеком; и, наконец, отдельно выделено требование объективной оценки эффекта. Это подтверждает, что описанная в обзоре повестка не является спецификой коммерческого сектора.

«Роль государства – задать чёткие правила игры, радикально снизить юридические риски использования ИИ в информационных сценариях, не связанных с прямым причинением физического вреда или с осуществлением намеренных информационных атак... Государство должно поощрять развитие отрасли, давая большую свободу и пространство для предпринимательского духа; возможно, через спонсирование инициатив и строительство инфраструктуры; регулирование и институционализация закономерно придут позже, когда рынок и практики созреют», - Мусаев Замир, Директор центра информационных технологий Новостные продукты Группы «Интерфакс»

Что мешает: барьеры глазами российских заказчиков

Полезную рамку для всего дальнейшего изложения дал один из экспертов-консультантов, разложив препятствия по трём категориям:

«Если разделить барьеры на пути ИИ-трансформации на группы: первая категория носит культурно-управленческий характер, вторая — методологический, ну и третья — технологический, точнее технологическо-архитектурный», — Лариса Малькова, Ахепix

Для этого высказывания довольно характерным является то, что технологические барьеры в нем названы последними. Это соответствует и остальному материалу интервью: собственно к моделям претензии у заказчиков есть, но при разборе конкретных проектов они, как правило, оказываются следствием состояния данных

и процессов. Поэтому в разделе о моделях говорится сравнительно немного, а там, где технологические ограничения действительно самостоятельны, — в мультимодальных сценариях и в безопасности агентных систем — они разобраны отдельно.

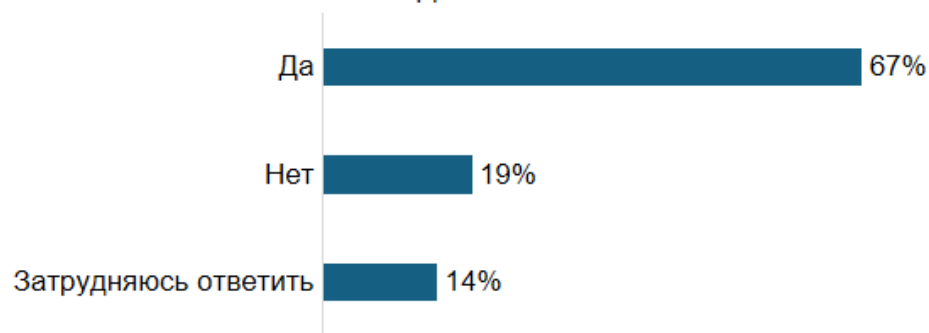
Особенность нынешнего момента заключается в том, что с большинством барьеров и проблем, связанных с внедрением генеративного ИИ, ни ИТ-сообщество, ни организации раньше не сталкивались, и практики их преодоления только формируются, как в России, так и за её пределами. Косвенно — не только в части барьеров — это подтверждается еще одним высказыванием эксперта:

«Ещё не выработаны фреймворки. Вот, допустим, в Accenture у нас был High Performance Banking — это процессы уровня 0-4, под которыми лежат и типовые КПЭ и референтная архитектура и много еще чего, которые говорят: если вы хотите быть успешным банком, у вас должно быть это, это и это... Такие фреймворки, подходы, best practices в отношении построения автономной компании, где ИИ — новая операционная система — они сейчас только вырабатываются», — Лариса Малькова, Axenix

Ловушка proof-of-concept

Тема повторяющихся циклов пилотов без выхода в промышленную эксплуатацию — одна из немногих, где публичная риторика и практика совпадают. По данным опроса, более двух третей респондентов сталкивались с ситуацией, когда ИИ-проекты в их организациях не доходили до устойчивого продуктивного состояния. Обычно эта цифра подаётся как признак незрелости рынка.

Были ли в вашей практике проекты внедрения генеративного ИИ, которые не доходили до продуктивной стадии?



Данные опроса «Были ли в вашей практике проекты внедрения генеративного ИИ, которые не доходили до продуктивной стадии?»

Интервью, в свою очередь, показывают, что высокий отсев пилотов далеко не всегда является провалом. В зрелых организациях он оказывается сознательно спроектированным свойством процесса:

«Мы разработали методологию того, куда должна двигаться архитектура по искусственному интеллекту в компании... Там прямо прописано, что если у нас 90% пилотов никуда дальше не пойдут — это нормально, это здоровая история. Самая главная история — это не двинуть в продуктив то, что никогда в жизни не окупится», — Ярослав Черный, «Арнест Юнирусь»

Это принципиально иной подход. Пилот здесь — не попытка внедрения, которая может не удасться, а инструмент дешёвой проверки гипотезы, и его основная ценность в том, чтобы вовремя отказаться от проекта, если его дальнейшее развитие экономически нецелесообразно. При такой постановке 90% отсева — показатель работающего фильтра, а не сломанного процесса.

Однако у высокого отсева есть и вторая природа, не имеющая отношения к экономике:

«Департамент IT довёл пилот до финального шага, мы понимаем, что это летит, но пока не получили согласования от совета директоров на применение данной технологии», — представитель фармацевтической компании

Здесь пилот успешен технически и, по-видимому, экономически — и всё равно не переходит в продуктив. Барьер лежит в управленческом контуре: решение о промышленном применении требует уровня согласования, к которому организация не готова, поскольку ответственность за работу автономной системы не распределена.

Есть и третья трактовка, переопределяющая само назначение пилота. Со стороны консультантов главным результатом называют не техническую демонстрацию, а готовность владельца бизнес-функции принять новый порядок работы:

«Мы сделали так называемый proof of concept, убедились, что идея реализуема. Но тут вопрос даже не в том, что это реализуемо, а в том, что владельцы соответствующих бизнес-функций не всегда к этому готовы ментально и морально. Потому что получается, что мы не приходим и автоматизируем процесс как есть, а мы этот процесс переосмыслием вообще, переворачиваем его и превращаем уже во что-то другое — с сильно меньшим количеством людей, которые в этом задействованы, и с гораздо более серьёзным уровнем автономии», — Лариса Малькова, Axenix

Если принять эту рамку, то пилот, доказавший работоспособность технологии, но не изменивший позицию владельца процесса, следует считать незавершённым, а не успешным. И тогда описанная выше ситуация с отсутствием согласования советом

директоров выглядит не бюрократической заминкой, а закономерным итогом пилота, в котором работа с готовностью заказчика просто не была запланирована.

К этим трём трактовкам поставщики добавляют четвёртую, смещающую проблему с отдельного пилота на переход к промышленному масштабу:

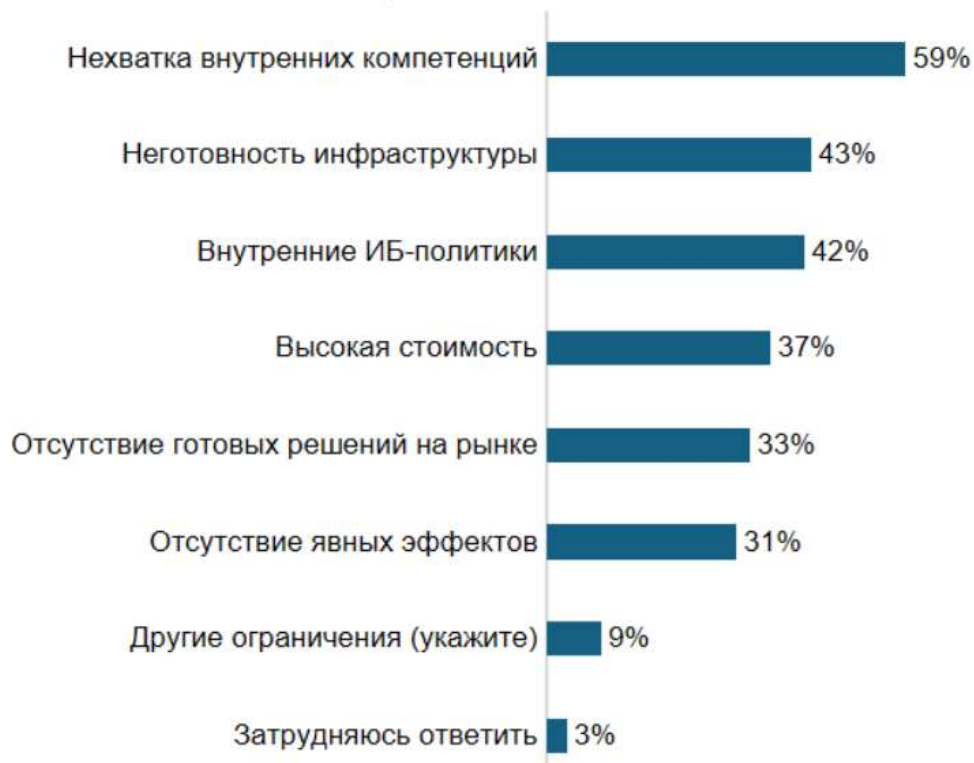
«Пилоты в этом году плюс-минус завершились. Все компании уже поняли, что в принципе ИИ работает, теперь они думают, как это внедрять в промышленных масштабах. У всех болит именно промышленное внедрение, качество и эффекты на этом уровне», — Артур Самигуллин, «Яндекс»

В этой рамке застревание на стадии пилота объясняется не только слабостью самих пилотов или неготовностью владельца процесса, но и отсутствием отраслевой методологии промышленной эксплуатации — её пока нет ни у заказчиков, ни у поставщиков.

