

Внедрение искусственного интеллекта в компании малого, среднего и крупного бизнеса: этапы, команды, инструменты, мировой и китайский опыт

Аналитический отчёт, сентябрь 2026

Резюме

В 2024–2026 годах искусственный интеллект перешёл из категории экспериментов в категорию обязательной компетенции бизнеса: регулярно используют ИИ уже 88% организаций мира, однако измеримую ценность в масштабе извлекают лишь 5–6% — разрыв между adoption и impact стал главной проблемой повестки. Отчёт показывает, что «бутылочное горлышко» внедрения носит организационный, а не технологический характер: успех на 70% определяется людьми и процессами (правило BCG 10-20-70), а сильнейшие корреляты результата — спонсорство CEO, фундаментальный редизайн процессов и системное измерение ценности. Все ведущие модели зрелости (Gartner, MIT CISR, BCG, Microsoft, Accenture, SEI) конвергируют к единой дорожной карте «подготовка → пилоты → операционализация → масштабирование → трансформация» с критическим переходом «пилот → масштаб», на котором застревает большинство компаний. Размер бизнеса определяет не ответ на вопрос «использовать ли ИИ», а архитектуру доступа, команды и governance: малый бизнес выигрывает за счёт встроенного в SaaS ИИ и низкого порога входа (\$10–100 в месяц), средний — за счёт компактных центров компетенций, крупный — за счёт платформенной модели hub-and-spoke и формализованного управления доступами (RBAC/SCIM/SSO, agent identity). Инструментальный рынок расслоился на open source стек и коммерческие платформы, причём 76% корпоративных сценариев покупаются, а не строятся, а вендорские внедрения успешны вдвое чаще внутренних разработок. Китайский опыт демонстрирует альтернативную модель массового внедрения — государственное форсирование, ценовая война API и открытые модели, — обеспечившую 45% масштабных внедрений против 38% в среднем по миру. Практическая часть отчёта даёт дорожные карты по размеру бизнеса, типологию ошибок и приоритеты по агентному ИИ — фронтиру 2026 года, где дисциплина governance должна опережать развёртывание.

Содержание

1. Введение: состояние внедрения ИИ в бизнесе в 2024–2026
2. Этапы внедрения ИИ и модели зрелости
3. Подходы и схемы внедрения по размеру бизнеса
4. Команды и организационные структуры для внедрения ИИ
5. Управление ИИ: governance, политики и распределение доступов сотрудников
6. Программные инструменты для управления ИИ и ИИ-процессами
7. Китайский опыт внедрения ИИ

1. Введение: состояние внедрения ИИ в бизнесе в 2024–2026

Период 2024–2026 годов стал точкой перелома в отношениях бизнеса и искусственного интеллекта: технология перешла из категории экспериментов в категорию обязательной компетенции, однако разрыв между «использованием» и «реальным эффектом» не только не закрылся, но и стал главной проблемой повестки. Настоящая глава обобщает статистику внедрения, калибрует конфликтующие оценки «провалов» (с обязательными оговорками об определениях) и формулирует центральный вопрос, на который отвечают последующие главы: как перейти от массового использования ИИ к измеримой ценности — и как этот переход различается для малого, среднего и крупного бизнеса.

1.1. Динамика adoption: взрывной рост

По данным McKinsey (State of AI), доля организаций, использующих ИИ хотя бы в одной бизнес-функции, после шести лет стагнации около 50% выросла до 72% в начале 2024 года, а использование генеративного ИИ почти удвоилось — с ~33% (2023) до 65% (2024) ¹. К ноябрю 2025 года (опрос 1 993 респондентов в 105 странах) регулярное использование ИИ достигло 88% организаций против 78% годом ранее ². Распределение компаний по фазам внедрения по данным того же опроса выглядит так: 32% — эксперименты, 30% — пилоты, 31% — масштабирование и только 7% — полноценное внедрение на уровне предприятия ³. Иными словами, «использующих» и «внедривших» — принципиально разные популяции: между первыми и вторыми лежит фильтр, который проходит менее чем одна компания из десяти.

Инвестиционный фон беспрецедентен: глобальные затраты на ИИ в 2024 году оценивались более чем в \$252 млрд (BCG), а прогноз Gartner на 2025 год — около \$1,5 трлн ИИ-расходов ⁴ (confidence: medium, агрегированная оценка). Для сравнения: предыдущий шестилетний цикл (2018–2023) характеризовался стабильным уровнем использования около 50% организаций и преимущественно узкими прикладными задачами (рекомендации, скоринг, компьютерное зрение). Генеративный слом 2023–2024 годов изменил саму экономику внедрения: универсальная модель с текстовым

- 1 McKinsey State of AI 2024 (анализ Libertify) — <https://www.libertify.com/interactive-library/mckinsey-state-of-ai-2024/> (2024 (отчёт McKinsey, май 2024))
- 2 McKinsey State of AI 2025 (анализ Libertify / SemanticOS / TAdviser) — <https://www.libertify.com/ar/interactive-library/mckinsey-state-of-ai-2025/> ; <https://semantics.io/blog/mckinsey-state-of-ai-2025-agents> (2025-11 (публикация отчёта))
- 3 TAdviser, «Внедрение ИИ в бизнесе» (пересказ McKinsey Global Survey 2025) — https://www.tadviser.ru/index.php/Статья:Внедрение_ИИ_в_бизнесе (2025 (обновлено 2026-07-30))
- 4 S&P Global Market Intelligence (март 2025); оценки глобальных ИИ-инвестиций BCG (\$252 млрд, 2024) и Gartner (~\$1,5 трлн, 2025) — <https://www.digitalapplied.com/blog/agent-ai-project-cancellations-gartner-40-percent-2026> (2025-03-14); см. также <https://talyx.ai/insights/enterprise-ai-implementation-failure>

интерфейсом оказалась применимой практически к любой функции, что и объясняет скачок adoption до 72–88%.

Характерно, что волна 2024–2026 годов отличается от предыдущих циклов ИИ-хайпа: во-первых, технология стала потребительски доступной (чат-интерфейс, подписка от \$20 в месяц), что сняло барьер входа для малого бизнеса; во-вторых, генеративный ИИ распространялся «снизу вверх» — через сотрудников, а не только через ИТ-департаменты, что породило феномен теневого использования; в-третьих, уже к 2025 году началась вторая волна — агентный ИИ, — не дождавшись завершения первой. По данным McKinsey, об использовании генеративного ИИ в 2025 году заявили 79% предприятий против 71% годом ранее ⁵ — рост продолжается, но темпы замедляются по мере насыщения: «использовать ИИ хотя бы где-то» перестаёт быть дифференциацией.

Отдельного внимания заслуживает региональная дифференциация: китайская модель (государственное форсирование + ценовая война на API + открытые модели) демонстрирует 45% масштабных внедрений против 38% в среднем глобально (McKinsey China; confidence: medium), что показывает: барьеры стоимости и компетенций снимаются не только рынком, но и институциональной организацией доступа.

1.2. Разрыв «adoption vs impact»

Ключевой парадокс периода — массовое использование при минимальном измеримом эффекте. McKinsey фиксирует: лишь 39% организаций видят какое-либо влияние ИИ на EBIT, причём у большинства из них вклад ИИ составляет менее 5% EBIT. Только ~6% организаций квалифицируются как «AI high performers» — с более чем 5% EBIT, атрибутируемой к ИИ ⁶. BCG в исследовании «The Widening AI Value Gap» (более 1 250 топ-менеджеров) даёт сходную картину: 5% компаний — «future-built», получающие существенную ценность от ИИ в масштабе, 35% — масштабируются и начинают генерировать ценность, а 60% — отстающие с минимальными выгодами ⁷. Ранее, в октябре 2024 года, BCG оценивала: только 26% компаний обладают компетенциями для перехода от PoC к реальной ценности, а 74% не получили ощутимой ценности от ИИ (опрос 1 000 СxО в 59 странах) ⁸.

Финансовые последствия разрыва измеримы. По BCG, «future-built» компании показывают 1,7-кратный рост выручки, 3,6-кратную трёхлетнюю совокупную акционерную доходность (TSR), 2,7-кратный ROIC и 1,6-кратную маржу EBIT по сравнению с

5 TAdviser, «Внедрение ИИ в бизнесе» (пересказ McKinsey Global Survey 2025) — https://www.tadviser.ru/index.php/Статья:Внедрение_ИИ_в_бизнесе (2025 (обновлено 2026-07-30))

6 McKinsey State of AI 2025 (анализ Libertify / SemanticOS / TAdviser) — <https://www.libertify.com/ar/interactive-library/mckinsey-state-of-ai-2025/> ; <https://semantics.io/blog/mckinsey-state-of-ai-2025-agents> (2025-11 (публикация отчёта))

7 BCG report (анализ Libertify; подтверждение SDXExec) — <https://www.libertify.com/interactive-library/ai-value-gap-enterprise-strategy-bcg-2025/> ; <https://www.sdcexec.com/software-technology/ai-ar/news/22951486/> (2025-09/10)

8 BCG report (анализ Libertify; подтверждение SDXExec) — <https://www.libertify.com/interactive-library/ai-value-gap-enterprise-strategy-bcg-2025/> ; <https://www.sdcexec.com/software-technology/ai-ar/news/22951486/> (2025-09/10)

отстающими; 6,2% прироста их выручки атрибутируется к ИИ против 1,2% у отстающих; 62% их ИИ-инициатив уже развёрнуты или масштабируются против 12% у отстающих, а время до эффекта (time-to-impact) — 9–12 месяцев против 12–18 месяцев⁹. Ретроспектива BCG за три года подтверждает тот же вывод на другой выборке: ИИ-лидеры показали 1,5-кратно более высокий рост выручки, 1,6-кратную акционерную доходность и 1,4-кратную рентабельность инвестированного капитала¹⁰.

Динамика 2024 → 2025 годов обнадеживает и тревожит одновременно: доля масштабирующихся компаний выросла с 22% до 35%, а стагнирующих сократилась с 25% до 14%¹¹ — середина рынка движется вверх; но доля «future-built» выросла лишь с 4% до 5%, то есть вершина пирамиды остаётся эксклюзивной клубовой категорией, куда переход единиц.

Поведенческий портрет high performers (McKinsey 2025) показывает, чем этот клуб отличается от остальных: они в 3,6 раза чаще проводят трансформационные изменения на уровне предприятия (а не точечные улучшения); 55% из них фундаментально перестроили рабочие процессы под ИИ — почти втрое чаще остальных компаний; и они примерно втрое чаще демонстрируют сильную вовлечённость топ-руководства¹². То есть элита отличается не бюджетами и не доступом к моделям (технологии доступны всем и дёшевеют), а глубиной организационной перестройки.

1.3. Почему большинство пилотов не дают ценности — с оговорками об определениях

Статистика «провалов» варьируется в широком диапазоне, и каждая цифра требует уточнения определения:

- **MIT NANDA, «The GenAI Divide» (август 2025; 150 интервью с руководителями, опрос 350 сотрудников, анализ 300 публичных внедрений):** ~95% корпоративных GenAI-пилотов не дали измеримого влияния на P&L при \$30–40 млрд корпоративных затрат¹³. **Важная оговорка:** «отсутствие измеримого P&L-эффекта» не равно «провалу проекта» — у большинства пилотов просто не было базовой линии для измерения; отчёт не является рецензируемым исследованием.

9 BCG report (анализ Liberty; подтверждение SDXExec) — <https://www.libertify.com/interactive-library/ai-value-gap-enterprise-strategy-bcg-2025/>; [https://www.sdcexec.com/software-technology/ai-ar/news/22951486/\(2025-09/10\)](https://www.sdcexec.com/software-technology/ai-ar/news/22951486/(2025-09/10))

10 BCG, “Where’s the Value in AI?” — <https://www.bcg.com/press/24october2024-ai-adoption-in-2024-74-of-companies-struggle-to-achieve-and-scale-value> (24 октября 2024)

11 BCG report (анализ Liberty; подтверждение SDXExec) — <https://www.libertify.com/interactive-library/ai-value-gap-enterprise-strategy-bcg-2025/>; [https://www.sdcexec.com/software-technology/ai-ar/news/22951486/\(2025-09/10\)](https://www.sdcexec.com/software-technology/ai-ar/news/22951486/(2025-09/10))

12 McKinsey State of AI 2025 (анализ Liberty / SemanticOS / TAdviser) — <https://www.libertify.com/ar/interactive-library/mckinsey-state-of-ai-2025/>; <https://semanticos.io/blog/mckinsey-state-of-ai-2025-agents> (2025-11 (публикация отчёта))

13 MIT NANDA report (анализы Trullion, agentmodeai, ontheground) — <https://agentmodeai.com/the-mit-genai-pilot-failure-claim/>; <https://ontheground.agency/resources/why-ai-projects-fail-mit-2025-genai-divide> (2025-08 (отчёт)); Fortune-материал 2025-08-21)

Независимый ориентир: IDC/Lenovo (март 2025) оценивает, что 88% ИИ-РоС не доходят до продакшена (~4 из 33) ¹⁴.

- **Gartner:** прогноз июля 2024 — не менее 30% GenAI-проектов будут заброшены после РоС к концу 2025 из-за плохих данных, слабого контроля рисков, растущих затрат и неясной ценности; ретроспектива января 2026 — фактически не менее 50% проектов заброшены ¹⁵¹⁶. Дополнительно: в среднем только 48% ИИ-проектов доходят до продакшена, путь от прототипа до продакшена занимает ~8 месяцев ¹⁷.
- **BCG (октябрь 2024):** 74% компаний не получили осязаемой ценности ¹⁸.
- **S&P Global Market Intelligence (март 2025, 1000+ респондентов):** 42% компаний забросили большинство ИИ-инициатив в 2025 году против 17% годом ранее; в среднем 46% РоС отбраковываются до продакшена ¹⁹.
- **RAND Corporation (2024):** до 80% ИИ-проектов терпят неудачу — почти вдвое выше, чем у не-ИИ IT-проектов ²⁰ (confidence: medium-high).

Воронка корпоративных GenAI-решений по MIT NANDA: 60% организаций оценивали решения, 20% дошли до пилота, только 5% — до промышленной эксплуатации. Оценки IDC/Lenovo (март 2025) рисуют ту же структуру отбора: 88% ИИ-РоС не доходят до продакшена, «выпускается» примерно 4 проекта из 33 ²¹. Ретроспективная оценка Gartner (январь 2026) — фактически не менее 50% GenAI-проектов заброшены после РоС к концу 2025 — оказалась существенно жёстче собственного прогноза 2024 года (≥30%), что иллюстрирует систематическое недооценивание рынком сложности перехода пилот → продакшен ²².

Как интерпретировать этот разброс цифр (30% / 50% / 74% / 88% / 95%) без противоречий:

- 14 MIT NANDA report (анализы Trullion, agentmodeai, ontheground) — <https://agentmodeai.com/the-mit-genai-pilot-failure-claim/> ; <https://ontheground.agency/resources/why-ai-projects-fail-mit-2025-genai-divide> (2025-08 (отчёт); Fortune-материал 2025-08-21)
- 15 Gartner Press Release (перепечатка HPCwire / Businessworld / IntelligentCIO) — <https://www.hpcwire.com/bigdatawire/this-just-in/gartner-predicts-30-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025/> ; <https://www.intelligentcio.com/me/2024/08/05/gartner-predicts-30-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025/> (2024-07-29/30)
- 16 Gartner — “Why 50% of GenAI Projects Fail — And How to Beat the Odds” (Arun Chandrasekaran) — <https://www.gartner.com/en/articles/genai-project-failure> (2026-01-26)
- 17 Gartner via Informatica blog — <https://www.informatica.com/blogs/the-surprising-reason-most-ai-projects-fail-and-how-to-avoid-it-at-your-enterprise.html> (2025-02-11)
- 18 BCG report (анализ Libertify; подтверждение SDXExec) — <https://www.libertify.com/interactive-library/ai-value-gap-enterprise-strategy-bcg-2025/> ; <https://www.sdcexec.com/software-technology/ai-ar/news/22951486/> (2025-09/10)
- 19 S&P Global Market Intelligence (март 2025); оценки глобальных ИИ-инвестиций BCG (\$252 млрд, 2024) и Gartner (~\$1,5 трлн, 2025) — <https://www.digitalapplied.com/blog/agent-ai-project-cancellations-gartner-40-percent-2026> (2025-03-14); см. также <https://talyx.ai/insights/enterprise-ai-implementation-failure>
- 20 RAND via timspark / talyx — <https://timspark.com/blog/why-ai-projects-fail-artificial-intelligence-failures/> (2024 (RAND report))
- 21 MIT NANDA report (анализы Trullion, agentmodeai, ontheground) — <https://agentmodeai.com/the-mit-genai-pilot-failure-claim/> ; <https://ontheground.agency/resources/why-ai-projects-fail-mit-2025-genai-divide> (2025-08 (отчёт); Fortune-материал 2025-08-21)
- 22 Gartner — “Why 50% of GenAI Projects Fail — And How to Beat the Odds” (Arun Chandrasekaran) — <https://www.gartner.com/en/articles/genai-project-failure> (2026-01-26)

исследования измеряют разные точки отсева (заброшен после PoC / не дошёл до продакшена / без осязаемой ценности / без измеримого P&L-эффекта) на разных популяциях и в разное время. Все они согласованы в одном: пропускная способность конвейера «идея → продакшен → ценность» в 2024–2026 годах низка, и сужается он преимущественно на организационных стыках.