Корневые барьеры: зрелость данных и процессов, компетенции в ИИ

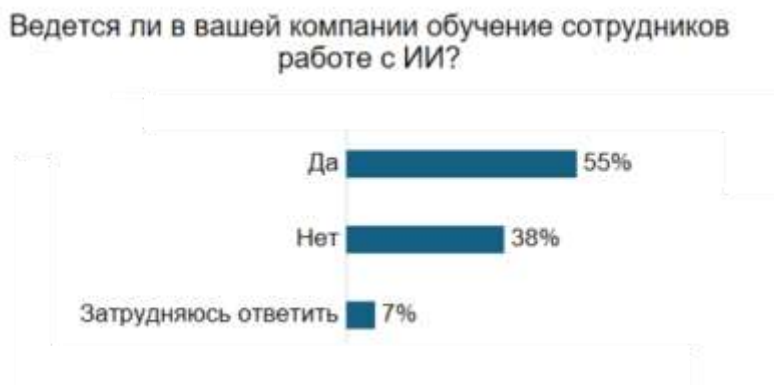
Почти две трети российских компаний (59%) называют дефицит нужных компетенций главным препятствием для внедрения ИИ — показатель даже выше глобального. Из этого утверждения напрашивается типовой вывод: нужно срочно обучать сотрудников работе с ИИ.

Какие ограничения вы видите в распространении и масштабировании технологий генеративного ИИ в рамках вашей организации?



Данные опроса «Какие ограничения вы видите в распространении и масштабировании технологий генеративного ИИ в рамках вашей организации?»

При этом обучение сотрудников работе с ИИ ведётся в 55% компаний, не ведётся — в 38%. То есть, более трети компаний обучение не ведут вовсе. Картина получается замкнутой: компетенций не хватает, их дефицит осознан, и при этом заметная часть рынка ничего не делает для их восполнения.



Данные опроса «Ведётся ли в вашей компании обучение сотрудников работе с ИИ?»

Интервью существенно дополняют эту картину. Компаниям зачастую не хватает не понимания ИИ как такового, а понимания собственных процессов и собственных данных.

«Разработчики и бизнес в момент реализации не заполняли описания сущностей, таблиц, мер — а для LLM важен контекст», — представитель торгово-производственной компании

Корпоративные базы знаний накапливались годами без ревизии противоречий, а корпоративные данные создавались в расчёте на то, что смысл полей известен разработчикам и не нуждается в описании. В традиционных информационных системах недостаток контекста частично компенсирует пользователь за счёт собственного опыта и знания процессов. Языковая модель такой возможности не имеет и опирается на доступные ей данные и их описание. В результате качество ответа определяется не характеристиками модели, а состоянием документации, которому в компаниях зачастую не уделяется достаточного внимания.

Наиболее радикальную формулировку той же проблемы дал эксперт со стороны интегратора:

«Насколько описанный, детально прописанный процесс... покрывает фактический процесс? В реальности процентов на 10–20, в лучшем случае. В

когда нужно что-то решить — делать или не делать — собирается куча людей... и делает миллиард шагов», — Дмитрий Трофимов, Сбер

То есть, формализованное описание процесса — то, на что опирается любая попытка автоматизации, — покрывает лишь малую долю того, что происходит в действительности. Остальное существует в виде неявных договорённостей, обходных путей и коллективных согласований, которые нигде не зафиксированы. Агент, спроектированный по документации, автоматизирует эти 10–20% и оказывается бесполезен ровно в тех ситуациях, ради которых его создавали.

Проблему усугубляет сама суть генеративного ИИ — готовность работы с нечеткими и даже противоречивыми данными. Прежние волны автоматизации хотя бы вынуждали компанию описать процесс — иначе систему было не настроить, и формализация происходила как побочный результат внедрения. Языковая модель такого принуждения не создаёт: она способна работать поверх неопisanного процесса, выдавая правдоподобные ответы. Именно поэтому неупорядоченность здесь дольше остаётся незамеченной и обнаруживается не на старте, а при попытке масштабирования.

Практический вывод может быть сформулирован следующим образом: инвентаризация процессов и данных является не подготовительной фазой ИИ-проекта, а его содержанием. Проект, стартующий без ответа на вопросы «описан ли процесс», «где лежат данные и что означают их поля», «кто владелец процесса», с высокой вероятностью упрётся в эти вопросы позже — когда бюджет уже израсходован, а сроки пройдены.

Экономика внедрения и разрыв между обещанным и подтверждённым эффектом

В публичном поле экономические результаты внедрения генеративного ИИ выглядят убедительно: на отраслевых конференциях и в материалах вендоров регулярно приводятся кейсы с впечатляющими показателями экономии. На их основе у рынка может складываться представление о том, какой эффект от внедрения технологии можно считать типичным.

Однако заказчики относятся к этим цифрам скептически — и объясняют, почему:

«...когда делаешь полноценную оценку с вендором, который это решение предлагает, оказывается, что эти два человека должны быть заменены двумя экспертами по ИИ, которые явно дороже — они нужны, чтобы дообучать модель, обновлять базу данных, проверять качество знаний... ты просто перекладываешь [затраты на] ресурс из одного департамента в другой», — представитель фармацевтической компании

Второй респондент описывает ту же ситуацию со стороны публичных кейсов:

«...ИИ сам по себе не приносит денег (кроме продавцов токенов или GPU)... Если вам скажут любой кейс на конференции — особенно если это было год назад — мы автоматизировали при помощи ИИ, у нас экономят 30 миллионов в год, а начинаешь задавать вопросы — там уже не 30, и уже не экономия, а расход», — представитель ритейлера спортивных товаров

Оговорка «особенно если это было год назад» существенна: она указывает, что рынок постепенно взрослеет и качество публичных заявлений повышается. Но накопленный скепсис никуда не делся и влияет на решения сегодня.

Третья реплика переводит проблему из плоскости расчётов в плоскость ответственности:

«...никто ответственность за результат взять не может. Если бы нам кто-то взял конкретную ответственность за решение конкретной бизнес-проблемы — то да, через столько-то времени вы получите ровно то, что вам требуется... Пока что в ИИ такого нет, пока только такие обещания воздушные», — представитель агропромышленного холдинга

Здесь сформулировано ключевое отличие ИИ-проектов от привычных ИТ-внедрений. При покупке традиционной системы поставщик отвечает за функциональность: система либо делает то, что записано в спецификации, либо нет. В ИИ-проекте результат носит вероятностный характер, зависит от данных заказчика и потому не поддаётся такой фиксации. Поставщики результат не гарантируют — и не могут гарантировать, а заказчики воспринимают отсутствие гарантий как признак незрелости предложения.

Дополнительная сложность связана с выбором самих показателей эффективности: вопрос заключается не только в величине достигнутого эффекта, но и в том, как именно его измерять. Даже в разработке ПО — сценарии с наиболее очевидным эффектом — общепринятой метрики не сложилось:

«Возьмём разработку программного обеспечения — одно из тех направлений, где импакт генеративного ИИ наиболее ярко проявлен... Но тоже возникает вопрос: а в чём меряются эти эффекты? В количестве разработанного кода? Немножко странно. В количестве реализованных требований бизнеса, выраженных в запросах/брифах/user stories в единицу времени (допустим квартал)? Такие метрики и сама экономика, включая токеномику — такая методология только формируется, и даже чемпионы такой вопрос ставят: а как это всё-таки правильно посчитать?», — Лариса Малькова, Axenix

При этом, нельзя не отметить и то, что некоторые российские компании уже нащупывают для себя подходящие методики оценки (хотя бы прогнозных) эффектов:

«Экономика проектов оценивается через юнит-экономику: влияние на маржинальность и возврат инвестиций по совокупной стоимости внедрения (ТСО); рассчитывается прогнозный LTV клиента (Predictive LTV); приоритет отдаётся проектам, которые влияют на потребительское качество продуктов и повышают их ценность для клиентов», - Мусаев Замир, Директор центра информационных технологий Новостные продукты Группы «Интерфакс»

Инфраструктура, безопасность и вопрос о качестве моделей

Среди ограничений в распространении и масштабировании генеративного ИИ неготовность инфраструктуры (43%) и внутренние ИБ-политики (42%) занимают второе и третье места — выше высокой стоимости (33%) и заметно выше отсутствия явных эффектов (31%).

Тот же сюжет виден и в ответах на второй вопрос анкеты — о причинах, по которым конкретные проекты не доходят до продуктивной эксплуатации. Здесь противоречие с политиками информационной безопасности названо в 37% случаев — чаще, чем отсутствие внутренних компетенций (33%), а отсутствие инфраструктуры — в 26%.

Этот вопрос задавался с более широким списком вариантов, и картина в нём выглядит иначе, чем в вопросе об ограничениях масштабирования. Первые три позиции занимают несовершенство моделей (52%), низкое качество данных (48%) и сложности интеграции с корпоративным ландшафтом (46%). Различие объясняется прежде всего постановкой вопросов: об ограничениях масштабирования респондентов спрашивали применительно к организации в целом, и вариант, относящийся к качеству самих моделей, в том списке не предлагался; здесь же речь идёт о причинах неудачи конкретного проекта, и такой вариант в списке был.

По каким причинам проекты внедрения генеративного ИИ не доходят до продуктивной эксплуатации?



Данные опроса «По каким причинам проекты внедрения генеративного ИИ не доходят до продуктивной эксплуатации?»

Первая позиция этого списка заслуживает отдельного комментария, поскольку на первый взгляд диссонирует с некоторыми выводами обзора. Несовершенство моделей — это причина в том виде, в каком её называет сам заказчик, и определить по ответу в анкете, что именно стояло за неудачей, невозможно. Интервью показывают, что значительная часть таких случаев имеет иную природу: модель отвечает неудовлетворительно не потому, что слаба сама по себе, а потому, что получает на вход неописанные данные и неформализованный процесс. Проект, остановленный по этой причине, в опросе почти наверняка будет отнесён к несовершенству моделей — просто потому, что видимым симптомом становится качество ответа. Показательно и соседство строк: две следующие позиции — низкое качество данных (48%) и сложности интеграции с корпоративным ландшафтом (46%) — описывают ровно ту же ситуацию, но уже названы своими именами.

Это не означает, что технологические ограничения отсутствуют. Они реальны, но, как показано ниже, сосредоточены в сравнительно узкой области: в мультимодальных сценариях — распознавании неструктурированных изображений и качестве открытых моделей синтеза речи, — а также в новых, ещё не изученных векторах атак на агентные системы. В типовом же текстовом корпоративном

сценарии решающим фактором оказывается не выбор модели, а состояние данных, процессов и контура, в который решение встраивается.

Отдельный интерес представляет анализ ответов о причинах в привязке к размеру компании. Так, в группе малых компаний «несовершенство моделей» как основную причину «застревания» проектов называют 62% респондентов, в то время как в группе больших компаний - всего 42%. В отношении ответа «Низкое качество данных» ситуация диаметрально противоположная - всего 31% респондентов из малых компаний считает эту причину значимой. При этом респонденты из крупных компаний указали эту причину в 83% случаев.



Это может объясняться тем, что компании малого размера пока что используют ГенИИ в основном для универсальных задач, где результат почти целиком зависит от самой модели (разработка ПО, аналитика и отчёты, маркетинг и контент, ИИ-ассистенты общего назначения). Крупные же компании все чаще идут дальше и применяют ГенИИ на своих корпоративных данных (документооборот, поиск по корпоративным данным (RAG); внутренняя ИТ-поддержка; HR-процессы). То есть, по мере роста зрелости корпоративных сценариев применения качество моделей становится все менее значимым фактором.

Вернёмся к инфраструктуре и требованиям безопасности. Интервью объясняют, почему эти два барьера стоят рядом и почему в России они тяжелее, чем на мировом рынке. Требования безопасности «заталкивают» проект в закрытый контур предприятия, а закрытый контур немедленно превращается в инфраструктурную задачу: нужны собственные серверы, ускорители и компетенции по их эксплуатации.

«Три препятствия: хорошие данные и трансформация процессов, окупаемость, которая никак не срастается, и безопасность, которая тоже пока непонятна», — Ярослав Черный, «Арнест Юнирусь»

Показательно, что безопасность здесь названа не в терминах политик и регламентов, а как область непредсказуемости — то же ощущение прозвучит ниже, в разговоре о неизвестных векторах атаки на агентные системы. Для агентов оно обосновано: их доступ к корпоративным системам создает новые классы рисков и расширяет поверхность атаки, а устоявшиеся механизмы контроля таких агентов пока отсутствуют.

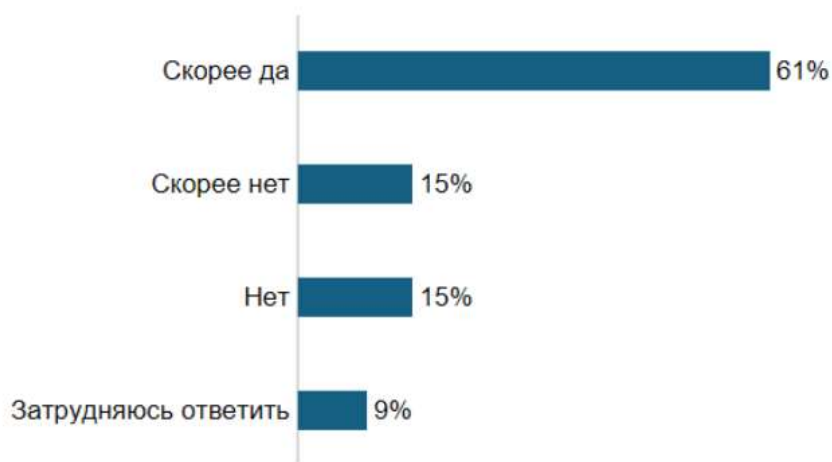
Инфраструктурная сторона того же барьера подробно разобрана выше, в разделе о структурных особенностях рынка: приоритет изолированного развёртывания удлинит цикл запуска, повышает порог входа и превращает пилот в инвестиционный проект. Данные количественного опроса показывают, на каком этапе этот барьер ощущается острее всего: неготовность инфраструктуры называют 51% компаний на стадии пилотов и лишь 21% компаний, активно использующих генеративный ИИ. Главный инфраструктурный порог, таким образом, приходится на вход, и компании, преодолевшие его, упоминают эту проблему заметно реже. Однако интервью предостерегают от того, чтобы считать вопрос закрытым после пилота: инфраструктура, достаточная для пилота на одной функции, может оказаться недостаточной, как только решение нужно распространить на организацию.

Практический вывод в этом вопросе может касаться последовательности шагов проекта. Требования информационной безопасности и архитектуру развёртывания следует согласовывать до старта пилота, а не после получения результата: именно на этом стыке останавливается заметная часть технически успешных проектов. Полезно заранее ответить на три вопроса: какие данные допустимо выпускать за периметр, какая конфигурация развёртывания приемлема для службы безопасности и хватит ли имеющихся вычислительных мощностей не для пилота, а для тиража. Проект, стартовавший без этих ответов, с высокой вероятностью упрётся в них на более поздней стадии, когда цена любых изменений будет гораздо выше.

Теневой ИИ: практика, требующая легализации, а не запрета

Использование сотрудниками личных аккаунтов в публичных моделях для рабочих задач обычно обсуждается в терминах риска утечки и нарушения политик — то есть как проблема информационной безопасности, требующая контроля и ограничений. Однако такой взгляд явно опровергается по результатам интервью с заказчиками: практически все компании осознают факт использования сотрудниками личного доступа к ИИ для решения рабочих задач и воспринимают его как норму. Это подтверждают и данные опроса:

Присутствует ли в вашей компании практика “теневого” использования ИИ?



Данные опроса «Присутствует ли в вашей компании практика „теневого” использования ИИ?»

Второй, более неожиданный результат опроса, — почти никто из собеседников не рассматривает запретительные меры как рабочий инструмент.

«...если человек захочет отнести данные вне корпоративной подписки, он это сделает. Просвещение важнее, чем наказание и попытка поймать все данные. В компании сейчас довольно взвешенная позиция: от того, что Claude узнал стоимость нашего проекта, мир не перевернется», — представитель деvelopepской компании

В реплике два тезиса, и оба заслуживают внимания. Первый — техническая невозможность полного контроля: у сотрудника всегда есть личное устройство, и запрет на корпоративном контуре просто выталкивает использование за его пределы, где оно становится невидимым. Второй — дифференциация чувствительности данных: не всякая корпоративная информация одинаково критична, и попытка защищать всё одинаково жёстко приводит к тому, что не защищается ничего.

В то время как служба безопасности видит в теновом ИИ нарушение политик, практики видят в нём рыночный сигнал: если сотрудники массово обходят запреты, значит, спрос велик, а легального предложения внутри компании нет. Это две разные интерпретации одного факта, ведущие к противоположным решениям.

Дальше - позиции, в которых теневое использование прямо переопределяется как ресурс:

«Теневой ИИ, конечно, есть. Мы стараемся, чтобы он был, наоборот, открытый, дать его всем. Но есть предостережения от руководства, что не надо с какими-то очень чувствительными данными куда-то вовне лезть. Для работы с чувствительными данными у нас есть внутренний контур», — представитель ритейлера спортивных товаров

«Нормально. У нас так и есть — хорошо, когда сами ставят эксперименты, находят инсайты... С нашей стороны ограничений нет: если сделают сами, потом заберём на централизованное управление», — представитель торгово-производственной компании

Последняя формулировка описывает целую модель внедрения, и она заслуживает того, чтобы быть названной. Сотрудники самостоятельно находят полезные сценарии, доводят их до работающего состояния на доступных им инструментах, а компания затем забирает удачные решения в управляемый контур. Такой путь бесплатен на стадии поиска, отбирает сценарии по факту востребованности, а не по предположению, и снимает с ИТ-подразделения необходимость угадывать потребности бизнеса. Издержка — период, в течение которого использование «теневого ИИ» остаётся неконтролируемым.

«Наш подход гибкий и стимулирует, а не подавляет внедрение новых технологий. Внутренние политики описывают преимущественно ограничения „куда нельзя“, пока конкретная технология или подход официально не регламентированы. После пилотирования на ограниченном круге задач компания охотно адаптирует технологию и переводит её в официальный контур – то есть теневое использование зачастую становится точкой входа для будущего официального внедрения, а не тем, с чем борются на подавление», - Мусаев Замир, Директор центра информационных технологий Новостные продукты Группы «Интерфакс»

Дефицит управления: агенты появляются раньше, чем модель управления ими

Незрелость управления ИИ обсуждается преимущественно как соответствие требованиям: политики, документация, оценка рисков, аудит. Российская ситуация в этом отношении близка к глобальной, а по ряду экспертных оценок разрыв между декларируемыми планами внедрения и реальной готовностью управлять рисками в нашей стране даже больше, чем на Западе.

Масштаб проблемы в мире уже измерен, и цифры стоит привести: они задают представление о том, к чему российские компании придут при сохранении нынешних темпов. Распространение агентов привело к росту числа «нечеловеческих» учётных записей (non-human identities) в корпоративных системах на 76%: агентам требуется

привилегированный доступ к базам данных, API и хранилищам документов, а традиционные практики управления доступом, созданные под пользователей-людей, с этим справляются плохо. Средний уровень зрелости управления ответственным ИИ — автоматизированная оценка рисков, governance-платформы, человеческий надзор — McKinsey оценивает в 2,3 балла по шкале 0–4, порог в 3 балла преодолела лишь треть организаций; при этом именно уровень инвестиций в управление коррелирует с бизнес-результатом.