Три независимых источника (MIT, Gartner, BCG) согласованы по направлению: **большинство инициатив не конвертируется в ценность, но причина — не технология.** Конвергентный вывод исследований: «бутылочное горлышко» — организационный переход от пилота к масштабу. Сильнейшие корреляты успеха — спонсорство CEO, фундаментальный редизайн процессов и оргмодель, тогда как технологии доступны и дешевеют. Правило BCG 10-20-70 фиксирует: успех ИИ на 10% зависит от алгоритмов, на 20% — от технологий и данных, на 70% — от людей и процессов ²³.

При этом ранние adopters, прошедшие фильтр, сообщают о значимых результатах: опрос Gartner (822 бизнес-лидера, конец 2023) фиксирует в среднем +15,8% выручки, –15,2% затрат и +22,6% продуктивности ²⁴. Задача компаний 2024–2026 годов — не «начать использовать ИИ», а попасть в те 5–6%, которые извлекают из него измеримую ценность.

1.4. Корневые причины разрыва

Анализ причин «пилотного чистилища» в исследованиях 2024–2026 годов устойчиво выделяет одни и те же факторы — и все они организационные, а не технологические:

1. **Отсутствие измерения.** Знаменитые «95%» MIT NANDA во многом объясняются тем, что у большинства пилотов не было базовой линии: ценность невозможно доказать post factum, если до внедрения никто не зафиксировал метрики процесса ²⁵.
2. **Качество и доступность данных, слабый риск-контроль, эскалация затрат** — триада причин отказа от проектов в формулировке Gartner ^{26,27}.
3. **Неправильное распределение бюджетов.** По MIT NANDA, 50–70% GenAI-

23 BCG via USDM blog — <https://www.usdm.com/resources/blogs/why-ai-adoption-is-failing-in-life-sciences> (2025-12-30)

24 Gartner Press Release (перепечатка HPCwire / Businessworld / IntelligentCIO) — <https://www.hpcwire.com/bigdatawire/this-just-in/gartner-predicts-30-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025/>; <https://www.intelligentcio.com/me/2024/08/05/gartner-predicts-30-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025/> (2024-07-29/30)

25 MIT NANDA report (анализы Trullion, agentmodeai, ontheground) — <https://agentmodeai.com/the-mit-genai-pilot-failure-claim/>; <https://ontheground.agency/resources/why-ai-projects-fail-mit-2025-genai-divide> (2025-08 (отчёт); Fortune-материал 2025-08-21)

26 Gartner Press Release (перепечатка HPCwire / Businessworld / IntelligentCIO) — <https://www.hpcwire.com/bigdatawire/this-just-in/gartner-predicts-30-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025/>; <https://www.intelligentcio.com/me/2024/08/05/gartner-predicts-30-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025/> (2024-07-29/30)

27 Gartner — “Why 50% of GenAI Projects Fail — And How to Beat the Odds” (Arun Chandrasekaran) — <https://www.gartner.com/en/articles/genai-project-failure> (2026-01-26)

бюджетов уходит в продажи и маркетинг, тогда как наибольший измеримый ROI даёт автоматизация бэк-офиса; BCG добавляет, что 62% ценности ИИ создаётся в основных бизнес-функциях.

4. **Теневое использование.** Между формальной политикой и реальным поведением существует разрыв: по MIT NANDA, 40% компаний купили официальные подписки, но 90% сотрудников пользуются ИИ через личные аккаунты²⁸; по BCG, 54% сотрудников готовы применять неавторизованные инструменты. Shadow AI делает формальные запреты бессмысленными без технического принуждения и управляемого легального доступа.
5. **Недоинвестирование в людей.** Почти половина сотрудников не получила никакого обучения ИИ (McKinsey 2025), хотя именно обучение называют главным стимулом использования.

Вывод введения: в 2024–2026 годах рынок прошёл фазу «технологического энтузиазма» и вошёл в фазу «дисциплины ценности». Разрыв между лидерами и отстающими (BCG: 5% против 60%) растёт, а не сокращается, и определяется он не выбором модели или вендора, а организационной зрелостью — чему и посвящена следующая глава.

Отдельно подчеркнём статус приведённой статистики. Часть цифр подтверждена несколькими независимыми источниками (разрыв adoption/impact, преимущество buy-over-build, правило 10-20-70, массовость Shadow AI) — это высокая степень уверенности. Часть опирается на один авторитетный источник или самоотчёты компаний (кейсы JPMorgan, Klarna, Walmart) — средняя уверенность. Наконец, ряд популярных метрик (ROI 3,7× для SMB, средний ROI агентных внедрений 171%) происходит из маркетинговых агрегаторов и используется в отчёте с явными оговорками. Конфликты между источниками (например, «95% провалов» MIT против «30–50%» Gartner) разрешаются не выбором «правильной» цифры, а уточнением определений — такой подход сохранён во всех главах отчёта.

2. Этапы внедрения ИИ и модели зрелости

2.1. Обзор ведущих моделей зрелости

В 2024–2026 годах консалтинг, академия и технологические вендоры предложили ряд моделей ИИ-зрелости, которые при внешних различиях демонстрируют удивительную конвергенцию. Модель зрелости для практика — это инструмент самодиагностики и коммуникации с руководством: она позволяет ответить на вопросы «где мы», «что дальше» и «почему инвестиции в ИИ пока не видны в отчётности», не изобретая терминологию заново.

Gartner — AI Maturity Model (5 уровней). Стадии: Foundational (ad hoc эксперименты с

28 MIT NANDA via agentmodeai / Trullion / pathopt — <https://agentmodeai.com/the-mit-genai-pilot-failure-claim/> ; <https://trullion.com/blog/why-95-of-ai-projects-fail-and-why-the-5-that-survive-matter/> (2025-08/09)

ограниченной координацией) → Emerging (ранние пилоты, растущий интерес топ-менеджмента) → Operational (ИИ встроен в отдельные процессы, назначены владельцы) → Scaled (ИИ-компетенции развёрнуты кросс-функционально, измеримый ROI) → Transformational (ИИ перестраивает принятие решений, операционную модель и конкурентное преимущество)²⁹. Для каждого перехода между стадиями Gartner предлагает набор практик (AI Roadmap Toolkit): от назначения владельцев и оценки рисков на ранних стадиях до измерения ROI и перестройки операционной модели на поздних. Ключевая эмпирическая находка: зрелость сильно коррелирует с долгоживучестью проектов — только 20% низкозрелых организаций удерживают ИИ-проекты в эксплуатации три года и более против 45% высокозрелых³⁰ (confidence: medium-high).

MIT CISR — Enterprise AI Maturity (4 стадии). Модель (Woerner, Sebastian, Weill; Future Ready Survey, 721 компания, обновление 2025): 1) Experiment and Prepare — 13% организаций; 2) Build Pilots and Capabilities — 23%; 3) Develop AI Ways of Working (индустриализация) — 46%; 4) AI Future Ready — 18%^{31,32}. Показательна динамика: в опросе 2022 года на первые две стадии приходилось 62% компаний — за три года доля «начинающих» сократилась примерно на 25 п.п., то есть рынок в целом движется по шкале зрелости вверх, хотя и с разной скоростью. Финансовые результаты по стадиям различаются радикально: компании стадии 4 опережают отрасль на +13,9 п.п. по росту выручки и +9,9 п.п. по прибыли; компании стадии 1 отстают на −26,5 п.п. и −15,1 п.п. соответственно. Самый сложный переход — не из стадии 1 в 2, а **из стадии 2 («пилоты») в стадию 3 («новые способы работы»);** MIT CISR выделяет четыре задачи этого перехода: согласование ИИ-инвестиций с целями, модульные платформы, создание AI-ready команд и синхронизация редизайна работы с возможностями ИИ³³.

BCG — AI Value Gap (3 категории). Прагматичная кластеризация: 5% «future-built», 35% «scaling», 60% «laggards»; динамика 2024 → 2025 — доля масштабирующихся выросла с 22% до 35%, стагнирующих сократилась с 25% до 14%³⁴. Главный предиктор зрелости по BCG — вовлечённость C-level: ~100% «future-built» компаний имеют глубоко вовлечённое руководство против 8% у отстающих; лидеры втрое чаще назначают Chief AI Officer³⁵.

29 Gartner — AI Maturity Model and AI Roadmap Toolkit — <https://www.gartner.com/en/chief-information-officer/research/ai-maturity-model-toolkit> (2026 (актуализировано))

30 Gartner research via puneetsinghal.com — <https://puneetsinghal.com/blog/gartner-ai-maturity-model> (2025-12-03)

31 MIT CISR research via Larridin; orgtp.com — <https://larridin.com/blog/ai-maturity-measurement> ; <https://orgtp.com/blog/the-mit-cisr-enterprise-ai-maturity-model> (2025-2026 (публикации анализов))

32 MIT CISR via orgtp.com; подтверждение MDPI review — <https://orgtp.com/blog/the-mit-cisr-enterprise-ai-maturity-model> ; <https://www.mdpi.com/2078-2489/17/7/646> (2026-06-21 (анализ); briefing MIT CISR — 2023-2025)

33 MIT CISR via orgtp.com; подтверждение MDPI review — <https://orgtp.com/blog/the-mit-cisr-enterprise-ai-maturity-model> ; <https://www.mdpi.com/2078-2489/17/7/646> (2026-06-21 (анализ); briefing MIT CISR — 2023-2025)

34 BCG report (анализ Libertify; подтверждение SDXExec) — <https://www.libertify.com/interactive-library/ai-value-gap-enterprise-strategy-bcg-2025/> ; <https://www.sdcexec.com/software-technology/ai-ar/news/22951486/> (2025-09/10)

35 BCG report (анализ Libertify; подтверждение SDXExec) — <https://www.libertify.com/interactive-library/ai-value-gap-enterprise-strategy-bcg-2025/> ; <https://www.sdcexec.com/software-technology/ai-ar/news/22951486/> (2025-09/10)

Microsoft — две модели. Inside Track: пятистадийная модель от «Awareness and foundation» (видение, спонсорство, AI Center of Excellence, принципы Responsible AI, готовность данных — «no AI without data») до агентных взаимосвязанных систем ³⁶. Характерно, что на первой стадии модель Microsoft требует не технологических инвестиций, а управленческих: смелое видение, раннее привлечение спонсоров-руководителей, формирование центра компетенций и оценку готовности данных. Отдельная Agentic AI Adoption Maturity Model описывает путь именно агентных внедрений: уровни 100 (Initial — незапланированные эксперименты) → 200 (Repeatable) → 300 (Defined) → 400 (Capable) → 500 (Efficient — «agent-first enterprise», организация, где агенты — первичный интерфейс работы) по пяти capability-пилонам ³⁷. Само выделение отдельной агентной модели зрелости в 2026 году — симптом того, что агентный ИИ стал самостоятельным треком трансформации.

Accenture — «The Art of AI Maturity» (2022, базовая модель). Выборка ~1 200 компаний: только 12% — «AI Achievers» (на развивающихся рынках — 17%), 63% — «AI Experimenters». Achievers в 3,5 раза чаще имеют более 30% выручки под влиянием ИИ, на 25% чаще масштабируют пилоты в продакшен, у 83% есть спонсорство CEO/топ-менеджмента. Прогноз Accenture: ИИ-трансформация произойдёт на 16 месяцев быстрее цифровой ³⁸. Хотя исследование датируется 2022 годом, его кластерная логика (Achievers / Builders / Innovators / Experimenters) остаётся базовым языком отрасли и хорошо согласуется с более поздними 5% «future-built» BCG и 6% «high performers» McKinsey: верхний слой зрелости устойчиво оценивается в диапазоне 5–12% в зависимости от строгости критерия.

SEI (CMU) + Accenture — AI Adoption Maturity Model (июнь 2026). Наиболее детальная рамка периода: 8 измерений (организационная стратегия, кадры и культура, реинжиниринг workflow, риск и говернанс, данные, инжиниринг, операции, экосистема) × 5 уровней (Exploratory → Implemented → Aligned → Scaled → Future-Ready, где верхний уровень означает «стабильно воспроизводимые и масштабируемые ИИ-инициативы, дающие предсказуемые инновации») ³⁹. Ценность модели — в мультидименсиональности: она делает явным, что компания может быть зрелой в инжиниринге и незрелой в говернансе, и что «бутылочное горлышко» у каждой организации своё.

Сводная характеристика моделей приведена в таблице:

Модель	Шкала	Число уровней	Ключевой акцент
Gartner AI Maturity Model	Foundational → Transformational	5	Корреляция зрелости с выживаемостью проектов (20% vs

36 Microsoft Inside Track blog — <https://www.microsoft.com/insidetrack/blog/enterprise-ai-maturity-in-five-steps-our-guide-for-it-leaders/> (2026-08-15 (дата индексации; публикация 2025-2026))

37 Microsoft Learn — <https://learn.microsoft.com/en-us/agents/adoption-maturity-model/> (2026-05-20)

38 Accenture Research PDF (The Art of AI Maturity — Growth Markets) — <https://www.accenture.com/content/dam/system-files/acom/custom-code/ai-maturity/Accenture-Art-AI-Maturity-GM.pdf> (2022 (отчёт))

39 CMU SEI news — <https://www.sei.cmu.edu/news/sei-and-accenture-release-ai-adoption-maturity-model-to-help-organizations-scale-ai-with-predictable-outcomes/> (2026-06-08)

Модель	Шкала	Число уровней	Ключевой акцент
MIT CISR Enterprise AI Maturity	Experiment & Prepare → AI Future Ready	4	45% за 3 года) Финансовый разрыв между стадиями; переход 2 → 3 как самый трудный
BCG AI Value Gap	Laggards / Scaling / Future-built	3 категории	Вовлечённость C-level как главный предиктор; правило 10-20-70
Microsoft Inside Track / Agentic	Awareness → агентные системы; 100 → 500	5 + 5	Данные и Responsible AI как фундамент; агентный трек
Accenture Art of AI Maturity	Experimenters → Achievers	4 кластера	12% Achievers; спонсорство CEO у 83% лидеров
SEI (CMU) + Accenture	Exploratory → Future-Ready	5 × 8 измерений	Предсказуемость результатов; мультидизименсиональ ность

2.2. Конвергентная дорожная карта

Все ведущие модели (Gartner, MIT CISR, McKinsey, Microsoft, Cohere) сходятся в типовой последовательности фаз ⁴⁰:

Этап	Содержание	Ключевые артефакты
1. Подготовка	Стратегия, спонсорство, аудит данных, ИИ-грамотность	Видение, спонсор C-level, AI CoE, оценка готовности данных
2. Эксперименты/пилоты	Изолированные PoC	Портфель пилотов, критерии успеха
3. Операционализация	Внедрение в отдельные процессы	Владельцы, метрики, говернанс, платформы/MLOps
4. Масштабирование	Кросс-функциональное развёртывание	Измеримый ROI, переиспользуемые

40 синтез Gartner / MIT CISR / McKinsey / Microsoft / Cohere (Enterprise AI Maturity Model: Experimentation → Tool adoption → Internal platforms → Strategic integrations → AI-native transformation) — <https://cohere.com/blog/enterprise-ai-maturity-model> ; <https://larridin.com/blog/ai-maturity-measurement> (2025-2026)

Этап	Содержание	Ключевые артефакты платформы
5. Трансформация	ИИ как ядро бизнес-модели	Перестроенные процессы, новые операционные модели

Критическое «бутылочное горлышко» всех исследований — переход **от пилотов (стадия 2) к масштабированию и новым способам работы (стадия 3)**. Cohere формулирует это как: «большинство компаний застревают между фазой Tool Adoption и фазой Internal Platforms, не сумев по-настоящему интегрировать ИИ в основной технологический стек»⁴¹.

Рассмотрим содержание каждой фазы подробнее.

Фаза 1. Подготовка. Задача — создать управленческий и информационный фундамент: сформулировать видение, привлечь спонсора на уровне СЕО/совета директоров, учредить центр компетенций (AI CoE), принять принципы ответственного ИИ и — критически — провести аудит готовности данных. Формула Microsoft «no AI without data»⁴² означает, что компании с разрозненными, неочищенными данными не должны начинать с моделей, а должны начинать с данных. Именно дефекты этого фундамента Gartner называет первой причиной отказа от проектов после PoC⁴³. Сюда же относится базовая ИИ-грамотность персонала: без неё последующие фазы упираются в человеческий фактор — главный барьер по всем исследованиям 2024–2026 годов.