В интервью с российскими заказчиками проблема выглядит приземлённее и проявляется на более раннем этапе. Речь идёт не о соответствии требованиям, а о том, что в компании появляется набор разрозненных решений, не связанных общей архитектурой:

«...мы понимаем, что здесь должен быть подход с нормальной платформой, а не разрозненный салат из агентов. Поэтому мы на условной паузе сейчас», — представитель девелоперской компании

К этому примыкает уже упоминавшаяся терминологическая проблема: когда «агентом» в компании называют и сценарного чат-бота, и автономную многошаговую систему, невозможно ни вести реестр, ни оценивать риски, ни планировать ресурсы, поскольку объекты управления не определены.

У этой позиции — отставания от мировой статистики по стадии распространения агентов — есть преимущество, которым можно воспользоваться. Компания, у которой пока три–четыре агента, может выстроить единый контур — каталог решений, владельцев, правила доступа и логирование — существенно дешевле, чем та, которой придётся наводить порядок в нескольких десятках разрозненных внедрений. Отставание в данном случае даёт возможность не повторять чужие ошибки.

Что ограничивает агентные мультимодальные системы

Всё изложенное выше относится к генеративному ИИ в целом. Но у агентных и мультимодальных систем есть собственный набор ограничений, не сводимый к общим проблемам с данными и экономикой.

Агенты-ассистенты имеют явный потолок эффективности. Прежде чем говорить об отдельных барьерах, полезно зафиксировать рамку, которую предложил один из экспертов-консультантов: путь зрелости внедрения агента в организации описывается как лестница, где первая ступень — работа в режиме ассистента, вторая — полуавтономная, третья — автономный конвейер. Различие между ступенями не количественное, а качественное:

«Первый этап — это ассистированная работа. То есть ты не передаёшь свои процессы, свои задачи, операции полностью в автономное плавание, а процесс у

тебя выглядит так же, как и был, просто у тебя есть помощник, какой-то ускоритель. На этом этапе, несмотря на то, что выигрыш в определённых задачах, по крайней мере в скорости их реализации, может быть вообще в разы, в любом случае, если брать любой процесс end-to-end, общий выигрыш будет процентов 10», — Лариса Малькова, Axenix

Эта оценка объясняет наблюдение, которое иначе выглядело бы парадоксом: отдельные операции ускоряются кратно, а сквозной эффект по процессу остаётся скромным. Причина в том, что сам процесс не меняется — ускоряются лишь его отдельные шаги, тогда как передачи, согласования и ожидания между ними сохраняются. Нелинейный эффект возникает только со сменой ступени:

«Какую функцию ни взять, выигрыш и эффект будет нелинейный, если у тебя существенная часть процесса перейдёт в автономное плавание. И то, что называется human in the loop, превратится в human on the loop, то есть над процессом», — Лариса Малькова, Axenix

Различие между «человеком в цикле» (human in the loop) и «человеком над циклом» (human on the loop) стоит держать в уме при чтении всего дальнейшего: подавляющее большинство российских внедрений находится на первой ступени, и описанные ниже барьеры — это, по существу, то, что мешает перейти на вторую. Отсюда же следует и трезвая оценка ожиданий: если проект спроектирован как построение ассистента, рассчитывать на кратный эффект по процессу не стоит, каким бы впечатляющим ни было ускорение отдельных операций.

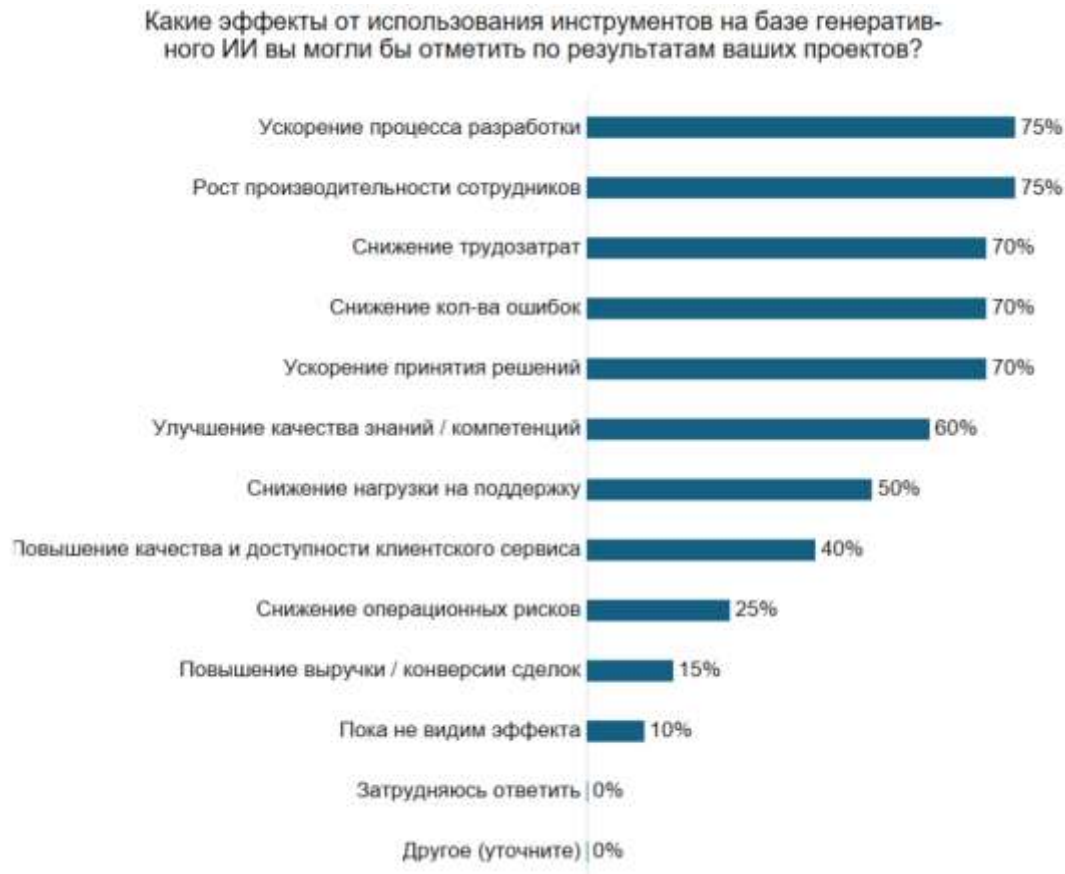
Показательно, что кейс, где эффект действительно оказался кратным, — уже описанное управление карьером в горнодобывающей отрасли — соответствует именно смене ступени: процесс не ускорили, а перестроили. Второй пример такого рода привели консультанты — мультиагентная проверка контрагентов:

«Процесс превращается в мультиагентский сценарий, когда действия происходят без участия человека. Например, анализ по всем источникам, потом несколько стадий проверки, и в итоге человеку на конечное решение выдаётся только суммирующая справка, причём со степенью уверенности: всё окей, степень уверенности 0,87. И при желании можно углубиться во всю трассировку... Этот исполнительский контур полностью уходит в автономное плавание, и остаётся только лицо, принимающее решение», — Лариса Малькова, Axenix

Здесь стоит обратить внимание на два элемента, делающие такую автономность приемлемой: явно выраженная степень уверенности системы в собственном выводе и сохранённая трассировка всех шагов. Именно они, а не качество модели, позволяют человеку остаться над процессом, а не внутри него.

Тезис об ограниченности эффектов ассистированного уровня зрелости косвенно подтверждается через профиль эффектов, которые компании отмечают по

результатам своих проектов. Верхние строки заняты ускорением и производительностью: ускорение процесса разработки (75%), рост производительности сотрудников (75%), снижение трудозатрат и снижение количества ошибок и ускорение принятия решений (по 70%). А вот повышение выручки или конверсии сделок отмечают лишь 15%.



Данные опроса «Какие эффекты от использования инструментов на базе генеративного ИИ вы могли бы отметить по результатам ваших проектов?»

Это ровно тот профиль, который описывает первая ступень лестницы зрелости. Операции внутри процесса ускоряются, и сотрудники это чувствуют, но устройство процесса не меняется, поэтому до верхнеуровневых финансовых показателей эффект не доходит. Пока картина эффектов выглядит так, ожидать кратной отдачи не следует: она появляется на следующей ступени, а до нее рынок пока не дошел.

Автономность агентов ограничивается сознательно. Практически везде автономность агента ограничена конкретными узкими шагами, а не сквозным выполнением задачи. Даже там, где формально работает агент, на входе и выходе цепочки почти всегда стоит человек.

«End-to-end процесс пока не выстроили... Сам процесс изучили, в середине агенты работают, но на входе и на выходе — люди», — представитель торгово-производственной компании

Это типичное решение: автономность не включается и не выключается целиком, а дозируется по величине последствий. Со стороны интегратора картина оценивается так же:

«Мы считаем, что наши агенты находятся на уровне L3 [автономности агентных систем], то есть выполняют определённую последовательность операций..., но требуют существенного контроля сотрудников. Агенты сами никуда не ходят, сами ничего не делают», — Дмитрий Филатов, «Сколково»

Дозированная автономность видна и в количественных данных. Больше половины компаний (55%) описывает своих агентов как дающих только подсказки и рекомендации, ещё 45% — как действующих по заранее заданным правилам. О полной автономии агентов или прохождении ими самостоятельно сложных цепочек действий заявили в сумме всего 5% опрошенных.