Фаза 2. Эксперименты и пилоты. Изолированные проверки гипотез с заранее зафиксированными критериями успеха и базовой линией. Главная ошибка фазы — «пилоты ради пилотов» без критериев продолжения/остановки: по S&P Global, 46% PoC отбраковываются до продакшена⁴⁴, и это нормально, если отбраковка — осознанное решение, а не тихое забвение.

Фаза 3. Операционализация. Самая недооценённая фаза: пилот превращается в часть реального процесса. Появляются назначенные владельцы результата, метрики в операционном контуре, правила говернанса (доступы, риски, качество), платформенная инфраструктура и MLOps. Именно здесь, по Gartner, до продакшена доходят лишь 48% проектов, а путь от прототипа занимает в среднем ~8 месяцев⁴⁵.

41 синтез Gartner / MIT CISR / McKinsey / Microsoft / Cohere (Enterprise AI Maturity Model: Experimentation → Tool adoption → Internal platforms → Strategic integrations → AI-native transformation) — <https://cohere.com/blog/enterprise-ai-maturity-model> ; <https://larridin.com/blog/ai-maturity-measurement> (2025-2026)

42 Microsoft Inside Track blog — <https://www.microsoft.com/insidetrack/blog/enterprise-ai-maturity-in-five-steps-our-guide-for-it-leaders/> (2026-08-15 (дата индексации; публикация 2025-2026))

43 Gartner Press Release (перепечатка HPCwire / Businessworld / IntelligentCIO) — <https://www.hpcwire.com/bigdatawire/this-just-in/gartner-predicts-30-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025/> ; <https://www.intelligentcio.com/me/2024/08/05/gartner-predicts-30-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025/> (2024-07-29/30)

44 S&P Global Market Intelligence (март 2025); оценки глобальных ИИ-инвестиций BCG (\$252 млрд, 2024) и Gartner (~\$1,5 трлн, 2025) — <https://www.digitalapplied.com/blog/agent-ai-project-cancellations-gartner-40-percent-2026> (2025-03-14); см. также <https://talyx.ai/insights/enterprise-ai-implementation-failure>

45 Gartner via Informatica blog — <https://www.informatica.com/blogs/the-surprising-reason-most-ai-projects-fail->

Фаза 4. Масштабирование. Кросс-функциональное развёртывание на переиспользуемых платформах с измеримым ROI. Разница с фазой 3 принципиальна: не «ещё один процесс», а «воспроизводимая машина внедрений», где каждый следующий кейс запускается быстрее и дешевле предыдущего за счёт общих данных, компонентов и говернанса. По McKinsey, до этой фазы дошли ~31% организаций; у компаний с выручкой >\$5 млрд — около 50%, у компаний <\$100 млн — 29% ⁴⁶. Четыре задачи перехода сюда из фазы пилотов по MIT CISR: согласование ИИ-инвестиций с бизнес-целями, архитектура модульных платформ, создание AI-ready команд и синхронизация редизайна работы с возможностями ИИ ⁴⁷.

Фаза 5. Трансформация. ИИ становится ядром бизнес-модели: перестраиваются принятие решений, операционная модель и источники конкурентного преимущества. Это удел 5–7% организаций: «future-built» по BCG (5%) и компаний с полноценным внедрением на уровне предприятия по McKinsey (7%) ^{48,49}.

Доказательством ценности полного прохождения дорожной карты служит «эффект выживаемости»: 45% высокореальных организаций удерживают ИИ-проекты в эксплуатации три года и дольше против 20% низкореальных ⁵⁰ — то есть зрелость конвертируется не только в эффект, но и в устойчивость эффекта.

2.3. Факторы успеха масштабирования

Конвергентные факторы, подтверждённые несколькими независимыми исследованиями:

1. **Вовлечённость CEO/C-level.** McKinsey: high performers в 3 раза чаще имеют сильную вовлечённость топ-руководства; BCG: ~100% лидеров против 8% отстающих; Accenture: 83% Achievers имеют спонсорство CEO ^{51,52,53}.

and-how-to-avoid-it-at-your-enterprise.html (2025-02-11)

46 McKinsey State of AI 2025 (анализ Libertify / SemanticOS / TAdviser) — <https://www.libertify.com/ar/interactive-library/mckinsey-state-of-ai-2025/> ;

<https://semantics.io/blog/mckinsey-state-of-ai-2025-agents> (2025-11 (публикация отчёта))

47 MIT CISR via orgtp.com; подтверждение MDPI review — <https://orgtp.com/blog/the-mit-cisr-enterprise-ai-maturity-model> ; <https://www.mdpi.com/2078-2489/17/7/646> (2026-06-21 (анализ)); briefing MIT CISR — 2023-2025)

48 BCG report (анализ Libertify; подтверждение SDXExec) — <https://www.libertify.com/interactive-library/ai-value-gap-enterprise-strategy-bcg-2025/> ; <https://www.sdcexec.com/software-technology/ai-ar/news/22951486/> (2025-09/10)

49 TAdviser, «Внедрение ИИ в бизнесе» (пересказ McKinsey Global Survey 2025) — https://www.tadviser.ru/index.php/Статья:Внедрение_ИИ_в_бизнесе (2025 (обновлено 2026-07-30))

50 Gartner research via puneetsinghal.com — <https://puneetsinghal.com/blog/gartner-ai-maturity-model> (2025-12-03)

51 McKinsey State of AI 2025 (анализ Libertify / SemanticOS / TAdviser) — <https://www.libertify.com/ar/interactive-library/mckinsey-state-of-ai-2025/> ; <https://semantics.io/blog/mckinsey-state-of-ai-2025-agents> (2025-11 (публикация отчёта))

52 BCG report (анализ Libertify; подтверждение SDXExec) — <https://www.libertify.com/interactive-library/ai-value-gap-enterprise-strategy-bcg-2025/> ; <https://www.sdcexec.com/software-technology/ai-ar/news/22951486/> (2025-09/10)

53 Accenture Research PDF (The Art of AI Maturity — Growth Markets) — <https://www.accenture.com/content/dam/system-files/acom/custom-code/ai-maturity/Accenture-Art-AI->

2. **Редизайн процессов, а не «плагин» ИИ.** 55% high performers фундаментально перестроили процессы под ИИ — почти втрое чаще остальных (~20%) ⁵⁴.
3. **Измерение ценности.** У лидеров BCG системное измерение ценности практикуют более 60% против 17% у отстающих ⁵⁵.
4. **Buy-over-build.** Вендорские специализированные внедрения успешны в ~67% случаев против ~33% у внутренних разработок (MIT NANDA; confidence: medium-high, отчёт не peer-reviewed) ⁵⁶.
5. **ИИ-грамотность и апскиллинг.** У лидеров BCG обучены более 50% сотрудников против ~20% у отстающих ⁵⁷.

Заметим, что все пять факторов — управленческие. Ни один не связан с выбором конкретной модели, вендора или архитектуры. Это эмпирическое подтверждение правила 10-20-70: компании, инвестирующие 70% усилий в алгоритмы и инструменты (типичная ошибка незрелых организаций), оптимизируют те 10–20%, которые меньше всего влияют на результат ⁵⁸. Практическая импликация для планирования ИИ-программы: бюджет и внимание руководства должны распределяться обратно пропорционально интуиции — большая часть ресурсов идёт в редизайн процессов, обучение, change management и систему измерения, а не в закупку технологий.

Дополнительный наблюдаемый паттерн 2024–2026 годов — конвергенция факторов успеха с факторами цифровой трансформации десятилетней давности (спонсорство, процессы, говернанс, таланты), но на сжатом горизонте: Accenture прогнозирует, что ИИ-трансформация произойдёт на 16 месяцев быстрее цифровой ⁵⁹, то есть у компаний меньше времени на прохождение тех же стадий зрелости.

2.4. Агентный ИИ как фронтир 2026

Следующий рубеж — агентные системы (ИИ-агенты, способные самостоятельно планировать и выполнять многошаговые задачи). McKinsey (2025): 62% организаций экспериментируют с ИИ-агентами, 23% масштабируют их хотя бы в одной функции, но ни в одной функции доля реально масштабированных агентных решений не превышает ~10%

Maturity-GM.pdf (2022 (отчёт))

54 McKinsey State of AI 2025 (анализ Libertify / SemanticOS / TAdviser) —

<https://www.libertify.com/ar/interactive-library/mckinsey-state-of-ai-2025/> ;

<https://semantics.io/blog/mckinsey-state-of-ai-2025-agents> (2025-11 (публикация отчёта))

55 BCG report (анализ Libertify; подтверждение SDXExec) — <https://www.libertify.com/interactive-library/ai-value-gap-enterprise-strategy-bcg-2025/> ; <https://www.sdcexec.com/software-technology/ai-ar/news/22951486/> (2025-09/10)

56 MIT NANDA via agentmodeai / Trullion / pathopt — <https://agentmodeai.com/the-mit-genai-pilot-failure-claim/> ; <https://trullion.com/blog/why-95-of-ai-projects-fail-and-why-the-5-that-survive-matter/> (2025-08/09)

57 BCG report (анализ Libertify; подтверждение SDXExec) — <https://www.libertify.com/interactive-library/ai-value-gap-enterprise-strategy-bcg-2025/> ; <https://www.sdcexec.com/software-technology/ai-ar/news/22951486/> (2025-09/10)

58 BCG via USDM blog — <https://www.usdm.com/resources/blogs/why-ai-adoption-is-failing-in-life-sciences> (2025-12-30)

59 Accenture Research PDF (The Art of AI Maturity — Growth Markets) —

<https://www.accenture.com/content/dam/system-files/acom/custom-code/ai-maturity/Accenture-Art-AI-Maturity-GM.pdf> (2022 (отчёт))

⁶⁰. Deloitte (декабрь 2025): только 11% организаций имеют production-ready агентные системы, 42% не имеют формальной стратегии агентного ИИ; Gartner CIO Survey 2026: агентов развернули 17% организаций ⁶¹. При этом ожидания рынка высоки: Gartner прогнозирует, что 40% корпоративных приложений будут включать задачевых агентов к концу 2026 года.

При этом Gartner прогнозирует: более 40% агентных ИИ-проектов будут отменены к концу 2027 года из-за эскалации затрат, неясной ценности и недостаточного контроля рисков; из тысяч «агентных» вендоров реально агентными Gartner считает лишь ~130 (явление «agent washing») ⁶². Агентный ИИ, таким образом, повторяет траекторию GenAI 2024–2025 годов: массовый энтузиазм на входе, жёсткий фильтр на выходе. Вывод для практики: переходить к агентным архитектурам следует только после прохождения стадии операционализации и при наличии централизованного реестра и идентичностей агентов (см. кейс Walmart в главе 3).

2.5. Ограничения моделей зрелости

К моделям зрелости следует относиться как к навигационным рамкам, а не как к точным измерительным приборам. Во-первых, распределения по стадиям различаются между моделями (MIT CISR относит 46% организаций к стадии индустриализации, тогда как McKinsey фиксирует масштабирование лишь у 31%, а BCG — у 35%) из-за разных опросных формулировок и выборок; прямое сопоставление некорректно. Во-вторых, все модели построены на самоотчётах руководителей, систематически завышающих зрелость своих организаций: сотрудники, по McKinsey, используют ИИ втрое чаще, чем оценивают руководители, но и глубина этого использования нередко ограничивается личной продуктивностью. В-третьих, часть данных поступает из вторичных пересказов платных первоисточников (что отражено в уровнях confidence соответствующих утверждений). Тем не менее конвергенция пяти независимых моделей на одной и той же логике фаз и одним и том же бутылочном горлышке — само по себе сильный сигнал: структура проблемы внедрения ИИ воспроизводится независимо от методологии.

2.6. Выводы главы

1. Модели зрелости Gartner, MIT CISR, BCG, Microsoft, Accenture и SEI/Accenture при разной терминологии описывают один и тот же путь: **подготовка** → **пилоты** → **операционализация** → **масштабирование** → **трансформация**. Эта конвергенция независимых методологий — самый надёжный факт данной главы.
2. Бутылочное горлышко локализовано на переходе **пилот** → **масштаб**: здесь

⁶⁰ McKinsey State of AI 2025 (анализ Libertify / SemanticOS) —

<https://www.libertify.com/interactive-library/state-of-ai-2025-mckinsey-report/> (2025-11)

⁶¹ Gartner press release via digitalapplied.com analysis — <https://www.digitalapplied.com/blog/agentic-ai-project-cancellations-gartner-40-percent-2026> (прогноз Gartner — 2025-06-25; анализ — 2026-07-11)

⁶² Gartner press release via digitalapplied.com analysis — <https://www.digitalapplied.com/blog/agentic-ai-project-cancellations-gartner-40-percent-2026> (прогноз Gartner — 2025-06-25; анализ — 2026-07-11)

застревают две трети организаций, здесь же проходят фильтры Gartner (48% доходят до продакшена) и MIT NANDA (5% до промышленной эксплуатации).

3. Факторы успеха конвергентны и управленчески: спонсорство CEO, редизайн процессов, системное измерение, покупка вместо самостоятельной разработки, массовое обучение. Распределение усилий — по правилу 10-20-70.
4. Зрелость измеримо окупается: разрыв между верхней и нижней стадиями — от +13,9 п.п. к росту выручки (MIT CISR) до 1,7× выручки и 3,6× TSR (BCG) — и проявляется в выживаемости самих ИИ-проектов (45% vs 20% за три года).
5. Агентный ИИ — фронт 2026 года, но и главная риск-зона: прогноз Gartner о >40% отменённых агентных проектов к 2027 году делает дисциплину зрелости ещё более актуальной, чем для классического GenAI.

3. Подходы и схемы внедрения по размеру бизнеса

Размер компании определяет не ответ на вопрос «использовать ли ИИ», а архитектуру доступа, модель внедрения и говернанс. Один и тот же инструмент (например, корпоративный чат с LLM) даёт малому бизнесу три фиксированные роли, а крупному — кастомный контроль доступа, аудит и управление идентичностями агентов; один и тот же сценарий (чат-бот поддержки) малому бизнесу обходится в \$29–79 в месяц, а крупному — в проект за миллионы долларов с интеграциями, говернансом и change management. Ниже — специфика трёх сегментов: малый и средний бизнес (раздел 3.1), крупный бизнес (раздел 3.2) и сводное сравнение (раздел 3.3).

3.1. Малый и средний бизнес

3.1.1. Статистика adoption: два ряда данных

Оценки проникновения ИИ среди SMB сильно различаются из-за разных определений «использования» и выборок; оба ряда приводятся с указанием природы источника.

Самоотчётные отраслевые опросы (оптимистичный ряд):

- США: 58% малых бизнесов используют генеративный ИИ в 2025 году против 40% в 2024 и 23% в 2023 — самое быстрое внедрение технологии со времён соцсетей (U.S. Chamber of Commerce, «Empowering Small Business Report», август 2025).
- США: использование выросло с 39% (2024) до 55% (2025), +41% г/г; среди компаний 10–100 сотрудников — с 47% до 68%; в 2026 году — уже 66% (Thryv, 540–561 респондент).
- Intuit QuickBooks: 78% малых бизнесов США регулярно используют ИИ (апрель 2026) против 48% в июле 2024; ежедневное использование выросло с 15% до 40%.
- Великобритания: 35% МСП активно используют ИИ в 2025 (25% в 2024), к марту 2026 — 54% (British Chambers of Commerce).
- Salesforce (n=2 500 руководителей SMB глобально): 75% лидеров SMB инвестируют в ИИ, 34% полностью внедрили его в операции; растущие SMB в 1,8 раза чаще инвестируют, чем сокращающиеся; 91% использующих сообщают о

росте выручки.

Строгие государственные замеры (консервативный ряд):

- США (Census Bureau BTOS / SBA Office of Advocacy, сентябрь 2025): 8,8% малых бизнесов (<250 сотрудников) использовали ИИ в производстве товаров и услуг на август 2025 (6,3% полугодом ранее); по всем функциям — 17–20%. При этом разрыв с крупными фирмами сократился с 1,8× (февраль 2024) до 1,2× (август 2025), поскольку adoption у крупных вышел на плато.
- ЕС (Eurostat, 2025): ИИ-технологии используют 20,0% предприятий с 10+ сотрудниками (8% в 2023); крупные — 55,0%, средние — 30,4%, малые — 17,0%. Лидеры — Дания (42%), Финляндия (37,8%), Швеция (35%).
- OECD (доклад к G7-2025): в 2024 году ИИ использовали 40% фирм с 250+ сотрудниками, 20,4% средних (50–249) и лишь 11,9% малых (10–49) — разрыв 3,4×, шире, чем для облака и IoT.
- Россия: 45% представителей малого бизнеса заявляют о применении ИИ (итоги федпроекта, 2024), однако по НИУ ВШЭ (2025) доля малых предприятий в объёме реализованных ИИ-инноваций — лишь 1,2% (66,3% приходится на крупные организации) — разрыв между декларацией и глубиной (confidence: medium-high).

Дополнительные региональные ориентиры: Канада — 45% малых бизнесов используют генИИ хотя бы для части задач (CFIB, февраль 2026), а строгий замер StatCan показывает рост использования ИИ в производстве с 6,1% (Q2 2024) до 19,2% (Q2 2026); Финляндия (SMB Barometer 2025, 4 553 компании) — только 12% МСП используют ИИ регулярно, 51% не используют вообще, 46% не видят релевантности ИИ своему бизнесу в ближайший год. Эти два примера хорошо иллюстрируют диапазон неопределённости: в зависимости от страны, определения и метода замера adoption SMB в 2025–2026 годах оценивается от ~12% до ~78%.

Вывод: SMB сокращают разрыв с enterprise по темпам, но по строгим замерам глубина использования всё ещё в ~3 раза ниже.

3.1.2. Типичные сценарии

Топ-3 сценария у малых бизнесов — маркетинг, клиентский сервис, административная работа; маркетинговая автоматизация — единственная область, где малые фирмы чаще обгоняют крупных по adoption (Intuit QuickBooks, SBA, U.S. Chamber, 2025). Среди использующих: 62% — анализ данных, 55% — генерация контента, 46% — customer engagement (чат-боты, автоответчики) (Thryv 2025). В ЕС среди использующих ИИ: 34,7% применяют его в маркетинге и продажах (сценарий №1), 31,1% — в административных процессах, 23,1% — в бухгалтерии/финансах (Eurostat 2025). ИИ-чат-боты стали вторым по распространённости технологическим инструментом малого бизнеса после поисковиков; HubSpot оценивает, что чат-боты закрывают до 70% клиентских запросов без оператора (confidence: medium).

Важная оговорка о глубине: половина работников малых бизнесов США используют ИИ, но 64% — для личной продуктивности (черновики, суммаризация), 26% — для

повторяющихся задач и только 6% — для автоматизации рабочих процессов (U.S. Chamber Foundation + Ipsos). Это «широкое, но мелкое» внедрение.

Структура спроса по сценариям устойчива между странами. В США: 54% малых бизнесов уже используют ИИ-маркетинговые инструменты, ещё 27% планируют в течение 12 месяцев; 42% используют генИИ-чат-ботов (разброс по штатам 13–71%); 77% SMB называют маркетинг и клиентский engagement приоритетом №1 для ИИ; рекрутинг применяют 19% использующих (Thryv, Salesforce, SBA, 2024–2025). В ЕС картина аналогична: маркетинг и продажи (34,7% использующих) → административные процессы и управление (31,1%) → бухгалтерия и финансы (23,1%) → производство (20,8%) → ИБ (19,8%) → логистика (6,1%), причём разрыв малых и крупных предприятий минимален именно в маркетинге и продажах (Eurostat 2025). Практический смысл: маркетинговые и клиентские сценарии — «низко висящий фрукт», требующий минимальной интеграции и данных, поэтому именно они и становятся точкой входа для компании без ИИ-команды.

3.1.3. Стек и стоимость: \$10–100 в месяц

Порог входа низкий. Типичный стартовый бюджет SMB — \$10–100/мес (53% SMB по Thryv 2026 тратят уже ≥\$100/мес). Ориентиры цен (2025–2026):

Категория	Инструменты	Цена от
Универсальный ассистент	ChatGPT	\$20/мес
Маркетинговый контент	Jasper	\$59–69/мес
Автоматизация связей	Zapier	\$29,99/мес
Транскрибация	Otter.ai	\$16,99/мес
Чат-боты поддержки	Tidio (Starter/Growth с Lyro AI)	\$29–59/мес
Поддержка (per-resolution)	Intercom Fin	\$0,99 за решённый запрос
CRM с ИИ	HubSpot Starter (Breeze Copilot)	\$15–20/место/мес

Серьёзная платформа уровня HubSpot Marketing+Sales Hub Professional для бизнеса 10–50 сотрудников — \$620–890/мес; реальный счёт 5-местной команды поддержки с полным ИИ: Tidio ~\$79/мес, Zendesk ~\$345/мес, Intercom \$250–495/мес (confidence: medium — часть цен из вторичных обзоров).

Экономика: 58% пользователей экономят >20 часов в месяц, 66% — \$500–2 000 в месяц (Thryv 2025); владельцы и менеджеры экономят в среднем ~7 часов в неделю; 70% пользователей сообщают о росте выручки, 55% — о снижении затрат, 92% говорят, что ИИ экономит время (Thryv 2026); канадские фирмы получают \$1,60 на каждый вложенный доллар, 55% выходят в плюс за 2 года (CFIB); встречающаяся оценка среднего ROI 3,7× (приписывается McKinsey) — confidence: medium/low, вторичный агрегатор. По Salesforce, 91% SMB, использующих ИИ, сообщают о росте выручки и 90% — о росте эффективности, причём растущие компании инвестируют в ИИ в 1,8 раза чаще сокращающихся — наблюдается двусторонняя причинность: ИИ поддерживает рост, а растущие компании имеют ресурс для экспериментов.

Заметим, что заявляемые эффекты SMB — это в основном самоотчёты владельцев, не верифицированные независимым измерением; однако порядок величин (десятки часов и сотни долларов в месяц на компанию) согласуется между несколькими независимыми опросами (Thryv, CFIB, Salesforce, QuickBooks), что повышает доверие к направлению эффекта, если не к точным цифрам.

3.1.4. Схема внедрения без ИИ-команды

Конвергентная практическая схема (синтез 90-дневного плана BizTech/Dario Markovic, 6-недельного roadmap LeadTruffle и рекомендаций Eaton & Associates; confidence: medium):

1. **Аудит (недели 1–2, дни 1–30 по 90-дневному плану).** Инвентаризация процессов, матрица «влияние/усилия», выбор одного высокочастотного низкосложного процесса (поддержка, follow-up маркетинга, счета), фиксация базовой линии метрик. Уже на этом шаге желательно заручиться спонсорством владельца бизнеса — по аналогии с фактором успеха №1 в enterprise.
2. **Приоритизация и выбор инструмента (дни 31–60).** 3–5 кандидатов под самый болезненный процесс; предпочтение встроенному ИИ в уже используемых SaaS (CRM, helpdesk, почта) — это снимает задачу интеграции, главный технический барьер малого бизнеса.
3. **Пилот (6–12 недель; дни 61–90).** Запуск с параллельным ручным режимом (страховка от деградации качества), обучение команды, сбор обратной связи, оптимизация настроек.
4. **Измерение ROI (недели 9–12).** Сравнение с базовой линией: часы, время ответа, конверсия, стоимость обработки одного запроса. Решение «расширять / доработать / остановить» принимается по цифрам, а не по ощущениям.
5. **Масштабирование.** Стандартные операционные процедуры (SOP), второй сценарий, 90-дневный план расширения, назначение ответственного за развитие ИИ-инструментов.

Принцип: «start small, validate value, then scale». Критично начинать с высокочастотных, низкосложных процессов (поддержка, follow-up маркетинга, выставление счетов) и обязательно фиксировать базовую линию до запуска — иначе ROI будет недоказуем, как у 95% корпоративных пилотов MIT NANDA.

3.1.5. Кейсы малого бизнеса

- **Parkable (Новая Зеландия, парковочный SaaS):** после настройки автоматизированных workflow в Slack и внедрения автономных агентов для типовых запросов конверсия выросла на 38%, время до конверсии сократилось на 50% (Salesforce; confidence: medium-high).
- **E-commerce-магазин:** +30% возвращённых продаж за счёт автоматических писем о брошенных корзинах и персональных рекомендаций; локальный сервисный бизнес — +25% записей за счёт SMS-напоминаний; B2B-агентство — −40% времени ручного follow-up благодаря автоматическому лид-скорингу. Большинство малых бизнесов окупают маркетинговую автоматизацию менее чем за 6 месяцев (CDM

Suite, 2025; confidence: medium — маркетинговый источник).

- **SaaS-стартап из Кембриджа:** ИИ-боты поддержки сократили время первого ответа с 6 часов до менее 1 минуты; консалтинговая фирма из Лидса тратит менее часа в неделю на бухгалтерию благодаря ИИ-буккипину (QuickBooks/Xero) (confidence: low-medium, непроверяемый вторичный кейс).
- **Обобщённый эффект полного workflow:** клиенты консалтинга OperatorLab экономят ~15 часов в неделю в течение 90 дней после установки первого слоя автоматизации; экономия >20 ч/мес характерна именно для автоматизированных повторяющихся процессов, а не для разовой генерации текстов (confidence: medium).

3.1.6. Барьеры и роль MSP/консультантов

Главный барьер — не деньги, а навыки. Eurostat: 70,3% предприятий, рассматривавших ИИ, но не внедривших, называют дефицит компетенций главным препятствием (малые — 70,9%, средние — 69,2%, крупные — 65,1%); далее — правовая неопределённость (53,6%), приватность (52,7%), стоимость (38,4%). Только 12% малых бизнесов инвестируют в ИИ-тренинги; 77% пользователей не имеют формальной стратегии применения; ~половина фирм, использующих ИИ, вложили \$0 в обучение и интеграцию (SBA). Второй барьер — восприятие нерелевантности: 82% микрофирм США (<5 сотрудников) считают ИИ неприменимым к ним; 78,1% канадских не-adopters: «ИИ не релевантен». Стоимость — главный барьер для 58% канадских малых фирм (Sage/CFIB). При этом опасения по ROI и безопасности данных снижаются (–40% г/г по Thryv) — рынок взрослеет, и барьеры смещаются от страхов к дефициту практических навыков.

По данным Reimagine Main Street / PayPal (1 000 малых бизнесов, май–июнь 2025), 76% малых бизнесов используют или изучают ИИ, но только 25% внедрили его в ежедневные операции и ~14% — полностью в ключевые процессы; 51% остаются в фазе «исследователей». Это подтверждает: «пилотное чистилище» — не только enterprise-феномен, но и структурная особенность SMB-сегмента.

Поскольку собственной ИИ-команды у SMB нет, рекомендуемая модель — партнёрство с MSP (managed service providers) и консультантами, ведущими структурированные пилоты с чёткими метриками успеха. Экономика такого партнёрства складывается из трёх факторов: консультант снимает барьер экспертизы (главный барьер №1 по Eurostat), приносит типовые сценарии и шаблоны измерения ROI, а также берёт на себя интеграцию с CRM/ERP — задачу, неподъёмную для малых команд. Для самих MSP ИИ-консалтинг для SMB в 2025–2026 годах стал одним из самых быстрорастущих направлений услуг.

Спрос на поддержку со стороны МСП структурирован: 73% малых бизнесов хотят больше обучающих ресурсов, 74% — более ясных доказательств ROI (Reimagine Main Street / PayPal, май–июнь 2025); практическое обучение — запрос №1. Обучение — также самая запрашиваемая форма господдержки среди МСП в 7 странах (опрос OECD, более 5 000 МСП, ноябрь 2025); существуют программы финансирования (Канада, BDC LIFT: \$500 млн, займы \$25 тыс.–5 млн под 2,25%). При этом 82% малых бизнесов считают внедрение ИИ необходимым для конкурентоспособности — мотивация есть, дефицит — в

инструментах реализации.

Обнадёживающий факт: внедрение редко сокращает штат — 95% британских МСП с ИИ сообщают об отсутствии влияния на численность персонала (ВСС, 2026).

3.2. Крупный бизнес

3.2.1. Специфика enterprise-сегмента

Крупные компании масштабируют ИИ заметно успешнее: ~50% компаний с выручкой >\$5 млрд достигли фазы scaling против 29% у компаний <\$100 млн (McKinsey 2025). Однако и здесь две трети застряли в пилотном режиме, а летом 2025 года крупные компании США (250+ сотрудников) впервые за два года снизили уровень внедрения ИИ в производственные процессы — с 14% до 12% (U.S. Census Bureau; confidence: medium). Это «плато» объясняется не потерей интереса, а фазой консолидации: после волны экспериментов 2023–2024 годов enterprise-сегмент вошёл в период отбраковки неэффективных инициатив (42% компаний свернули большинство ИИ-инициатив по S&P Global ⁶³) и перенастройки портфеля. Бюджеты несопоставимы с SMB: стоимость внедрения GenAI-моделей корпоративного класса — \$5–20 млн (Gartner) ⁶⁴, что при 48-процентной вероятности доведения проекта до продакшена делает управление портфелем ИИ-инициатив финансово значимой дисциплиной.

Отличительные черты enterprise-сегмента: наличие собственных данных и инфраструктуры (одновременно актив и источник долга), регуляторные ограничения (в 2024–2025 регуляторика стала барьером №1 по Deloitte), сложность интеграции с legacy-контуrom, многоуровневый говернанс доступа (RBAC, аудит, идентичности агентов) и — как следствие — более длинные циклы внедрения: 12–18 месяцев до эффекта у отстающих против 9–12 месяцев у лидеров (BCG).

3.2.2. Организационная модель: hub-and-spoke

Для крупных компаний доминирующей оргмоделью внедрения стала «hub-and-spoke»: центральный хаб (AI Center of Excellence) задаёт стандарты, платформы и говернанс, а «спицы» — встроенные в бизнес-подразделения команды — отвечают за внедрение в процессы. Модель снимает главное противоречие enterprise-внедрения: полная централизация убивает скорость и контакт с процессами, полная децентрализация порождает дублирование, риски и «agent sprawl» (кейс Walmart ниже). По данным,

63 S&P Global Market Intelligence (март 2025); оценки глобальных ИИ-инвестиций BCG (\$252 млрд, 2024) и Gartner (~\$1,5 трлн, 2025) — <https://www.digitalapplied.com/blog/agent-ai-project-cancellations-gartner-40-percent-2026> (2025-03-14); см. также <https://talyx.ai/insights/enterprise-ai-implementation-failure>

64 Gartner Press Release (перепечатка HPCwire / Businessworld / IntelligentCIO) — <https://www.hpcwire.com/bigdatawire/this-just-in/gartner-predicts-30-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025/> ; <https://www.intelligentcio.com/me/2024/08/05/gartner-predicts-30-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025/> (2024-07-29/30)

подтверждённым двумя независимыми ветками исследования, эта модель ассоциирована с ростом эффективности внедрений: Deloitte фиксирует +73%, IBM — +36% ROI (confidence: high как направление; точные цифры — из отраслевых сводок). Ключевая новая роль в этой модели — AI product manager, соединяющий бизнес-процесс, данные и возможности моделей. Вспомогательный маркер институционализации: доля организаций с ролью Chief AI Officer выросла с 26% (2025) до прогнозируемых 76% (2026) по IBM (confidence: medium).

Что делают успешные enterprise-лидеры (конвергентные практики BCG и McKinsey):

- концентрируют усилия: меньше инициатив, но глубже — в core-функциях, где создаётся 62% ценности (BCG);
- перестраивают процессы, а не «прикручивают» ИИ к существующим: 55% high performers фундаментально переработали workflow ⁶⁵;
- измеряют ценность системно: >60% лидеров против 17% отстающих ⁶⁶;
- масштабируют быстрее: time-to-impact 9–12 месяцев против 12–18 месяцев у отстающих ⁶⁷;
- обучают персонал массово: >50% сотрудников против ~20% ⁶⁸;
- преследуют трансформационные, а не инкрементальные цели: high performers в 3,6 раза чаще проводят изменения на уровне предприятия ⁶⁹;
- строят внутренние платформы, которые делают каждое следующее внедрение дешевле предыдущего (переиспользуемые данные, компоненты, шаблоны говернанса).