Данные опроса «Насколько автономными являются ваши ИИ-агенты?» (вопрос с множественным выбором)

Таким образом, при оценке экономики агентных проектов целесообразно исходить

из автоматизации отдельных этапов процесса при сохранении участия человека, а не из полной автоматизации всей цепочки. Расчеты, предполагающие полное исключение человека из процесса, пока не находят подтверждения в российской практике.

Агент — новый тип сотрудника, для которого нет модели прав и ответственности. Этот барьер находится на стыке организационной и технической проблематик: агент действует между функциями и потому оказывается «ничьим» и никому не подотчетным.

«Первое и главное — организационные вопросы и человеческий фактор: отсутствие владельца сквозного процесса — агент работает "между" функциями, и оказывается ничьим... Второе — безопасность и комплаенс: агент с доступом к системам — новый тип "сотрудника", для которого нет готовых моделей прав, аудита и ответственности», — Александр Вибе, «Сбер»

Для традиционной информационной системы вопрос прав решается через роль пользователя. Агент же — не пользователь и не сервис в привычном понимании. Он действует от чьего-то имени, но самостоятельно. Практическое проявление этой неопределённости описывает другой эксперт:

«Проблема — дать агенту разрешение отключить клиенту приложение на 25 минут. Поэтому есть смысл делать такую функциональность своими руками, если готового решения купить нельзя», — Дмитрий Трофимов, «Сбер»

В обобщённом виде этот барьер формулируется как противоречие, снять которое пока не удаётся ни организационно, ни юридически:

«Ты должен из своих рук какой-то процесс, набор операций, набор действий отдать, но продолжать нести за это ответственность... Самая ключевая задача наших пилотов, консалтинга как раз в этом — продумывание всего механизма, как это превратить в доверенную историю», — Лариса Малькова, Axenix

Отсюда же вытекает и вопрос ответственности, который заказчики формулируют прямо:

«В конечном итоге, когда [за процессом] стоит человек, он несёт ответственность, и кто-то отвечает за эту ошибку. На текущий момент, если это будет делать искусственный интеллект, вопрос — кто несёт ответственность», — представитель фармацевтической компании

«Никогда не доверим полностью: работу с закрытой информацией и персональными данными без надлежащих правовых оснований и согласований; самостоятельное принятие бизнес-решений; полную автономию без ограничения набора доступных инструментов и прав доступа...

Ответственность за итоговый результат всегда лежит на сотруднике, а весь сгенерированный контент подлежит обязательной проверке человеком», - Мусаев Замир, Директор центра информационных технологий Новостные продукты Группы «Интерфакс»

Добавляется и незрелость самой модели угроз:

«Мы даже, наверное, пока не совсем понимаем все вектора атаки, которые с нами могут случиться — настолько незрелая технология, что она может повести себя вообще непонятно как», — Ярослав Черный, «Арнест Юнирусь»

Опасение обоснованно: с переходом к агентным архитектурам поверхность атаки расширяется принципиально. OWASP (Open Worldwide Application Security Project) выделяет три ключевых вектора: захват цели, то есть подмена целевых инструкций агента; злоупотребление инструментами, когда агент вызывает API вне первоначального назначения; и злоупотребление привилегиями. Сохраняет актуальность и prompt injection — скрытые инструкции во внешних документах или на веб-страницах, переопределяющие поведение агента. Общее у всех трёх векторов то, что они не имеют аналогов в защите традиционных систем: атакуется не код и не канал передачи данных, а цель, которую агент считает своей.

Фокус внимания смещается с вопроса «насколько точна модель?» к вопросу «кто отвечает за ошибочные действия системы?» При этом установить причины конкретного действия агента зачастую можно только с помощью глубокого технического анализа, что существенно усложняет аудит его работы.

Интеграция с унаследованными системами. Мировая дискуссия описывает подключение агента к корпоративным системам как решённую задачу — необходимо использовать MCP и готовые коннекторы. Российская практика упирается в контур, для которого таких коннекторов нет:

«Интеграция с 1С — мы плотно сидим на 1С, а там API нет, нужно поднимать веб-сервисы... Систем много, каждое 1С-решение — отдельная конфигурация», — представитель торгово-производственной компании

Это прямое продолжение сюжета об отсутствии экосистемы агентных компонентов: там, где западный заказчик берёт готовый коннектор, российский разворачивает интеграционный проект — причём отдельный для каждой конфигурации.

Нужна не модель, а обязанка вокруг неё. Из сказанного выше следует вывод, что объектом внедрения следует считать не отдельную языковую модель, а платформу, обеспечивающую прослеживаемость действий, наблюдаемость системы и контроль стоимости вычислений.

«Это вопрос не только LLM, это вопрос действительно полноценной платформы и выстраивания всей обязанки, которая в профессиональной среде

получила терминологию harness... Появляется такое понятие, как «токеномика», потому что стоит все дорого, и если ты будешь, грубо говоря, забивать гвозди микроскопом, то это нецелесообразно. Значит, встает вопрос, как перераспределять нагрузку: если это более простая задача — направить её на более маленькую, более дешёвую модель», — Лариса Малькова, Axenix

Это замыкает тезис о коммодитизации моделей, появившийся в начале обзора. Если модели сблизилась по качеству и различие между ними для большинства задач не принципиально, то предметом инженерной работы становится всё остальное: маршрутизация запросов между моделями разного размера и стоимости, сбор трассировок, контроль качества ответов, управление правами. Именно эта обвязка, а не выбор модели, определяет и стоимость эксплуатации, и возможность отдать процесс агенту, сохранив за собой ответственность.

Ограничения мультимодальных сценариев. Здесь барьеры делятся на качественные и экономические. Качество распознавания неструктурированных изображений остаётся слабым местом практически у всех, кто пробовал такие сценарии:

«Хуже с изображениями... там, где нужно распарсить PDF, хорошо справляется open source парсер, но... нужно сверить подпись на скане с подписью в договоре. Здесь не смогли добиться хорошего результата», — представитель торгово-производственной компании

«Есть бот, который смотрит дизайн этикеток... и в половине случаев ошибается», — представитель торгово-производственной компании

Отдельная проблема — качество открытых моделей в речевых задачах, где разрыв с коммерческими решениями пока не закрыт:

«Модели синтеза голоса недостаточно хороши, чтобы вот так с коробки заменить решение от Яндекса... Опенсорсные модели доступные пока не дотягивают по качеству», — представитель ритейлера спортивных товаров

Более ресурсоемкие модели могут обеспечивать более высокое качество, но требуют больше времени на обработку запроса, тогда как для голосового взаимодействия критична скорость реакции. Поэтому в таких сценариях приходится использовать более компактные и быстрые модели, сознательно жертвуя частью качества рассуждения. Это ограничивает применимость агентных цепочек в голосе сильнее, чем в текстовых каналах.

Второй по частоте барьер — стоимость обработки на потоке:

«Обработка аудио и особенно видео в разы дороже текстовой; на больших потоках... стоимость становится главным тормозом», — Александр Вибе, «Сбер»

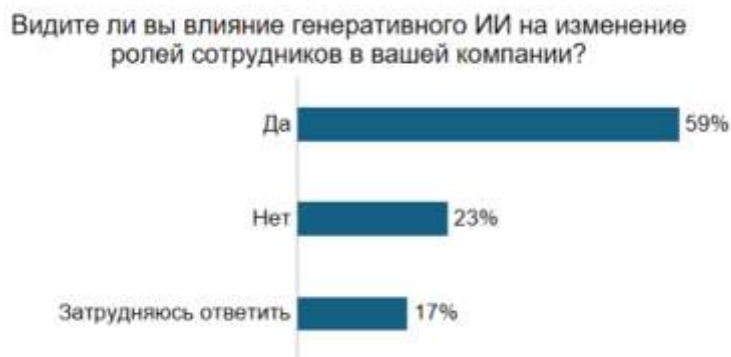
К этому добавляется ограничение, существенное прежде всего для регулируемых отраслей: инструменты интерпретируемости для мультимодальных моделей заметно отстают от инструментов, разработанных для текстовых. Понять, почему система пришла к конкретному выводу по изображению или аудиозаписи, сложнее — а значит, сложнее и обосновать принятое на этом основании решение перед регулятором или в спорной ситуации.

Сочетание перечисленных ограничений объясняет, почему российская мультимодальность закрепились именно в парах «речь плюс текст» и «документ плюс текст»: там качество достаточно, стоимость обработки приемлема, а результат проверяем. Видео и неструктурированные изображения остаются в пилотной зоне не из-за отсутствия интереса, а потому что не проходят сразу по нескольким параметрам.

Влияние генеративного ИИ на роли и компетенции сотрудников

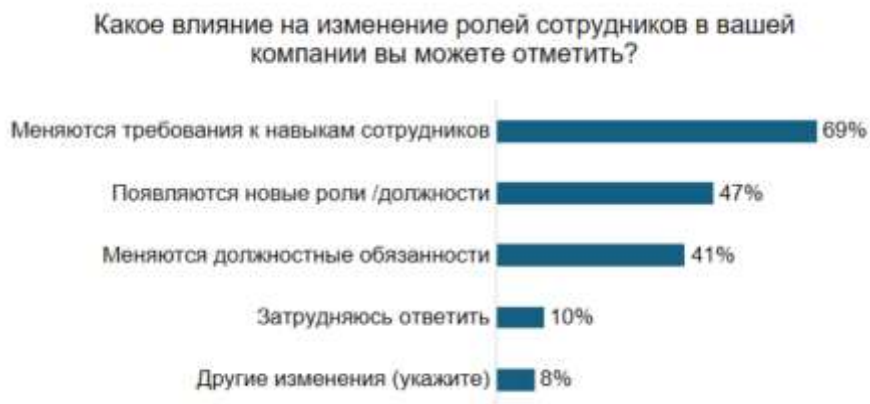
Тема изменения ролей сотрудников в связи с внедрением в компаниях генеративного ИИ до сих пор появлялась в обзоре по касательной. Между тем количественный опрос даёт по этому сюжету достаточно материала, чтобы разобрать его отдельно.

Изменения уже видны большинству. Влияние генеративного ИИ на изменение ролей сотрудников отмечают 59% респондентов. Не видят его 23%, ещё 17% затрудняются ответить.



Данные опроса «Видите ли вы влияние генеративного ИИ на изменение ролей сотрудников в вашей компании?»

Меняются прежде всего требования, а не штатное расписание. Среди тех, кто такое влияние отмечает, 69% говорят об изменении требований к навыкам сотрудников, 47% — о появлении новых ролей и должностей, 41% — об изменении должностных обязанностей.

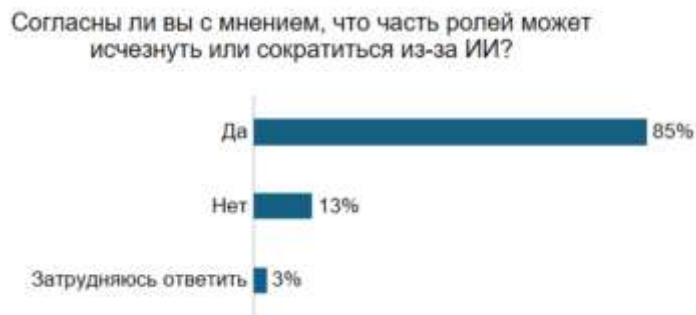


Данные опроса «Какое влияние на изменение ролей сотрудников в вашей компании вы можете отметить?»

Порядок здесь показателен. Первым меняется то, что не требует формальных решений: от людей начинают ожидать других навыков. Обязанности и должности перестраиваются медленнее, поскольку относятся к формальным кадровым процедурам, требующим времени и согласований. В результате возникает переходный период, когда сотрудник работает уже по-новому, а его функции описаны все еще по-старому. Именно этот разрыв порождает знакомую ситуацию: человек фактически руководит работой ИИ-инструментов, но в его должностной инструкции таких задач нет, а значит, нет ни оценки, ни ответственности, ни признания.

«Смещение фокуса сотрудников с экспертного на предпринимательский: технологичность и набор доступных технологий перестали быть ограничением, стало можно быстро экспериментировать, прототипировать, прорабатывать гипотезы и решать давно откладывавшиеся задачи... Ощущение такое, будто в штате появились высококвалифицированные консультанты и коучи вместе с крепкими мидл-исполнителями, которые позволяют быстро ставить задачи и оперативно получать результат», - Мусаев Замир, Директор центра информационных технологий Новостные продукты Группы «Интерфакс»

Сокращение ролей воспринимается как данность. Наиболее единодушный ответ во всём опросе: 85% согласны с мнением, что часть ролей может исчезнуть или сократиться из-за ИИ.



Данные опроса «Согласны ли вы с мнением, что часть ролей может исчезнуть или сократиться из-за ИИ?»

Здесь стоит сопоставить две цифры, поскольку вместе они говорят больше, чем по отдельности. Исчезновение части ролей считают вероятным 85%. Реальный эффект от внедрений при этом сегодня описывается языком ускорения, а не высвобождения: даже снижение трудозатрат (70%) означает, что прежняя работа делается быстрее, а не то, что её делает меньше людей, и влияние на выручку отмечает малое количество респондентов (15%).

Разрыв между ожиданием и наблюдаемым эффектом объясняется тем же переходом между ступенями зрелости, о котором говорилось выше. Пока ИИ ассистирует, роли сохраняются — меняется лишь содержание работы внутри них. Исчезновение ролей относится к следующей ступени, где процесс перестраивается, а человек перемещается из исполнительского контура в надзорный. Ожидания сотрудников, таким образом, опережают фактическое положение дел — и это само по себе является управленческой задачей: тревога возникает раньше, чем изменения, и работать с ней приходится тоже раньше.

«Бойтесь не того, что вас ИИ заменит, а что придут люди, которые умеют с ИИ работать», — представитель торгово-производственной компании

Из сложившейся ситуации можно сделать три практических вывода. Первый: пересмотр должностных требований и системы оценки стоит планировать одновременно с внедрением инструментов, а не после него — иначе сотрудники будут работать по одним правилам, а оцениваться по другим. Второй: обучение является не сопутствующей активностью, а условием получения эффекта, поскольку главным ограничением масштабирования названы именно компетенции. Третий: коммуникация об изменении ролей нужна на той стадии, когда сокращений ещё не происходит, — ожидания в компаниях уже сформировались, и молчание их не снимает.

Мнения экспертов

В этом разделе представлены мнения экспертов со стороны предложения — представителей «Сбера», фонда «Сколково» и Ахепiх, полученные в ходе глубинных интервью для настоящего исследования. Их оценки дополняют позицию заказчиков, представленную выше. Если заказчики рассматривают внедрение ИИ на примере собственной организации, то эксперты, работающие с разными компаниями, могут сопоставлять опыт десятков проектов, отраслей и сценариев. Их наблюдения во многом совпадают с выводами об особенностях российского рынка, практике внедрения и существующих барьерах, но позволяют рассмотреть эти вопросы на уровне рынка в целом.

Андрей Дмитриев, руководитель дирекции трансформации клиентов СiВ, «Сбер»

О разной степени зрелости заказчиков:

«Примерно 10–15% холдингов даже не пытаются внедрять сейчас какие-то пилоты: находятся в режиме ожидания... Вторая часть, самая крупная, — процентов 60–65 — это те, кто пытается проводить пилоты..., но зачастую эти пилоты носят несистемный характер... Последняя группа — 10–15% — это продуктивные внедрения, передовики»

О предпочтении российских моделей в регулируемых сценариях:

«Если компания не хочет дальше иметь интересный разговор с регулирующими и контролирующими органами и хочет стабильно работающий, скажем, чат-бот с правильными ответами, соответствующий российскому законодательству, то лучше, конечно, использовать российскую модель»

О балансе между запретом и легализацией теневого ИИ:

«С одной стороны, можно очень сильно ограничить сотрудников, но у тебя всегда есть свои гаджеты, ты можешь на них всё это делать... С другой стороны, все сошлись во мнении, что лучше дать сотрудникам максимальную возможность использовать модели — создать ту инфраструктуру, предоставить её в общее пользование... Здесь баланс — между обучением, развитием и использованием и корпоративной кибербезопасностью»

О том, что даёт устойчиво высокий процент успешных пилотов:

«Когда мы реализуем полный цикл пилота — то есть когда мы вместе с клиентом отбираем процесс, который можно пропилотировать, помогаем ему достичь этого экономического эффекта и правильно его посчитать, правильно найти экономические драйверы, — тогда у нас 90 и более процентов успеха»

О попытках заказчиков переложить ответственность на интегратора:

«Есть и экстремальные сценарии, когда крупные холдинги даже внутри договора на разработку за вендором, за интегратором фиксируют ответственность и влияние на производственные KPI, которые сами же считают»

Дмитрий Трофимов, управляющий директор департамента развития корпоративного бизнеса, «Сбер»

О коммодитизации мощных моделей для большинства бизнес-задач:

«...сейчас для 90% применений, по которым считается бизнес-кейс, вообще не нужны такие мощные фронтирные модели — нужны лёгкие версии на пару сотен миллионов параметров, которые прекрасно справляются с рутинным высоконагруженным куском задачи»

О требовании к экономическому обоснованию проектов:

«Если говорить об ИИ как о бизнесе — это высоконагруженные, рутинные, частотные куски процесса с большим количеством операций, и компания готова туда инвестировать, но уже требует конкретный бизнес-кейс с гарантией, сколько это сэкономит или принесёт»

О скорости трансформации ролей:

«Ещё в прошлом году у каждой команды в крупном холдинге был свой дизайнер. Сейчас, думаю, уже у половины команд его нет — не нужен. Через полгода никто не придёт на рабочее совещание с картинкой, нарисованной руками, — разве что для какой-то по-настоящему важной истории. Так произошло меньше чем за год»

О первой волне избыточного инвестирования в тяжёлые экспертные системы:

«Явный паттерн — это первая волна, когда ИИ только появился и все бросились собирать экспертные системы... Даже крупнейшие холдинги строили такие системы — они как-то продолжают существовать, туда было вложено много инвестиций, но по интенсивности потребления они практически не используются. Например, в агрономии: на предприятиях в России этих специалистов всего 5–7, и системы строились из опасения, что кто-то из них уйдёт и знания потеряются. Но этого не случилось, агрономы такими системами не пользуются — они и так всё знают»

Александр Вибе, директор Центра ИИ-решений для B2B, «Сбер»

О смещении проектного фокуса в сторону core-процессов:

«Мне кажется, агентные сотрудники, агентные процессы... станут частью типового ПО. Соответственно, весь проектный фокус и следующие свершения будут именно на core-процессах компании — тех, что просто так у вендора не купишь, потому что они более-менее кастомные: у двух девелоперов не бывает одинаковых процессов»

О сегментации заказчиков по зрелости:

«Я вижу три чётких уровня с точки зрения разделения по зрелости. Лидеры — банки, телеком, ИТ, e-commerce: продуктивные масштабированные внедрения с измеримой экономикой. Компании на этапе масштабирования — сюда чаще попадают крупная промышленность, ритейл, страхование, логистика и энергетика... Осторожное большинство — это часть промышленности, строительство, транспорт, фарма и госсектор... Основные тормоза — не сами модели, а данные, информационная безопасность, нехватка компетенций и сложность менять устоявшиеся процессы»