3.2.3. Buy vs build: 67% против 33%

Структурная находка MIT NANDA: проекты с привлечением специализированных вендоров/партнёров успешны в ~67% случаев, внутренние (in-house) разработки — лишь в ~33% (по некоторым интерпретациям ~22%) ⁷⁰ (confidence: medium-high — отчёт не peer-reviewed). Эмпирическое правило для enterprise: покупать специализированные решения для стандартных задач и строить собственные лишь там, где ИИ — источник

65 McKinsey State of AI 2025 (анализ Libertify / SemanticOS / TAdviser) —

<https://www.libertify.com/ar/interactive-library/mckinsey-state-of-ai-2025/> ;

<https://semantics.io/blog/mckinsey-state-of-ai-2025-agents> (2025-11 (публикация отчёта))

66 BCG report (анализ Libertify; подтверждение SDXExec) — <https://www.libertify.com/interactive-library/ai-value-gap-enterprise-strategy-bcg-2025/> ; <https://www.sdcexec.com/software-technology/ai-ar/news/22951486/> (2025-09/10)

67 BCG report (анализ Libertify; подтверждение SDXExec) — <https://www.libertify.com/interactive-library/ai-value-gap-enterprise-strategy-bcg-2025/> ; <https://www.sdcexec.com/software-technology/ai-ar/news/22951486/> (2025-09/10)

68 BCG report (анализ Libertify; подтверждение SDXExec) — <https://www.libertify.com/interactive-library/ai-value-gap-enterprise-strategy-bcg-2025/> ; <https://www.sdcexec.com/software-technology/ai-ar/news/22951486/> (2025-09/10)

69 McKinsey State of AI 2025 (анализ Libertify / SemanticOS / TAdviser) —

<https://www.libertify.com/ar/interactive-library/mckinsey-state-of-ai-2025/> ;

<https://semantics.io/blog/mckinsey-state-of-ai-2025-agents> (2025-11 (публикация отчёта))

70 MIT NANDA via agentmodeai / Trullion / pathopt — <https://agentmodeai.com/the-mit-genai-pilot-failure-claim/> ; <https://trullion.com/blog/why-95-of-ai-projects-fail-and-why-the-5-that-survive-matter/> (2025-08/09)

дифференциации и есть соответствующие компетенции. JPMorgan — пример осознанного «build» при колоссальных ресурсах; большинство компаний такими ресурсами не располагают.

Причины преимущества «buy» объяснимы: специализированный вендор приносит готовое решение, уже встроенное в целевой процесс, с отлаженными интеграциями и накопленной на десятках внедрений экспертизой; внутренняя разработка, напротив, упирается в дефицит талантов, недооценку трудоёмкости MLOps и конкуренцию ИИ-команды за приоритеты с основным ИТ. Критерии выбора «build» в 2024–2026 годах обычно сводятся к трём условиям: (1) задача — источник устойчивого конкурентного преимущества; (2) есть требования к суверенитету данных или регуляторные ограничения, не закрываемые вендором; (3) в компании уже существует зрелая data/ML-функция. Невыполнение хотя бы двух из трёх — аргумент за «buy». Отдельный тренд зрелых enterprise (и массовый тренд в Китае) — собственный стек на открытых моделях; выбор между «встроенным ИИ в SaaS» и «self-hosted open source» определяется компетенциями и требованиями к данным, при этом совокупная стоимость владения открытыми и закрытыми решениями сближается.

3.2.4. Кейсы

JPMorgan Chase. Годовой техбюджет ~\$18 млрд, ~\$2 млрд — на ИИ; платформа LLM Suite развёрнута примерно для 250 000 сотрудников, около половины пользуются ею ежедневно; 450+ ИИ-кейсов в продакшене; заявленный рост финансовых выгод от ИИ — 30–40% в год. Система COiN (запущена ещё в 2017): анализ 12 000 кредитных соглашений в год, экономия 360 000 юрист-часов, –80% ошибок. Урок: масштаб достигается платформенным подходом и многолетними инвестициями в данные и компетенции — JPMorgan строил ИИ-контур почти десятилетие до появления ChatGPT, и LLM Suite легла на готовый фундамент, а не на пустое место (confidence: high по масштабу; цифры — преимущественно самоотчёт компании).

Klarna. ИИ-ассистент (с OpenAI), запуск февраль 2024: за первый месяц 2,3 млн диалогов (~2/3 всех обращений), эквивалент нагрузки 700 FTE, время решения с 11 минут до <2 минут, повторные обращения –25%, оценка \$40 млн прироста прибыли в 2024; к Q3 2025 — эквивалент ~853 FTE и ~\$60 млн экономии. Однако в мае 2025 CEO Себастьян Семятковски признал, что компания «пережала» с сокращением затрат и качество упало — Klarna вернулась к гибридной модели с людьми для сложных кейсов. Урок: агрессивная замена людей без контроля качества приводит к откату; устойчивая модель — гибридная (все цифры — самоотчёт Klarna; «700 агентов» — расчётный эквивалент нагрузки, не увольнения).

Walmart. Стратегия «super agents» (июль 2025): консолидация десятков разрозненных ботов в четыре супер-агента — Sparky (покупатели), агент сотрудников, Marty (партнёры/поставщики), агент разработчиков. Причина — «agent sprawl»: множество одноцелевых агентов не масштабируется на уровне предприятия. Урок: агентная архитектура требует централизованной оркестрации с самого начала.

Siemens. Industrial Copilot (с Microsoft, Azure OpenAI): 100+ клиентов в Европе и США

(Schaeffler, thyssenkrupp Automation Engineering — первый глобальный rollout с начала 2025); более 120 000 инженеров TIA Portal получили GenAI-ассистента; с мая 2025 — переход к ИИ-агентам (Planning/Engineering/Operations/Services Copilot); Maintenance Copilot Senseye: –25% времени реактивного обслуживания в пилотах. Урок: в промышленности ИИ встраивается в существующие инженерные платформы, а не развёртывается «с нуля»; партнёрство с гиперскейлером (Microsoft) решает инфраструктурную задачу, а собственная экспертиза Siemens — задачу предметной ценности. Это классическая buy-and-partner-конфигурация, соответствующая находке MIT NANDA о 67% успеха вендорских внедрений.

Прочие верифицированные кейсы ROI (компиляция, цифры — из публичных заявлений компаний; confidence: medium):

- **Morgan Stanley:** DevGen.AI проанализировал 9 млн+ строк legacy-кода и сэконобил разработчикам ~280 000 часов; в wealth management добровольное принятие ИИ-ассистента достигло 98% — редкий пример почти полного органического adoption без принуждения.
- **Salesforce:** ИИ-агент для контрактов сэконобил >\$5 млн на внешних юристах; внутренний Agentforce обрабатывает ~32 000 диалогов в неделю с 83% resolution rate.
- **General Mills:** оптимизация цепочки поставок (5 000+ отгрузок в день), >\$20 млн экономии с FY2024 — пример ценности в core-функции, а не в поддерживающих.

Средний заявленный ROI агентных внедрений ~171% (в США ~192%); 74% руководителей получили ROI в первый год (confidence: medium — агрегированные маркетинговые сводки, к точным значениям следует относиться осторожно).

Кейсы провалов. Zillow: алгоритмический выкуп недвижимости обернулся >\$500 млн чистых убытков, ликвидацией подразделения и увольнением 25% персонала. Commonwealth Bank of Australia: замена 45 операторов голосовыми ИИ-агентами провалилась — банк извинился и нанял людей обратно (confidence: medium-high).

Сравнение успешных и провальных кейсов выявляет общий паттерн: успешные внедрения (JPMorgan, Siemens, Morgan Stanley, General Mills) объединяют платформенный подход, долгую подготовку данных, встраивание в существующий процесс и сохранение человека в контуре критических решений; провальные (Zillow, CBA, первая фаза Klarna) — попытку заменить ИИ целый процесс или функцию «одним ходом», без гибридного контура и без измеримой базовой линии. Даже показательный «успех» Klarna в 2025 году потребовал коррекции курса в сторону гибридной модели — что подтверждает: устойчивая конфигурация в клиентских функциях на горизонте 2024–2026 годов — «ИИ для типового, человек для сложного».

3.2.5. ROI и метрики

По мета-своду исследований 2025–2026 годов только 5–8% компаний сообщают об измеримой финансовой отдаче в масштабе; IBM CEO study: 25% заявляют о «достигнутом ожидаемом ROI», при этом 56% CEO признают отсутствие значимого финансового

эффекта (confidence: medium-high). Deloitte (Q4 2024): большинство компаний ожидают, что менее 30% ИИ-экспериментов достигнут полного масштаба за 6 месяцев; при этом 78% планируют наращивать инвестиции, а регуляторика стала барьером №1 (28% → 38% между волнами опроса). Важный контраст: почти все организации, сумевшие масштабировать инициативы, видят измеримую отдачу — лидируют ИТ (28%) и кибербезопасность (44% превысили ROI-ожидания) — то есть проблема не в отсутствии ценности как таковой, а в способности довести инициативу до масштаба, где ценность становится измеримой.

Контраст заявлений и реальности объясняется расщеплением метрик: на уровне отдельных команд и кейсов эффекты фиксируются часто (экономия часов, ускорение циклов), но на уровне предприятия они растворяются — локальная экономия времени не конвертируется в P&L без редизайна процесса (перераспределения нагрузки, изменения норм, отказа от аутсорсинга). Именно поэтому McKinsey называет переработку workflow «рычагом №1», а MIT NANDA фиксирует: у 95% пилотов эффект не измерим на уровне P&L ⁷¹⁷².

Ключ к измеримому ROI — наличие базовой линии до внедрения. Быстрее всего ROI доказуем в клиентском сервисе (АХТ, cost per resolution, repeat contact rate — недели); дольше всего — в supply chain и ERP-интеграциях (12+ месяцев). Парадокс распределения бюджетов по MIT NANDA: 50–70% GenAI-бюджетов идёт в продажи и маркетинг, тогда как наибольший ROI — в автоматизации бэк-офиса. BCG добавляет: 62% ценности ИИ создаётся в основных (core) бизнес-функциях, а не в поддерживающих.

Типология метрик корпоративного ИИ включает четыре слоя (Master of Code, мета-обзор 18 кейсов и 16 бенчмарков; confidence: medium):

Слой метрик	Примеры	Горизонт доказуемости
Финансовые	Рост выручки, вклад в EBIT, ROIC, экономия затрат	6–18 месяцев
Операционные	Сэкономленные часы, cycle time, error rate, стоимость обработки	Недели — месяцы
Клиентские	NPS, конверсия, время решения, repeat contact rate	Недели (клиентский сервис — самый быстрый ROI)
Метрики внедрения	Доля масштабированных процессов, время pilot → value, adoption rate, отношение продакшен-кейсов к пилотам	Непрерывно

71 McKinsey State of AI 2025 (анализ Libertify / SemanticOS / TAdviser) — <https://www.libertify.com/ar/interactive-library/mckinsey-state-of-ai-2025/> ; <https://semantics.io/blog/mckinsey-state-of-ai-2025-agents> (2025-11 (публикация отчёта))

72 MIT NANDA report (анализы Trullion, agentmodeai, ontheground) — <https://agentmodeai.com/the-mit-genai-pilot-failure-claim/> ; <https://ontheground.agency/resources/why-ai-projects-fail-mit-2025-genai-divide> (2025-08 (отчёт); Fortune-материал 2025-08-21)

Практический принцип: скорость доказуемости ROI напрямую коррелирует с тем, насколько измеримой была базовая линия до внедрения (MIT NANDA). Поэтому зрелые программы начинаются не с выбора модели, а с выбора процесса, у которого уже есть честная «до»-метрика.

3.2.6. Change management: решающие 70%

Исследования 2025–2026 сходятся: главный барьер масштабирования — не технологии и даже не сопротивление сотрудников, а инерция и рассогласованность руководства (McKinsey, AI in the Workplace 2025). Сотрудники используют ИИ втрое чаще, чем оценивают руководители; 48% сотрудников назвали формальное обучение главным стимулом использования, но почти половина получили минимальное обучение или не получили его вовсе. BCG AI at Work 2025 (10 635 сотрудников, 11 стран): только 36% считают обучение достаточным; использование фронтлайн-работниками застопорилось на 51% («silicon ceiling»); 54% готовы применять неавторизованные ИИ-инструменты (Shadow AI; у Gen Z и миллениалов — 62%); при этом сильная поддержка руководства поднимает позитивное отношение к ИИ с 15% до 55%.

Диагностирован «donut hole» менеджмента: C-level инвестирует, молодые сотрудники используют ИИ нативно, а средний менеджмент — именно тот слой, который должен перестроить процессы, — буксует (Wharton/GBK; Gallup: только 9% сотрудников США чувствуют себя «очень комфортно» с ИИ; поддержка менеджера — сильнейший предиктор принятия ИИ командой). Скрытое использование — рациональная стратегия: исследование Melbourne–KPMG (48 000 работников) показало, что 48% признаются в использовании ИИ вопреки политике компании, опасаясь роста норм или сокращений; рекомендации — gain-sharing, гарантии «no-layoff-from-disclosure», привязка стимулов к результатам, а не к часам.

Практические нормативы: минимум 30% бюджета трансформации — на управление изменениями (коммуникация, обучение, работа с сопротивлением); организации, достаточно инвестирующие в людей и процессы, достигают 5-кратно большего прироста выручки (BCG AI@Scale); структурированный change management повышает шансы на успех трансформации в 6 раз (McKinsey Rewired); типовой цикл — 12–24 месяца (confidence: medium).

Из этого следует набор рабочих практик change management для ИИ-программ: (1) публичное и постоянное спонсорство CEO — сигнал, поднимающий позитивное отношение к ИИ с 15% до 55%; (2) системное формальное обучение с выделенным временем, а не «в свободное время»; (3) работа со средним менеджментом как с ключевым слоем: цели, полномочия и стимулы менеджеров должны быть привязаны к перестройке процессов, а не только к текущим KPI; (4) легализация теневого использования через предоставление управляемого доступа и понятных правил — вместо запретов; (5) экономическое разделение выгоды с сотрудниками (gain-sharing) и гарантии занятости при раскрытии автоматизационных приёмов; (6) измерение принятия (adoption rate, доля масштабированных процессов) наравне с финансовыми метриками.

Типовые ошибки enterprise-внедрений (свод Т-Банк «Секреты бизнеса», январь 2026): отсутствие стратегии и KPI, завышенные ожидания ROI, внедрение ИИ в «сырые» процессы, полная зависимость от подрядчиков, точечные инструменты «для галочки» (confidence: medium). По BCG и Gartner, от 60% до 80% проектов внедрения ИИ не достигают заявленных бизнес-эффектов или остаются на уровне пилотов (confidence: medium). Российский контекст воспроизводит те же причины: по оценкам, >80% ИИ-проектов в России терпят крах из-за хаотичных данных и отсутствия интеграций; 51% крупных российских организаций признают неготовность ИТ-инфраструктуры к нейросетям (confidence: low-medium).

Суммируя практический блок для крупного бизнеса: (1) назначить спонсора на уровне СЕО и рассмотреть роль CAIO; (2) строить hub-and-spoke-оргмодель с центром компетенций; (3) по умолчанию покупать специализированные решения, строить собственные только в зонах дифференциации; (4) направлять бюджет в бэк-офис и core-функции, а не только в продажи и маркетинг; (5) фиксировать базовую линию до каждого пилота; (6) выделять минимум 30% бюджета трансформации на change management; (7) для агентных инициатив — сразу проектировать централизованный реестр и оркестрацию, чтобы избежать agent sprawl.

3.3. Сравнительная таблица подходов по размеру бизнеса

Сравнение трёх сегментов показывает: различия касаются не целей (все хотят роста эффективности и выручки), а архитектуры доступа к ИИ, глубины ресурсов и природы главных рисков. Малый бизнес компенсирует отсутствие ресурсов скоростью и стандартными SaaS-инструментами; крупный бизнес компенсирует инерцию платформенными инвестициями и оргмоделью. При этом фактор успеха №1 — вовлечённое руководство и редизайн процессов — идентичен для всех сегментов.

Параметр	Малый бизнес (<50 сотр.)	Средний бизнес (50–249)	Крупный бизнес (250+ / enterprise)
Adoption (строгий замер, 2025)	ЕС: 17,0%; OECD: 11,9%	ЕС: 30,4%; OECD: 20,4%	ЕС: 55,0%; OECD: 40%
Типовой путь	Встроенный ИИ в SaaS (CRM, helpdesk, почта)	SaaS + отдельные платформы, первые интеграции	Платформенная стратегия, собственные LLM-порталы, CoE
Бюджет входа	\$10–100/мес	\$100–1 000/мес	\$5–20 млн на GenAI-внедрение (Gartner)
Команда	Нет ИИ-команды; владелец/маркетолог + MSP/консультант	Ответственный за ИИ/цифровизацию, частичный аутсорсинг	AI CoE, hub-and-spoke, Chief AI Officer
Buy vs build	Только buy (готовые SaaS-инструменты)	Преимущественно buy, точечная кастомизация	Buy-over-build как правило (67% vs 33% успеха, MIT)

Параметр	Малый бизнес (<50 сотр.)	Средний бизнес (50–249)	Крупный бизнес (250+ / enterprise) (NANDA); build — при зрелых компетенциях
Стартовые сценарии	Маркетинг/контент, чат-боты поддержки, документы, анализ данных	Поддержка, продажи (лид-скоринг), финансы/бухгалтерия	Бэк-офис автоматизация (наибольший ROI), клиентский сервис, core-функции (62% ценности)
Цикл пилота	6–12 недель, один процесс	3–6 месяцев, 1–3 пилота	~8 месяцев от прототипа до продакшена; только 48% проектов доходят до продакшена
Главный барьер	Навыки (70,9% не-adopters EC), восприятие нерелевантности (82% микрофирм), стоимость (58%)	Дефицит навыков (41% — главный барьер, Канада), интеграция с CRM/ERP	Change management, инерция руководства, регуляторика (барьер №1 по Deloitte), редизайн процессов
Измеримый эффект	Экономия >20 ч/мес у 58%; \$500–2 000/мес у 66%; ROI ~\$1,6–3,7 на \$1 (confidence: medium)	Рост выручки у 91% использующих (Salesforce, самоотчёт)	ROI у 5–8% в масштабе; лидеры: 1,7× рост выручки, 3,6× TSR (BCG)
Риск-зона	Зависимость от вендора, отсутствие обучения (инвестируют 12%)	«Пилотное чистилище» при отсутствии владельца	Agent sprawl, Shadow AI (54–90%), «donut hole» среднего менеджмента
Господдержка/экосистема	Программы финансирования (BDC LIFT \$500 млн), MSP	Отраслевые консультанты, вендорские программы	Собственные программы обучения, стандарты говернанса (ISO/IEC 42001)

Методологические оговорки к таблице. Показатели adoption в первой строке — строгие государственные замеры (Eurostat, OECD), дающие консервативную нижнюю границу; по самоотчётным опросам (US Chamber, Thryv, QuickBooks, BCC) значения для малого бизнеса в 2–4 раза выше (55–78%) — оба ряда данных приведены в разделе 3.1.1. Оценки ROI для SMB — преимущественно самоотчёты (confidence: medium). Цены инструментов актуальны на 2025–2026 годы и быстро меняются. Часть enterprise-кейсов основана на публичных заявлениях самих компаний.