О конце эпохи «пилотов ради пилотов»:

«Что изменилось за последний год, и это самое важное изменение на рынке: эпоха "пилотов ради пилотов" закончилась. В 2023–2024 годах проекты запускали "чтобы не отстать", с 2025-го финансовые директора требуют оценку эффекта ещё на входе»

О новой роли интеграторов:

«С приходом ГенИИ роль интеграторов заметно изменилась: от "внедрить систему" они переходят к роли архитектора изменений и долгосрочного технологического партнёра. Для ИИ это особенно важно, потому что ценность создаётся не одной моделью, а связкой из данных, процессов, безопасности, настройки, контроля качества и эксплуатации»

О теневого ИИ как рыночном сигнале:

«Присутствует повсеместно — и предпочитаю относиться к этому не столько как к угрозе, а как к рыночному сигналу: если сотрудники массово обходят корпоративные запреты, чтобы использовать ИИ, значит, спрос огромный, а легального предложения внутри компании нет... Рабочая модель — "запретить опасное, дать легальную альтернативу": корпоративный ассистент в защищённом контуре с контролем доступа и модерацией»

О выборе между локальным, облачным и гибридным развёртыванием:

«Регулируемые отрасли и чувствительные данные — локальные конфигурации (on prem); облако выбирают для быстрых пилотов и некритичных данных, гибрид — оптимум для крупного нерегулируемого бизнеса, позволяющий значительно сэкономить на железе. Наша корпоративная платформа GigaCowork предлагает все три варианта — локальный, облачный и гибридный — и мы видим, что каждая конфигурация находит своих клиентов»

О границах доверия к ИИ-агентам:

«Граница доверия тут, в классическом понимании, проходит по трём осям: обратимость последствий, цена ошибки, проверяемость результата. От себя добавлю четвёртую, управленческую: доверие к ИИ-агенту строится так же, как

доверие к подрядчику — через историю безотказной работы и готовность вернуться за повторной услугой»

О фактическом уровне автономности решений:

«С точки зрения автономности, основной массив решений ограничивается подсказки и действия по жёстким правилам с подтверждением человеком. Быстро растёт слой цепочек действий с контрольными точками: агент планирует и исполняет, человек утверждает ключевые шаги. Полная автономия технически давно доступна, но применяется только там, где действия обратимы и цена ошибки низка: классификация, маршрутизация, черновики, массовый контент»

Дмитрий Филатов, заместитель директора центра «Сколково-Скай»

О концентрации рынка вокруг экосистем:

«У нас есть некая концентрация предложения, есть четыре крупных экосистемы: Сберовская, Яндексовская, МТС, и сейчас Т-технологии активно закрепляют роль на рынке. Соответственно рынок моделей и платформ складывается вокруг этих четырех экосистем — меньше конкуренции, но внутри выбранных систем проще интеграция»

О критериях перехода пилотов в продуктив:

«В продуктив чаще всего переходят проекты, где участвуют четыре момента: во-первых, обратимость — результат является предварительным черновиком, а не самостоятельным решением; есть один владелец процесса, а не несколько; данные уже собраны, находятся в одном месте, они машиночитаемые, их не надо долго обрабатывать; и цена ошибки не такая высокая... Получается, что в продуктив уходят не самые крутые, не самые умные сценарии, а те, где есть пока обратимый результат»

О государственном спросе как самостоятельном драйвере рынка:

«У нас по-другому ведет себя государство: сам государственный спрос является самостоятельным драйвером, потому что есть отраслевые показатели по внедрению ИИ... Но, с другой стороны, есть риск внедрения ради внедрения, ради показателей»

О более зрелых подходах к теневому ИИ:

«Есть более умный подход — когда пытаются создать управляемый контур, делая шлюз или роутер к разрешенным моделям, логирование, очистку данных, и разрешают сотрудникам так или иначе экспериментировать с открытыми инструментами. Есть еще более зрелая стратегия — когда есть разрешенный каталог, где находятся агенты, сценарии или модели, которыми позволительно пользоваться»

Об ограниченном доступе к моделям как драйвере on-premise-развёртываний:

«У нас есть специфика — во-первых, доступ к моделям: у заказчиков, у которых есть деньги в нашей стране, сейчас ограничен доступ к фронтирным моделям, работа с облаком тоже большой вопрос, соответственно рынок пошел в сторону on-premise и работы с моделями с открытым кодом. Мы больше работаем с собственным железом, а это означает более длинный цикл запуска проектов, но с другой стороны компания меньше зависит от внешних вендоров»

Лариса Малькова, руководитель практики Data & AI, Axenix

О 2025 годе как точке перелома:

«2025 год был годом перелома, когда от хайпа, каких-то слов компании перешли к делу... Все, кто в начале 25-го начинал пилотами активно по разным направлениям, у них сейчас действительно накоплена уже неплохая практика, подходы, понимание, как это делать, как правильно, как неправильно»

О рациональности выжидательной позиции:

«Достаточно большой процент выжидающих резонно говорит: ребята, если мне столько всего необходимо сделать, чтобы получить вот этот 10-процентный выигрыш, ещё и мерить его непонятно как, — пусть чемпионы прокладывают нам путь методом проб и ошибок. Чемпионам, первопроходцам всегда будет дорого. Мы тут немножко подождём, когда будут выработаны методологии, подходы, best practices, фреймворки измерения эффектов, — вот мы тогда и двинемся»

О различии между культурной зрелостью компаний и стартовой точкой внедрения:

«Компании действительно находятся на разном уровне зрелости с точки зрения своей цифровизации, культуры, тех или иных процессов. И когда мы говорим, что мы начинаем с некой стартовой точки и сейчас пойдём в ассистированную историю, потом в полуавтономную..., то на самом деле мы за точку отсчёта уже берём достаточно высокий уровень. А это далеко не везде так. Кто-то ещё находится на уровне «минус один», «минус два»... Эти культурно-управленческие проблемы никуда не делись»

О новых ролях и о людях, формулирующих намерение:

«Раньше мы говорили : люди, процессы, технологии. А сейчас мы людей заменили уже на акторов, то есть на участников... Искусственный интеллект сам по себе ничего не хочет, у него нет собственной мотивации на что-то. А вот формулирование того, что ты хочешь получить, причём в виде, который будет понятен не человеку, а понятен машине, — это, безусловно, очень важный [новый] навык»

О необходимости независимого контроля за парком агентов:

«Конечно, будут люди, которые должны следить за парком агентов — не происходит ли там какого-то сдвига в их поведении, нежелательного, и так далее. Потому что это уже сильно более самостоятельные сущности с недетерминированным

поведением... Желательно, знаете, как в авиации: двойной, тройной контуры контроля и обратной связи»

Дмитрий Иванов, руководитель платформы разработки SourceCraft, «Яндекс»

О том, как ИИ усиливает специалистов:

«Моё глубокое убеждение: если ты сильный специалист, то AI тебя довольно сильно усиливает, создаёт дополнительный рычаг»

О командах, состоящих из людей и агентов:

«Следующий виток эволюции — мы тоже сейчас над этим работаем — заключается в том, что помимо ассистента у тебя в команде будут автономные агенты, специализированные под определённый ряд работ»

О переделке корпоративных систем под работу агентов:

«Есть запрос на то, чтобы существующие интерфейсы становились agent-friendly, то есть информационные системы перестраивались так, чтобы с ними мог работать агент. Хороший критерий — когда новый человек приходит в команду и может через диалог узнать всё, что нужно; когда тебе не нужно идти к людям за знаниями, которые хранятся в головах, а можно обратиться к агенту»

Артур Самигуллин, [должность уточняется], «Яндекс»

О горизонтальных функциях как удачной точке входа:

«Юрист получает экономию времени на подготовке договора, анализе прецедентов, законодательства и других данных, при этом финальные выводы перепроверяет сам. В Яндексе есть примеры внедрения Нейроюриста — юристы хорошо оценивают этот инструмент».

О квалификации входящего лида как о первом выстрелившем кейсе в продажах:

«Один из первых кейсов, который выстрелил и у нас, и в мире, — автоматизация входящего лида. Агенту дают доступ к базе знаний компании, внутренним источникам и вебу, а дальше он квалифицирует лид, объясняет, что это за клиент и как его соотнести с продуктами. Продавец получает уже готовый, описанный лид и работает с клиентом»

Об универсальных ассистентах как о следующем запросе рынка:

«Универсальные системы, универсальные агенты — суперинтересная вещь. Система получает доступ к вертикальным, горизонтальным и сервисным инструментам, а сотрудник выступает постановщиком задачи. Мы видим, что с каждой итерацией развития моделей это всё лучше работает: они становятся всё более универсальными и всё лучше решают задачи, к которым им дают доступ. Главное — правильно дать этот доступ и описать, как работать в системах. Вот за этим, мне кажется, будущее».

Практические рекомендации для менеджера

Ниже собраны выводы прикладного характера, следующие из материала обзора. Они не претендуют на роль методологии — общепринятой методологии в отрасли пока не сложилось — но описывают те решения, которые в российской практике чаще всего отделяют проекты, дошедшие до продуктива, от остановившихся на стадии пилота.

1: начинать не с выбора вендора, а с инвентаризации процессов и данных. Формализованное описание процесса покрывает, по оценке интеграторов, 10-20% того, что происходит в действительности; корпоративные базы знаний накапливались годами без ревизии противоречий, а поля в хранилищах данных остались без описаний, поскольку их смысл был очевиден разработчикам. Человек восполняет недостающий контекст из опыта, языковая модель — нет. Прежние волны автоматизации хотя бы принуждали компанию описать процесс, иначе систему было не настроить; генеративный ИИ такого принуждения не создаёт — он выдаёт правдоподобные ответы поверх неопisanного процесса, и неупорядоченность обнаруживается не на старте, а при попытке масштабирования. Поэтому инвентаризация — не подготовительная фаза ИИ-проекта, а его содержание, и она должна быть отражена в сроках и бюджете.

2: точку входа выбирать по четырём признакам. Практика показывает, что в промышленную эксплуатацию переходят не самые технологически амбициозные сценарии, а те, где результат обратим и носит характер черновика, а не окончательного решения; цена ошибки невысока; данные уже собраны в одном месте, машиночитаемы и не требуют долгой обработки; у процесса есть единственный владелец, а не несколько. Клиентский сервис стал самым зрелым сценарием именно потому, что удовлетворяет всем четырём условиям сразу и вдобавок даёт готовую метрику эффекта. Проверка замысла по этим четырём признакам занимает несколько часов и позволяет заранее оценить, к какой категории относится задуманный проект.

3: экономику считать до старта и целиком. Публичные цифры экономии заказчики оценивают скептически, и на то есть две причины, причём вторая важнее первой: помимо приукрашивания результатов, в отрасли не сложилось общепринятой методологии измерения — даже в разработке ПО, сценарии с наиболее очевидным эффектом, нет согласия о том, что именно измерять. Расчёт следует вести в модели полной стоимости владения, включая роли, необходимые для эксплуатации: специалистов, которые дообучают модель, обновляют базу знаний и проверяют качество ответов. Иначе экономия оборачивается переносом затрат из одного подразделения в другое. Пилот, остановленный на таком расчёте — не провал, а сэкономленные ресурсы.

4: требования безопасности и архитектуру развёртывания согласовывать до пилота, а не после результата. Полезно заранее ответить на три вопроса: какие данные допустимо выпускать за периметр, какая конфигурация развёртывания

приемлема для службы безопасности и хватит ли имеющихся вычислительных мощностей для последующего тиража. **Последний вопрос ключевой: неготовность инфраструктуры — самый заметный барьер на стадии пилотов (её называют 51% пилотирующих компаний против 21% активно использующих), а мощности, рассчитанные на одну функцию, могут оказаться недостаточными, когда решение нужно распространить на организацию.**

5: строить управленческий контур, пока агентов ещё немного. Компания, у которой сейчас три-четыре агента, выстроит каталог решений, реестр владельцев, правила доступа и логирование существенно дешевле, чем та, которой придётся наводить порядок в нескольких десятках разрозненных внедрений. Отставание от мировой статистики здесь даёт редкую возможность не повторять чужие ошибки. Начать стоит с терминологии: пока «агентом» в компании называют и сценарный чат-бот, и автономную многошаговую систему, невозможно ни вести реестр, ни оценивать риски, ни планировать ресурсы — объекты управления просто не определены.

6: теневому ИИ противопоставлять легальную альтернативу, а не запрет. Запрет технически не работает: у сотрудника всегда есть личное устройство, и ограничение на корпоративном контуре лишь выталкивает использование за его пределы, где оно становится невидимым. Рабочая модель — запретить действительно опасное и дать сопоставимую по удобству легальную замену: корпоративного ассистента в защищённом контуре с контролем доступа, шлюз к разрешённым моделям с логированием и очисткой данных, каталог одобренных сценариев. Массовый обход запретов стоит читать как рыночный сигнал: спрос велик, а легального предложения внутри компании нет — возможно, здесь и находится зерно будущего успешного проекта.

7: роли, систему оценки и обучение менять одновременно с внедрением инструментов. Требования к навыкам меняются быстрее описания должностных обязанностей, и между ними лежит зона, где сотрудник работает по-новому, а описан по-старому: фактически руководит работой ИИ-инструментов, но в должностной инструкции этого нет, а значит, нет ни оценки, ни признания, ни ответственности. При этом более трети компаний обучение навыкам ИИ не ведут вовсе, хотя сами же называют дефицит компетенций главным ограничением. Отдельная задача — коммуникация: исчезновение части ролей считают вероятным 85% опрошенных, тогда как фактически пока что речь идет об ускорении операций и увеличении производительности, а не о сокращениях людей. Тревога сотрудников возникает раньше изменений, и работать с ней приходится тоже раньше.

Общее у всех семи пунктов одно: ни один из них не требует ни выбора модели, ни немедленного доступа к передовым технологиям. Все они находятся в зоне контроля самой компании — и именно поэтому определяют исход проекта сильнее всего.

«Каждой компании сейчас предстоит пройти такую трансформацию, которая по своим масштабам, импакту и степени сложности будет сложнее, чем была

Заключение

Технологический выбор перестал быть главным вопросом для заказчиков. Для основной массы бизнес-сценариев фронтальные модели избыточны — рутинные высокочастотные операции успешно закрываются моделями среднего класса. Предметом инженерной работы стала не сама модель, а обвязка вокруг неё: маршрутизация запросов между моделями разной стоимости, трассировка, контроль качества ответов, управление правами и стоимостью вычислений.

Российский рынок отличается от мирового не темпом, а устройством. Особая роль государства как заказчика, регулятора и соинвестора; концентрация вокруг трёх-четырёх корпоративных экосистем вместо стартапов; приоритет изолированного развёртывания; открытый код как основной способ доступа к передовым возможностям; фактически свёрнутый венчурный контур — это устойчивые характеристики рынка, а не временные обстоятельства. Отставание сокращается само, структурные особенности — нет. Выбирая модель, компания нередко выбирает заодно облако, инструменты и интегратора, поэтому решение о поставщике стоит принимать на уровне архитектуры, а не отдельной закупки. А там, где западный заказчик берёт готовый коннектор из библиотеки, российский разворачивает интеграционный проект — одинаковые по названию работы имеют принципиально разную трудоёмкость.

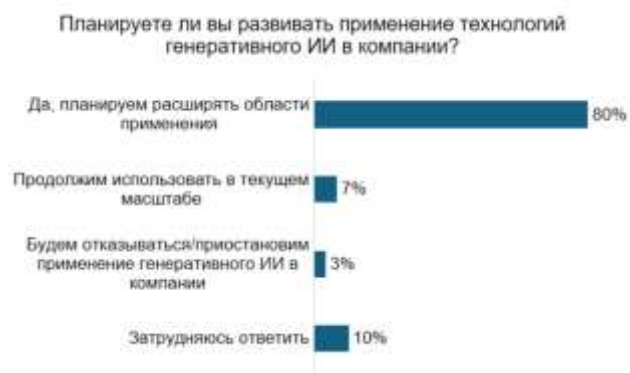
Практика использования сложилась там, где результат достижим и проверяем. В продуктив уходят не самые технологически амбициозные сценарии, а те, где результат обратим, цена ошибки невысока, данные уже собраны в одном месте, а у процесса есть единственный владелец. Этим объясняется и лидерство клиентского сервиса и автоматизации разработки, и то, что российская мультимодальность закрепились в двух опорных парах — «речь плюс текст» и «документ плюс текст», где качество достаточно, стоимость обработки приемлема, а вывод поддаётся проверке.

Барьеры внедрений - не только технологии. Компаниям не хватает понимания собственных процессов и качества собственных данных: формализованное описание процесса покрывает, по оценке интеграторов, 10-20% того, что происходит в действительности, а языковая модель, в отличие от человека, не восполняет недостающий контекст из опыта. Требования безопасности заталкивают проект в закрытый контур, а закрытый контур немедленно превращается в инфраструктурную задачу. Экономика стала входным условием: с 2025 года финансовые директора требуют оценку эффекта ещё на старте, при этом общепринятой методологии измерения в отрасли не сложилось. Наконец, ответственность за результат остаётся на стороне заказчика — поставщик не может гарантировать бизнес-эффект, зависящий от чужих данных и чужих процессов.

Основной эффект ещё впереди, и он достигается не заменой модели на более совершенную, а сменой устройства процесса. Пока ИИ ассистирует, не меняя устройства процесса, сквозной выигрыш составляет около 10%. Нелинейная, «взрывная» отдача появится на следующей ступени, где агентный контур уйдет в «автономное плавание», а человек переместится из цикла принятия решений на позицию супервизора. Такие проекты пока единичны, и их цена — перестройка процесса целиком, то есть речь идет уже не об ИТ-проекте, а об организационном изменении. Один из экспертов называет предстоящее компаниям «второй цифровой трансформацией», сложнее первой по масштабу, влиянию и степени неопределённости; оценка выглядит обоснованной хотя бы потому, что готовых фреймворков, аналогичных тем, что задавали ориентир в прошлую волну, пока что не существует.

Российский рынок находится в не самом удобном, но продуктивном положении: технологические ограничения перестали быть определяющими для большинства корпоративных сценариев, а организационные ещё не стали привычными. Это означает, что результат внедрения сегодня зависит преимущественно от решений, находящихся в зоне контроля самой компании.

И последнее наблюдение, важное для оценки перспективы. При всех перечисленных трудностях — незрелых данных, спорной экономике, отсутствующих фреймворках — рынок не разворачивается назад. Расширять области применения генеративного ИИ планируют 80% компаний, сохранить текущий масштаб — 7%, и лишь 3% собираются отказаться от технологии или приостановить её применение.



Данные опроса «Планируете ли вы развивать применение технологий генеративного ИИ в компании?»

Неудачи отдельных проектов не приводят к разочарованию в технологии в целом: компании корректируют ожидания и подходы, но не отказываются от дальнейшего развития этого направления. Для тех, кто сегодня решает, стоит ли начинать собственные проекты с генеративным ИИ, это, возможно, самая содержательная цифра всего обзора.