Итог главы. Малый бизнес выигрывает за счёт скорости и низкого порога входа — его задача: один пилот с базовой линией и честным измерением ROI за 6–12 недель. Крупный бизнес выигрывает за счёт масштаба, но проигрывает в организационной инерции — его задача: спонсорство CEO, редизайн процессов, платформенная оргмодель и минимум 30% бюджета на change management. Общий знаменатель для всех размеров: 70% успеха — люди и процессы, а не модели.

Если предыдущие главы отвечали на вопросы «что происходит» и «как различаются подходы по размеру бизнеса», то следующий блок глав посвящён вопросу «кто и как этим управляет»: сначала — команды и организационные структуры, затем — governance и распределение доступов.

4. Команды и организационные структуры для внедрения ИИ

Технологии искусственного интеллекта к 2025–2026 годам перестали быть главным барьером внедрения: 88% организаций уже используют ИИ, но только около трети сумели выйти за пределы пилотов⁷³. Разрыв «adoption vs impact» — отправная точка любого разговора о командах: если технология доступна всем, а ценность извлекают немногие, то дифференцирующим фактором становится не стек, а организация. Правило BCG 10-20-70 формулирует это количественно: 10% усилий в проектах ИИ приходится на алгоритмы, 20% — на технологии и данные, и 70% — на людей, процессы и организационные изменения; при этом 74% компаний не способны масштабировать ценность ИИ, и около 70% проблем носят людско-процессный, а не технический характер⁷⁴. Настоящая глава рассматривает организационные структуры, роли и программы обучения, которые отличают компании, успешно масштабирующие ИИ: центры компетенций AI CoE (4.1), операционные модели (4.2), ключевые роли (4.3), размер команд по стадиям и масштабу бизнеса (4.4) и программы AI-грамотности (4.5). Сквозной тезис главы: организационный дизайн подбирается под зрелость и размер компании — универсальной «правильной» структуры не существует, но существует типовая эволюционная траектория.

4.1. AI Center of Excellence: функции и назначение

AI Center of Excellence (AI CoE) — выделенное организационное подразделение, централизующее экспертизу, governance и приоритизацию инициатив. Возникновение этой организационной формы — прямой ответ на типовую ситуацию 2023–2024 годов: ИИ-эксперименты велись десятками в разных подразделениях, без общих стандартов, повторно «изобретая» одни и те же компоненты и создавая неучтённые риски.

73 AI Assembly Lines, “What Is an AI Center of Excellence?” — <https://aiassemblylines.com/post/ai-center-of-excellence> (2026-04-09)

74 Clixie / CreativeMarketing (синтезы BCG 2024 “The Leader’s Guide to Transforming with AI”) — <https://www.clixie.ai/blog/10-20-70-rule-ai> ; <https://creativemarketing.ai/blog/10-20-70-rule-ai-budget-allocation> (2026-05-18 ; 2026-07-16)

Каноническая модель выделяет четыре ключевые функции CoE: (1) governance — стандарты, политики, контроль рисков; (2) приоритизация use cases — отбор инициатив по ценности и осуществимости; (3) shared enablement — общие платформы, инструменты и переиспользуемые компоненты; (4) развитие талантов — обучение и наращивание AI-компетенций⁷⁵. Каждая функция отвечает на конкретный провальный сценарий: без governance — регуляторные и репутационные инциденты; без приоритизации — «зоопарк» пилотов без ценности; без shared enablement — каждое подразделение заново покупает и строит одно и то же; без развития талантов — зависимость от единичных носителей экспертизы. Российская практика (DataFinder) фиксирует те же задачи и дополнительно выделяет четыре варианта реализации: центральный CoE с каталогами и сервисами, федеративная модель доменных CoE, встроенный CoE в продуктовые команды и гибрид; типовые роли в RACI-матрице — глава CoE, архитектор, MLOps-инженеры, специалисты по данным и доменные эксперты⁷⁶ (confidence: medium). Сходство российской и глобальной типологий показывает, что дизайн CoE стал интернациональным стандартом де-факто: различия сводятся к масштабу и степени формализации, а не к набору функций.

Исследование McKinsey (State of AI, март 2025) показывает, что функции governance централизуются охотнее остальных: риск/комплаенс и data governance чаще всего полностью централизованы в модели CoE, тогда как технические таланты и adoption-функции обычно гибридные (частично централизованы). Малые компании (выручка <\$500 млн) чаще централизуют всё⁷⁷. Это имеет прямое практическое следствие: малому и среднему бизнесу не нужно выбирать сложную операционную модель — централизация для них не вынужденная мера, а эмпирически доминирующий паттерн. Гибридность же остальных функций отражает компромисс, к которому приходят крупные организации: таланты и внедрение неудобно держать в центре (они должны быть близко к бизнесу), но и полностью отдавать их юнитам рискованно (дублирование, потеря стандартов) — отсюда частично централизованные схемы с выращиванием экспертизы на местах. Состав ядра зрелого CoE включает CAIO, кросс-функциональный steering committee, AI/ML-инженеров и data scientists, специалистов по AI-этике, AI product managers, governance-специалистов и business translators⁷⁸. Ключевой фактор успеха — активное спонсорство со стороны топ-менеджмента: по данным McKinsey, организации с активным executive sponsorship AI-инициатив в 3,3 раза чаще создают ценность выше затрат⁷⁹. По оценкам Deloitte, операционализация ИИ в масштабе через AI CoE способна дать до +20% к показателю

75 AI Assembly Lines, “What Is an AI Center of Excellence?” — <https://aiassemblylines.com/post/ai-center-of-excellence> (2026-04-09)

76 DataFinder, “Центры компетенций AI: структура, задачи и взаимодействие” — <https://datafinder.ru/products/centry-kompetenciya-ai-struktura-zadachi-i-vzaimodeystvie> (n/a (доступ 2025–2026))

77 McKinsey, “The State of AI: How organizations are rewiring to capture value” — <https://www.mckinsey.de/capabilities/quantumblack/our-insights/the-state-of-ai-how-organizations-are-rewiring-to-capture-value> (2025-03-12)

78 Monetizely, “The Agentic AI Center of Excellence” — <https://www.getmonetizely.com/articles/the-agentic-ai-center-of-excellence-how-should-you-structure-your-organization-for-success> (2025-08-30)

79 Monetizely, “The Agentic AI Center of Excellence” — <https://www.getmonetizely.com/articles/the-agentic-ai-center-of-excellence-how-should-you-structure-your-organization-for-success> (2025-08-30)

EBITDA (confidence: medium)⁸⁰.

Важно, что эффективный CoE мультидисциплинарен: помимо стратегов ИИ и data-специалистов, он включает юристов, специалистов по безопасности и change-менеджеров⁸¹. Deloitte подчёркивает, что CoE должен быть «встроен и близок к бизнес-императиву», а не существовать как изолированный исследовательский центр⁸². Это принципиальное требование дизайна: CoE, построенный как «башня из слоновой кости» и отбирающий проекты у бизнеса, воспроизводит ту самую проблему «бутылочного горлышка», которую призван решать. Именно поэтому в зрелых моделях центр отвечает за enablement (платформа, стандарты, обучение), а не за монополию на исполнение.

Мультидисциплинарность также снимает ложную дихотомию «инновации против контроля»: когда юристы и специалисты по безопасности встроены в CoE с самого начала, оценка рисков происходит на этапе приоритизации, а не в виде блокирующей проверки перед запуском — это сокращает и время вывода решений, и число поздно обнаруженных проблем.

Контекст внедрения подчёркивает важность этих структур: по данным McKinsey, приведённым в обзоре AI Assembly Lines, 88% организаций используют ИИ, однако лишь около одной из трёх сумела масштабировать его за пределы proof of concept⁸³. CoE и есть основной организационный механизм преодоления этого разрыва: он превращает разрозненные эксперименты в управляемый портфель инициатив с приоритизацией, переиспользуемыми компонентами и контролем рисков. При этом сам факт создания CoE не гарантирует результата: центр без мандата на приоритизацию становится консультационной службой, а центр без enablement-функции — «отделом запретов». Именно поэтому четыре функции следует рассматривать как неделимый набор: выпадение любой из них воспроизводит один из известных провальных архетипов.

4.2. Операционные модели: централизованная, федеративная, hub-and-spoke

Выбор операционной модели — второе по значимости организационное решение после назначения ответственного лидера: модель определяет, где принимаются решения о портфеле инициатив, кто владеет платформой и стандартами и как распределяется ответственность за результат. Практика 2024–2026 годов выделяет три доминирующих архетипа AI operating model: централизованная, hub-and-spoke и федеративная⁸⁴.

80 IntuitionLabs, “How to Build an AI Center of Excellence in Biotech (2026)” — <https://intuitionlabs.ai/articles/ai-center-of-excellence-biotech> (2026-06-26)

81 Киберленинка, “От пилота к масштабу: как гибридные архитектуры и модели управления определяют успех внедрения ИИ в бизнесе” — <https://cyberleninka.ru/article/n/ot-pilota-k-masshtabu-kak-gibridnye-arhitektury-i-modeli-upravleniya-opredelyayut-uspeh-vnedreniya-iskusstvennogo-intellekta-v> (2026-09-02)

82 AI Assembly Lines, “How Do You Build a Lean AI Center of Excellence? A 4-Phase Mid-Market Design” — <https://aiassemblylines.com/post/build-lean-ai-center-excellence-4-phase-mid-market-design> (2026-07-01)

83 AI Assembly Lines, “What Is an AI Center of Excellence?” — <https://aiassemblylines.com/post/ai-center-of-excellence> (2026-04-09)

84 AI Assembly Lines, “Enterprise AI Operating Model” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> (2026-08-31)

Централизованная модель — все AI-ресурсы, экспертиза и принятие решений сконцентрированы в едином центре. McKinsey фиксирует, что более 50% компаний используют централизованную модель для генеративного ИИ даже при децентрализованной data/analytics-функции⁸⁵. Она оптимальна для ранних стадий зрелости и небольших организаций: «организации на ранних стадиях почти всегда выигрывают от централизованной или слабо выраженной hub-and-spoke модели»⁸⁶. Её сильные стороны — единые стандарты, отсутствие дублирования и быстрый старт; ограничение — центр становится узким местом, когда число инициатив из бизнес-юнитов превышает его пропускную способность.

Hub-and-spoke («хаб и спицы») — доминирующая модель масштабирования. Хаб (CoE) владеет платформой, стандартами, governance и enablement — но, что критично, не gatekeeping-функцией; спицы (бизнес-юниты) отвечают за приоритизацию, delivery и бизнес-результаты. Формулировка «the hub enables rather than gates» — ключевая деталь дизайна: как только хаб начинает блокировать инициативы вместо их ускорения, бизнес-юниты уходят в shadow AI или требуют возврата к децентрализации⁸⁷. Типичный переход к этой модели происходит при наличии более 15–20 активных инициатив в трёх и более бизнес-юнитах⁸⁸. Численные свидетельства преимущества модели: по данным Deloitte, организации на hub-and-spoke показывают на 73% лучшие результаты масштабирования по сравнению с чисто централизованными или децентрализованными структурами (confidence: medium, вторичный источник)⁹⁰; IBM зафиксировала, что CAIO в централизованных и hub-and-spoke моделях достигают на 36% более высокого ROI, чем в децентрализованных⁹¹. Независимое российское исследование подтверждает направление: компании, успешно масштабирующие ИИ, в 3 раза чаще используют hub-and-spoke⁹². Характерно, что сама Deloitte не только исследует модель, но и применяет её: компания запустила глобальный AI Infrastructure CoE (сентябрь 2025) и инвестирует более \$3 млрд в

85 AI Assembly Lines, “Enterprise AI Operating Model” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> (2026-08-31)

86 AI Assembly Lines; Quick AI Lab, “Centralized, Federated, or Hub-and-Spoke?” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> ; <https://www.quickailab.com/articles/aom-02-ai-operating-model> (2026-08-31 ; 2026-07-01)

87 AI Assembly Lines; Quick AI Lab, “Centralized, Federated, or Hub-and-Spoke?” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> ; <https://www.quickailab.com/articles/aom-02-ai-operating-model> (2026-08-31 ; 2026-07-01)

88 AI Assembly Lines, “Enterprise AI Operating Model” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> (2026-08-31)

89 AI Assembly Lines; Quick AI Lab, “Centralized, Federated, or Hub-and-Spoke?” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> ; <https://www.quickailab.com/articles/aom-02-ai-operating-model> (2026-08-31 ; 2026-07-01)

90 UD.hk, “How to Build an AI Center of Excellence (CoE) 2026”; Internal, “Best AI Strategy Frameworks” — <https://www.ud.hk/en/blogs/insight/article/2026-06-05-ai-center-of-excellence-framework> ; <https://internal.ai/best-ai-strategy-frameworks> (2026-06-05 ; 2026)

91 AI Assembly Lines, “Enterprise AI Operating Model” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> (2026-08-31)

92 Киберленинка, “От пилота к масштабу: как гибридные архитектуры и модели управления определяют успех внедрения ИИ в бизнесе” — <https://cyberleninka.ru/article/n/ot-pilota-k-masshtabu-kak-gibridnye-arhitektury-i-modeli-upravleniya-opredelyayut-uspeh-vnedreniya-iskusstvennogo-intellekta-v> (2026-09-02)

генеративный ИИ к 2030 году ⁹³.

Федеративная модель — доменные CoE в бизнес-юнитах при минимальном центре; подходит организациям с высокой зрелостью ИИ ⁹⁴. Её предпосылка — наличие в каждом юните зрелых команд, способных не только эксплуатировать, но и развивать ИИ-решения, а также договорённости о единых стандартах безопасности и данных. Без этих предпосылок федеративная модель дегенерирует в хаос несовместимых стеков; именно поэтому она рассматривается как конечная точка эволюции, а не стартовая конфигурация ⁹⁵.

Сравнительная характеристика трёх моделей обобщена в таблице ниже (синтез ^{96,97,98}):

Критерий	Централизованная	Hub-and-spoke	Федеративная
Кто владеет платформой и стандартами	Центр (CoE)	Хаб (CoE)	Доменные CoE
Кто отвечает за delivery	Центр	Спицы (бизнес-юниты)	Бизнес-юниты
Оптимальная зрелость ИИ	Начальная	Средняя и выше	Высокая
Оптимальный размер компании	<1 000 сотрудников; МСП	1 000–20 000	>20 000; зрелые корпорации
Главный риск	Узкое горлышко, оторванность от бизнеса	Коллапс в централизацию без сети чемпионов	Дублирование, разноречивость стандартов
Эмпирика эффективности	Базовый вариант старта	Deloitte: +73% к результатам масштабирования; IBM: +36% ROI для CAIO	Подходит организациям высокой зрелости ⁹⁹

93 UD.hk, “How to Build an AI Center of Excellence (CoE) 2026”; Iternal, “Best AI Strategy Frameworks” — <https://www.ud.hk/en/blogs/insight/article/2026-06-05-ai-center-of-excellence-framework> ; <https://internal.ai/best-ai-strategy-frameworks> (2026-06-05 ; 2026)

94 AI Assembly Lines, “Enterprise AI Operating Model” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> (2026-08-31)

95 AI Assembly Lines, “Enterprise AI Operating Model” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> (2026-08-31)

96 AI Assembly Lines, “Enterprise AI Operating Model” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> (2026-08-31)

97 AI Assembly Lines; Quick AI Lab, “Centralized, Federated, or Hub-and-Spoke?” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> ; <https://www.quickailab.com/articles/aom-02-ai-operating-model> (2026-08-31 ; 2026-07-01)

98 Киберленинка, “От пилота к масштабу: как гибридные архитектуры и модели управления определяют успех внедрения ИИ в бизнесе” — <https://cyberleninka.ru/article/n/ot-pilota-k-masshtabu-kak-gibridnye-arhitektury-i-modeli-upravleniya-opredelyayut-uspeh-vnedreniya-iskusstvennogo-intellekta-v> (2026-09-02)

99 AI Assembly Lines, “Enterprise AI Operating Model” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> (2026-08-31)

Таблица намеренно фиксирует не только преимущества, но и «точки отказа» каждой модели. Практический выбор между ними сводится к трём вопросам: (1) сколько одновременных инициатив и в скольких юнитах уже идёт — при >15–20 инициативах в 3+ юнитах централизованная модель становится узким местом¹⁰⁰¹⁰¹; (2) есть ли в юнитах люди, способные нести роль чемпионов, — без них hub-and-spoke не работает¹⁰²; (3) насколько отрасль регулируема — в регулируемых секторах централизованный governance даёт лучшие compliance-результаты, что сдвигает оптимум в сторону более сильного хаба даже при высокой зрелости¹⁰³. Ответы на эти вопросы объективируют решение, которое иначе принимается политически — по балансу сил между центром и юнитами.

Эволюция модели по зрелости. Консенсус исследований: траектория ведёт от централизации к гибриду по мере роста зрелости¹⁰⁴. Механика эволюции прозрачна: на старте централизация неизбежна, поскольку экспертиза существует в единственном экземпляре; по мере того как программы AI literacy и сеть чемпионов «разносят» компетенцию по юнитам, маргинальная ценность центра смещается от исполнения к платформе и стандартам — и модель естественным образом перестраивается в hub-and-spoke. Попытка форсировать этот переход до того, как компетенция распространилась, и есть «преждевременный переход», статистика которого приведена ниже. В фарме, например, ~60% компаний стартуют с централизованного хаба и ~40% — с децентрализованных спиц, при этом hub-and-spoke оказывается наиболее устойчивой во времени¹⁰⁵. Однако преждевременный переход опасен: 71% ранних переходов на hub-and-spoke откатываются в течение 18 месяцев со средними невозвратными потерями \$2,3 млн (confidence: medium, единичный источник)¹⁰⁶. Критическая инфраструктура перехода — сеть AI Champions в бизнес-юнитах: её отсутствие — наиболее частая причина коллапса hub-and-spoke обратно в де-факто централизацию¹⁰⁷. Дополнительные слои лидерства под CAIO включают AI Program Lead (операционное управление портфелем, вендоры, governance-процессы) и упомянутую сеть чемпионов, обеспечивающую доменную кредитность и базовую AI-грамотность на местах¹⁰⁸.

100AI Assembly Lines, “Enterprise AI Operating Model” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> (2026-08-31)

101AI Assembly Lines; Quick AI Lab, “Centralized, Federated, or Hub-and-Spoke?” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> ; <https://www.quickailab.com/articles/aom-02-ai-operating-model> (2026-08-31 ; 2026-07-01)

102AI Assembly Lines, “Enterprise AI Operating Model” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> (2026-08-31)

103AI Assembly Lines; Quick AI Lab, “Centralized, Federated, or Hub-and-Spoke?” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> ; <https://www.quickailab.com/articles/aom-02-ai-operating-model> (2026-08-31 ; 2026-07-01)

104Киберленинка, “От пилота к масштабу: как гибридные архитектуры и модели управления определяют успех внедрения ИИ в бизнесе” — <https://cyberleninka.ru/article/n/ot-pilota-k-masshtabu-kak-gibridnye-arhitektury-i-modeli-upravleniya-opredelyayut-uspeh-vnedreniya-iskusstvennogo-intellekta-v> (2026-09-02)

105IntuitionLabs, “How to Build an AI Center of Excellence in Biotech (2026)” — <https://intuitionlabs.ai/articles/ai-center-of-excellence-biotech> (2026-06-26)

106AI Advisory Practice, “AI Operating Model for Enterprise” — <https://aiadvisorypractice.com/blog/ai-operating-model-enterprise-transformation> (2025-12-02)

107AI Assembly Lines, “Enterprise AI Operating Model” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> (2026-08-31)

108AI Assembly Lines, “Enterprise AI Operating Model” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model>

Отдельный управленческий вывод касается governance в регулируемых отраслях: централизованный governance даёт лучшие compliance-результаты, поэтому даже при федеративной модели функции риск-контроля и соответствия целесообразно сохранять в хабе ¹⁰⁹. Это согласуется с данными McKinsey о том, что риск/комплаенс и data governance централизуются чаще остальных AI-функций ¹¹⁰.

Ориентир выбора модели от размера компании (confidence: low-medium, по аналогии специализированных функций): менее 1 000 сотрудников — централизованная; 1 000–5 000 — централизованная или hub-and-spoke; 5 000–20 000 — hub-and-spoke; более 20 000 — hub-and-spoke или федеративная ¹¹¹. Российское исследование сходится: для МСП предпочтительна централизация, для крупных корпораций — гибрид ¹¹².

Не менее важен вопрос «когда переходить», чем «к какой модели». Диагностические сигналы готовности к hub-and-spoke: более 15–20 активных инициатив; три и более бизнес-юнита с собственными use cases; наличие в юнитах сотрудников, способных стать AI Champions; сформированные в центре платформа, стандарты и governance-процессы, которые можно «отдать» спицам без потери контроля. Обратные сигналы преждевременности: инициативы пока носят экспериментальный характер, сеть чемпионов не построена, governance не отделён от исполнения. Статистика откатов (71% в течение 18 месяцев при средних потерях \$2,3 млн ¹¹³) делает эту диагностику финансово значимой процедурой, а не организационной формальностью: смена операционной модели — это дорогая трансформация, а не реорганизация на бумаге.

4.3. Ключевые роли

Организационная структура оживает только через конкретные роли. Рынок 2024–2026 годов сформировал новый контур AI-ролей — от C-level лидерства (CDAO/CAIO) до специализированных позиций (AI Product Manager, AI Governance Officer) — и одновременно показал обратный процесс: часть «модных» ролей первой волны (прежде всего prompt engineer) сжимается и растворяется в широких компетенциях. Ниже рассматривается каждая роль с данными о её распространённости, эффективности и типовом размещении в структуре. Сквозная закономерность: ценность роли измеряется не редкостью навыка, а положением в цепочке создания ценности — роли, связывающие

operating-model (2026-08-31)

109AI Assembly Lines; Quick AI Lab, “Centralized, Federated, or Hub-and-Spoke?” —

<https://aiassemblylines.com/resources/ai-initiatives-operating-model> ;

<https://www.quickailab.com/articles/aom-02-ai-operating-model> (2026-08-31 ; 2026-07-01)

110McKinsey, “The State of AI: How organizations are rewiring to capture value” —

<https://www.mckinsey.de/capabilities/quantumblack/our-insights/the-state-of-ai-how-organizations-are-rewiring-to-capture-value> (2025-03-12)

111TestParty, “Accessibility Team Structure Models” (аналогичная орг. логика специализированных функций)

— <https://testparty.ai/blog/accessibility-team-structure-models> (n/a)

112Киберленинка, “От пилота к масштабу: как гибридные архитектуры и модели управления определяют успех внедрения ИИ в бизнесе” — <https://cyberleninka.ru/article/n/ot-pilota-k-masshtabu-kak-gibridnye-arhitektury-i-modeli-upravleniya-opredelyayut-uspeh-vnedreniya-iskusstvennogo-intellekta-v> (2026-09-02)

113AI Advisory Practice, “AI Operating Model for Enterprise” — <https://aiadvisorypractice.com/blog/ai-operating-model-enterprise-transformation> (2025-12-02)

технологии с бизнес-результатом и управлением изменениями (AI PM, change manager, чемпионы), дефицитнее и влиятельнее ролей, замкнутых на саму технологию.

CDAO / CAIO

Роль Chief Data & Analytics Officer / Chief AI Officer стала центральной в оргструктурах ИИ. По данным Gartner (CDAO Agenda Survey 2025, n=504), 70% CDAO несут основную ответственность за AI-стратегию и operating model; доля CDAO, отчитывающихся напрямую CEO, выросла с 21% в 2024 до 36% в 2025 году ¹¹⁴. Рост отчётности напрямую CEO — индикатор перехода ИИ из разряда ИТ-инициатив в разряд корпоративной стратегии. Gartner выделяет три траектории эволюции роли (Expert D&A Leader, Connector CDAO, Pioneer CDAx — слияние CDO/CDAO/CAIO) и прогнозирует, что к 2027 году 75% CDAO, не воспринимаемых как критичные для успеха ИИ, потеряют C-level титул ¹¹⁵. Траектории различаются центром тяжести: Expert D&A Leader углубляет техническую и аналитическую экспертизу функции; Connector CDAO строит связи между данными, ИИ и бизнес-подразделениями; Pioneer CDAx отражает консолидацию смежных C-level ролей в единое лидерство данных и ИИ. Выбор траектории — не кадровый нюанс, а структурное решение о том, где в организации «живёт» ИИ: в технической вертикали, в горизонтальной трансформации или в едином C-level портфеле. Иными словами, роль либо становится центральной для AI-трансформации, либо поглощается.

Проникновение выделенной роли CAIO растёт взрывными темпами: по IBM IBV (Q1 2025, >600 CAIO, 22 страны), только 26% организаций имели CAIO (рост с 11% в 2023), а обновление IBM CEO Study 2026 фиксирует уже 76% организаций с CAIO; при этом компании с CAIO показывают на 20% более высокий ROI AI-программ и на 29% меньшие потери от AI-сбоев ¹¹⁶¹¹⁷. 57% CAIO назначены из внутренних кадров ¹¹⁸ — то есть рынок пока опирается на переподготовку собственных лидеров, а не на внешний найм.

Эффективность роли зависит от позиционирования: CAIO с отчётностью CEO/COO дают +10% ROI на AI-расходы и +24% к вероятности опережать конкурентов по инновациям ¹¹⁹.

114Gartner Press Release — <https://www.gartner.com/en/newsroom/press-releases/2025-05-12-gartner-survey-finds-seventy-percent-of-cdaos-are-responsible-for-artificial-intelligence-strategy-and-operating-model> (2025-05-12)

115EWSolutions (со ссылкой на Gartner), “Chief Data Officer vs. CIO vs. CDAO” — <https://www.ewsolutions.com/chief-data-officer-vs-cio-vs-cdao/> (2026-06-16)

116IBM IBV, “How Chief AI Officers deliver AI ROI”; IBM Newsroom, “CEOs are Reshaping C-suite Roles for the AI Era” — <https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/chief-ai-officer> ; <https://newsroom.ibm.com/2026-05-04-ibm-study-ceos-are-reshaping-c-suite-roles-for-the-ai-era> (2025-07-13 ; 2026-05-04)

117IBM IBV, “How Chief AI Officers deliver AI ROI”; IBM Newsroom, “CEOs are Reshaping C-suite Roles for the AI Era” — <https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/chief-ai-officer> ; <https://newsroom.ibm.com/2026-05-04-ibm-study-ceos-are-reshaping-c-suite-roles-for-the-ai-era> (2025-07-13 ; 2026-05-04)

118IBM IBV, “How Chief AI Officers deliver AI ROI”; IBM Newsroom, “CEOs are Reshaping C-suite Roles for the AI Era” — <https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/chief-ai-officer> ; <https://newsroom.ibm.com/2026-05-04-ibm-study-ceos-are-reshaping-c-suite-roles-for-the-ai-era> (2025-07-13 ; 2026-05-04)

119AI Assembly Lines, “Enterprise AI Operating Model” — <https://aiassemblylines.com/resources/ai-initiatives->

Для малого и среднего бизнеса аналогом полноценного CAIO служит fractional CAIO в составе lean CoE (см. 4.4): та же ответственность за стратегию и governance при доле полной ставки ¹²⁰.

AI Product Manager

AI product manager — продуктовый лидер, определяющий, где ИИ создаёт ценность, переводящий бизнес-цели в требования и ведущий инициативы от пилота к продакшену; он координирует product, engineering, data science, legal, risk и топ-менеджмент, балансируя feasibility/value/risk ¹²¹¹²². AI PM не строит модели, но обязан понимать работу AI-систем — accuracy, drift, bias, latency как KPI ¹²³. Это принципиально меняет профиль продуктовой роли: если классический продакт-менеджер управляет детерминированным поведением ПО, то AI PM управляет вероятностной системой, у которой качество деградирует (drift), а ошибки распределены неравномерно (bias); отсюда требование владеть метриками модели на уровне постановки целевых значений, а не только приёмки. По оценкам практиков, AI PM и change manager — две наиболее часто отсутствующие и одновременно наиболее влияющие роли в enterprise AI-команде; полная команда покрывает четыре слоя: strategy/governance, platform/infrastructure, model development, business application ¹²⁴. Дефицит именно этих ролей объясним: технические специальности нанимаются по понятным рынку шаблонам, тогда как AI PM требует редкой комбинации продуктового мышления, понимания ограничений моделей и навыка управления заинтересованными сторонами. Для малого и среднего бизнеса практическая импликация состоит в том, что функцию AI PM на ранних стадиях обычно выполняет руководитель направления или владелец продукта, но выделение этой ответственности — хотя бы частичной — критично уже на стадии growth.

ML-инженеры и «вымирание» prompt-инженера как титула

Инженерное ядро команд составляют ML-инженеры, data-инженеры и MLOps-специалисты. Характерная эволюция рынка 2024–2026 — сокращение отдельной роли

operating-model (2026-08-31)

120AI Assembly Lines, “How Do You Build a Lean AI Center of Excellence? A 4-Phase Mid-Market Design” — <https://aiassemblylines.com/post/build-lean-ai-center-excellence-4-phase-mid-market-design> (2026-07-01)

121Boston University, “AI Product Manager in Enterprise Organizations”; Data Science PM, “What is an AI Product Manager?” — <https://www.bu.edu/cds-faculty/programs-admissions/masters-in-enterprise-ai-online/omeai-blogs/role-of-ai-product-manager-in-enterprise-organizations/> ; <https://www.datascience-pm.com/ai-product-manager/> (2026-08-20 ; 2024-12-21)

122Boston University, “AI Product Manager in Enterprise Organizations”; Data Science PM, “What is an AI Product Manager?” — <https://www.bu.edu/cds-faculty/programs-admissions/masters-in-enterprise-ai-online/omeai-blogs/role-of-ai-product-manager-in-enterprise-organizations/> ; <https://www.datascience-pm.com/ai-product-manager/> (2026-08-20 ; 2024-12-21)

123Boston University, “AI Product Manager in Enterprise Organizations”; Data Science PM, “What is an AI Product Manager?” — <https://www.bu.edu/cds-faculty/programs-admissions/masters-in-enterprise-ai-online/omeai-blogs/role-of-ai-product-manager-in-enterprise-organizations/> ; <https://www.datascience-pm.com/ai-product-manager/> (2026-08-20 ; 2024-12-21)

124Alice Labs, “AI Operating Model: Structure Your Organization for AI at Scale” — <https://alicelabs.ai/en/insights/ai-operating-model> (2026-05-23)

«prompt engineer»: LinkedIn зафиксировал падение профилей с этим титулом на 40% (середина 2024 — начало 2025); вакансии достигли пика ~0,3% AI-листингов в 2024 и снизились; Gartner прогнозирует, что к 2026 году 70% предприятий будут использовать AI-автоматизацию промптов. Спрос смещается к более широким ролям — AI workflow design, integration, automation (+25% вакансий) ¹²⁵ (confidence: medium). Навык промптинга не исчезает — он становится частью компетенций широкого круга ролей, а не отдельной профессией; при формировании команды компаниям следует нанимать «процессно-ориентированных» специалистов (интеграция, автоматизация workflow), а не узких носителей титула. Это удешевляет и ускоряет формирование команд: базовый навык работы с промптами массово прививается через программы AI literacy (см. 4.5), а узкая инженерная экспертиза концентрируется в ML/MLOps-ролях. Для рекрутинга это означает смену фильтров: вместо «умеет писать промпты» в описаниях вакансий всё чаще значится «проектирует AI-усиленные процессы», а проверка AI-профессиональности постепенно становится стандартным элементом найма вообще для всех ролей (см. прогноз Gartner о 75% процессов найма к 2027 году в 4.5 ¹²⁶).

Change Manager и сеть AI Champions

Формально отсутствующая в большинстве оргдиаграмм, но критичная роль — change manager: по оценкам практиков, именно она наряду с AI PM чаще всего отсутствует и при этом сильнее всего влияет на результат ¹²⁷. Задача роли — перестройка процессов, работа с сопротивлением и доведение внедрённых инструментов до фактического использования. Операционным продолжением change-функции в бизнес-юнитах служит сеть AI Champions: сотрудники с доменной кредитностью, обученные на уровне power users и выполняющие роль «первой линии» adoption. Именно отсутствие champion-сети — наиболее частая причина коллапса hub-and-spoke обратно в де-факто централизацию ¹²⁸: без местных носителей компетенции спицы не способны самостоятельно развивать инициативы, и вся работа стекает обратно в хаб. Для среднего бизнеса сеть чемпионов — самый дешёвый элемент всей оргмодели: это не новые ставки, а выделенная доля времени существующих сотрудников плюс их обучение (см. сегментацию программ в 4.5).

AI Governance Officer

AI Governance Officer отвечает за разработку и поддержку AI governance framework,

¹²⁵AI Plain English, “Why I Said Goodbye to Prompt Engineering (And What’s Next)” —

<https://ai.plainenglish.io/why-i-said-goodbye-to-prompt-engineering-and-whats-next-8eb18fe82c27> (2025-09-01)

¹²⁶IBM Think (анализ Gartner 2026 predictions); DataGalaxy; Seltech —

<https://www.ibm.com/think/insights/gartner-2026-tech-predictions-implications> ;

<https://www.datagalaxy.com/en/blog/gartners-top-5-data-analytics-predictions-for-new-wave-data-teams-in-2026/> ; <https://www.seltech.io/blog/gartner-2026-predictions-navigating-era-ai-agents-and-data-integrity> (2025-10-20 ; 2026-06-24)

¹²⁷Alice Labs, “AI Operating Model: Structure Your Organization for AI at Scale” —

<https://alicelabs.ai/en/insights/ai-operating-model> (2026-05-23)

¹²⁸AI Assembly Lines, “Enterprise AI Operating Model” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> (2026-08-31)

мониторинг соответствия EU AI Act и отраслевым требованиям, оценку рисков моделей (bias, explainability, robustness), обучение сотрудников responsible AI, взаимодействие с регуляторами и аудитом, отчётность совету директоров; роль кросс-функциональна (tech, data science, legal, compliance, product) ¹²⁹. При этом AI governance — распределённая ответственность: борд несёт ultimate accountability; CDAO/CAIO владеет стратегией, реестром use cases, риск-тиерами и model cards; CISO — AI control baseline и red-teaming промпт-инъекций; DPO — DPIA; также участвуют Legal/Compliance, Risk/Internal Audit, Procurement, HR и AI Ethics Committee ¹³⁰. Такая конструкция означает, что даже в компании без выделенного AI Governance Officer эти функции должны быть явно распределены между существующими ролями — иначе governance остаётся «ничейным». Распределённая модель требует и ритма: реестр use cases и риск-тиры (CDAO/CAIO), контрольный базис безопасности (CISO) и DPIA (DPO) должны регулярно сводиться в отчётность борду, который несёт конечную ответственность ¹³¹. По McKinsey (n=1 491, 101 страна), oversight AI governance со стороны CEO — фактор, наиболее коррелирующий с EBIT-эффектом от gen AI (особенно в крупных компаниях); в 28% организаций governance курирует CEO, в 17% — совет директоров; в среднем два лидера совместно отвечают за AI governance ¹³². Эти цифры имеют двойное прочтение: с одной стороны, высокая доля CEO-oversight подтверждает стратегический статус темы; с другой — более чем в половине организаций governance ИИ не поднят ни до CEO, ни до борда, что при экстерриториальных регуляторных рисках (см. 5.1) выглядит системным пробелом.

4.4. Размер команд по стадиям и масштабу бизнеса

Размер AI-команды определяется стадией зрелости и масштабом компании (confidence: medium) ¹³³:

Стадия	Размер команды	Типовой состав
Pilot / seed	1–3 человека	AI-лид + 1–2 инженера
Growth	5–8 человек	2–3 ML-инженера, 1–2 data engineer, MLOps, AI PM
Scale	15–20 человек	Несколько подов + AI-архитектор
Enterprise	20+ человек	Платформенная команда + поды + CAIO + governance

129Taylor Root, “AI Governance Officer job description” — <https://www.taylorroot.com/resources/job-description-templates/ai-governance-officer-job-description/> (2026-04-28)

130Knostic, “Primer: AI Governance Roles and Responsibilities”; Diligent, “AI governance: A guide for boards” — <https://www.knostic.ai/blog/ai-governance-roles-stakeholders> ; <https://www.diligent.com/resources/blog/ai-governance> (2025-10-15 ; 2026-06-12)

131Knostic, “Primer: AI Governance Roles and Responsibilities”; Diligent, “AI governance: A guide for boards” — <https://www.knostic.ai/blog/ai-governance-roles-stakeholders> ; <https://www.diligent.com/resources/blog/ai-governance> (2025-10-15 ; 2026-06-12)

132McKinsey, “The State of AI” (PDF) — https://www.mckinsey.com/~media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai/2025/the-state-of-ai-how-organizations-are-rewiring-to-capture-value_final.pdf (2025-03)

133Futureproofing.dev, “AI-Native Team Structure: Roles and Composition” — <https://www.futureproofing.dev/resources/ai-native-team/ai-native-team-structure> (2026-06-01)

Логика роста команды отражает переход от «героических усилий» одного-двух энтузиастов к индустриальной модели: на стадии pilot критичны скорость и фокус на одном измеримом use case; на стадии growth появляется разделение труда (данные, модели, эксплуатация, продукт); на стадии scale команда дробится на автономные поды с общим архитектором; на enterprise-уровне добавляются платформенная команда и формальные роли governance¹³⁴. Подовая структура (pod) — кросс-функциональная мини-команда вокруг конкретного продукта или процесса — повторяет на уровне ИИ-команд логику hub-and-spoke на уровне организации: общий хаб даёт платформу и стандарты, поды отвечают за бизнес-результат в своём домене. Для enterprise-стадии типичен четырёхслойный состав команды: strategy/governance, platform/infrastructure, model development, business application¹³⁵ — по сути, микрокопия полной операционной модели внутри одной функции.

При чтении таблицы важно учитывать две оговорки. Во-первых, это ориентиры ядра AI-функции, а не полная численность вовлечённых: сеть чемпионов, ИБ, юристы и доменные эксперты добавляют эквивалент нескольких ставок, распределённых по организации. Во-вторых, переход между стадиями редко бывает линейным: типичный паттерн — стагнация на уровне pilot, пока не появятся AI PM и MLOps-функция, то есть рост численности без восполнения дефицитных ролей не сдвигает стадию. Поэтому решение о расширении команды следует принимать не по числу пилотов, а по наличию в портфеле инициатив, готовых к продакшену, но упирающихся в отсутствующие компетенции¹³⁶¹³⁷.

Для mid-market (средний бизнес) разработан формат lean AI CoE — 3–5 человек: fractional CAIO / AI strategy lead, 1–2 встроенных практика и program manager; такой lean CoE покрывает все четыре функции центра компетенций без издержек полноценного подразделения¹³⁸ (confidence: medium). Это принципиально важно для среднего бизнеса: полноценный CoE enterprise-образца ему недоступен по бюджету, но fractional-модель позволяет получить governance, приоритизацию и enablement на приемлемых издержках. Ключевые дизайн-решения lean-формата: fractional CAIO берёт на себя стратегию и представление интересов ИИ на уровне правления; встроенные практики работают непосредственно в бизнес-командах, обеспечивая близость к «бизнес-императиву», которую требует Deloitte¹³⁹; program manager ведёт портфель инициатив, вендоров и ритм governance-процессов. При росте портфеля lean CoE естественным образом эволюционирует в хаб hub-and-spoke модели — без болезненной реорганизации, поскольку практики уже работают внутри юнитов и фактически предфигурируют сеть чемпионов.

134Futureproofing.dev, “AI-Native Team Structure: Roles and Composition” —

<https://www.futureproofing.dev/resources/ai-native-team/ai-native-team-structure> (2026-06-01)

135Alice Labs, “AI Operating Model: Structure Your Organization for AI at Scale” —

<https://alicelabs.ai/en/insights/ai-operating-model> (2026-05-23)

136Alice Labs, “AI Operating Model: Structure Your Organization for AI at Scale” —

<https://alicelabs.ai/en/insights/ai-operating-model> (2026-05-23)

137Futureproofing.dev, “AI-Native Team Structure: Roles and Composition” —

<https://www.futureproofing.dev/resources/ai-native-team/ai-native-team-structure> (2026-06-01)

138AI Assembly Lines, “How Do You Build a Lean AI Center of Excellence? A 4-Phase Mid-Market Design” —

<https://aiassemblylines.com/post/build-lean-ai-center-excellence-4-phase-mid-market-design> (2026-07-01)

139AI Assembly Lines, “How Do You Build a Lean AI Center of Excellence? A 4-Phase Mid-Market Design” —

<https://aiassemblylines.com/post/build-lean-ai-center-excellence-4-phase-mid-market-design> (2026-07-01)

Крупные компании (выручка >\$500 млн) перестраиваются быстрее: они вдвое чаще создают выделенные команды для AI adoption и чаще внедряют role-based обучение¹⁴⁰. Это подтверждает общий вывод: масштаб бизнеса определяет не решение «использовать ли ИИ», а архитектуру команд и governance. Практический ориентир выбора состава для разных размеров бизнеса:

- **Малый бизнес (до ~100 сотрудников):** выделенная команда не обязательна — достаточно 1–3 человек, совмещающих AI-функции с основными ролями, плюс внешний консультант или fractional-лид; операционная модель — де-факто централизованная¹⁴¹¹⁴².
- **Средний бизнес:** lean CoE 3–5 человек¹⁴³, централизованная или начальная hub-and-spoke модель; критично назначить ответственного за AI-стратегию (по Gartner, эту роль в 70% случаев берёт CDAO-эквивалент¹⁴⁴).
- **Крупный бизнес:** команда 20+, CAIO с отчётностью CEO/COO (что даёт измеримый прирост ROI¹⁴⁵), сеть AI Champions, выделенный AI Governance Officer¹⁴⁶.

Сравнение трёх сегментов обнажает общий паттерн: у малого бизнеса ключевой ресурс — время одного-двух энтузиастов, и ошибка стоит недели работы; у среднего — связность lean-команды с бизнесом, и ошибка стоит потерянного года пилотов; у крупного — согласованность хаба и спиц, и ошибка измеряется миллионами долларов откатов¹⁴⁷. Поэтому инвестиционный приоритет различается: малый бизнес вкладывается в доступ и обучение, средний — в lean CoE и чемпионов, крупный — в платформенную команду, governance и управление трансформацией модели.

4.5. Программы AI literacy и правило BCG 10-20-70

Обучение — обязательный компонент оргструктуры: ни одна операционная модель не работает, если сотрудники бизнес-юнитов не способны сформулировать задачу для ИИ, оценить результат и соблюсти ограничения по данным. При этом данные McKinsey 2025

140Genioux Facts (синтез McKinsey State of AI 2025) — <https://blog.geniouxfacts.com/2025/03/g-f23387-rewiring-for-ai-value-key.html> (2025-03)

141Киберленинка, “От пилота к масштабу: как гибридные архитектуры и модели управления определяют успех внедрения ИИ в бизнесе” — <https://cyberleninka.ru/article/n/ot-pilota-k-masshtabu-kak-gibridnye-arhitektury-i-modeli-upravleniya-opredelyayut-uspeh-vnedreniya-iskusstvennogo-intellekta-v> (2026-09-02)

142Futureproofing.dev, “AI-Native Team Structure: Roles and Composition” — <https://www.futureproofing.dev/resources/ai-native-team/ai-native-team-structure> (2026-06-01)

143AI Assembly Lines, “How Do You Build a Lean AI Center of Excellence? A 4-Phase Mid-Market Design” — <https://aiassemblylines.com/post/build-lean-ai-center-excellence-4-phase-mid-market-design> (2026-07-01)

144Gartner Press Release — <https://www.gartner.com/en/newsroom/press-releases/2025-05-12-gartner-survey-finds-seventy-percent-of-cdaos-are-responsible-for-artificial-intelligence-strategy-and-operating-model> (2025-05-12)

145AI Assembly Lines, “Enterprise AI Operating Model” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> (2026-08-31)

146Taylor Root, “AI Governance Officer job description” — <https://www.taylorroot.com/resources/job-description-templates/ai-governance-officer-job-description/> (2026-04-28)

147AI Advisory Practice, “AI Operating Model for Enterprise” — <https://aiadvisorypractice.com/blog/ai-operating-model-enterprise-transformation> (2025-12-02)

года показывают глубину проблемы: перестройка workflow — сильнейший из 25 измеренных факторов ЕВІТ-эффекта от ИИ, однако только 21% организаций фундаментально перестроили свои процессы ¹⁴⁸. Иначе говоря, недостаточно «обучить людей пользоваться чатом» — требуется перепроектирование рабочих процессов вокруг ИИ, и именно здесь кадровый разрыв максимален. В этом разделе рассматриваются три уровня темы: эталонные программы крупных работодателей, методика их построения и макропрогнозы рынка навыков, определяющие, почему инвестиции в AI-грамотность перестали быть дискреционными.

Эталонные корпоративные программы

Практики крупных работодателей демонстрируют три различных дизайна программ ¹⁴⁹:

- **KPMG** — программа GenAI 101 и обязательный тренинг «Trusted AI» для всех сотрудников: массовый базовый слой плюс нормативный слой ответственного использования;
- **PwC** — геймифицированная программа PowerUp с более чем 9 000 участников в каждой ежемесячной игре (США и Мексика): ставка на вовлечение и регулярную практику вместо разовых курсов;
- **Ally Financial** — ежеквартальные «AI Days» и внутреннее AI-сообщество: обучение через сообщество практиков и регулярные события.

Методика и сегментация обучения

Лучшие практики построения программ AI literacy (в том числе под требования ст. 4 EU AI Act): структурированный e-learning в сочетании с воркшопами; разделение на общий модуль и role-/organization-specific тренинги; градация уровней от beginner до advanced; игровые и story-driven методы; employee-led сообщества; интеграция в онбординг; программы «AI Governance Champion / Responsible AI Champion» ¹⁵⁰. Требование ст. 4 EU AI Act о «достаточном уровне AI-грамотности» действует с 2 февраля 2025 года (см. 5.1), поэтому для компаний в скоупе регулятора обучение — не добровольная инициатива HR, а юридическое обязательство ¹⁵¹. Сегментация аудитории (по Writer): общая рабочая сила (базовая AI literacy), power users / citizen automators (чемпионы, строящие департаментские агенты), технические и governance-команды (API, безопасность, оркестрация) ¹⁵². Такая трёхслойная сегментация позволяет не перетренировывать большинство и не недотренировывать ключевые группы: общему потоку достаточно курса уровня GenAI

148Genioux Facts (синтез McKinsey State of AI 2025) — <https://blog.geniouxfacts.com/2025/03/g-f23387-rewiring-for-ai-value-key.html> (2025-03)

149Great Place To Work, “How the 100 Best Companies Are Training Their Workforce for AI” — <https://www.greatplacetowork.com/resources/blog/100-best-training-workforce-ai> (2025-01-31)

150Trail, “How to build AI Literacy with AI Trainings Best Practices” — <https://www.trail-ml.com/blog/building-ai-literacy-under-the-ai-act-best-practices> (2025-08-26)

151Alice Labs (обзор официального таймлайна ЕК); Scalewise; Haffa.ai — <https://alicelabs.ai/en/insights/eu-ai-act-timeline-2026> ; <https://scalewise.com/resources/eu-ai-act-timeline-ai-vendors-developers-2025-2026/> (2026-05–2026-08)

152Writer, “Agentic AI adoption: A leader’s roadmap” — <https://writer.com/blog/responsible-ai-adoption-and-training/> (2025-09-10)

101, чемпионам нужны практикумы по построению агентов, а техническим и governance-командам — углублённая подготовка по безопасности и соответствию.

Рынок навыков и прогнозы

Долгосрочный фон, на котором строятся корпоративные программы, задают прогнозы рынка труда и навыков: по Gartner, к 2027 году 75% процессов найма будут включать тестирование AI-профессиональности, а организации с AI-грамотными руководителями покажут на 20% лучшие финансовые результаты; AI literacy станет обязательной частью онбординга в крупных организациях ¹⁵³. Одновременно Gartner прогнозирует, что к 2026 году 50% организаций введут «AI-free» оценки для защиты критического мышления — то есть обучение строится не только на использовании ИИ, но и на сохранении автономной экспертизы ¹⁵⁴. McKinsey фиксирует масштаб перекалфикации: 44% организаций переобучили до 5% сотрудников из-за ИИ за прошедший год, 52% планируют переобучить более 10% персонала в ближайшие три года ¹⁵⁵. По WEF Future of Jobs 2025, ИИ и big data — самые быстрорастущие навыки; 39% ключевых навыков изменятся к 2030 году; 63% работодателей называют дефицит навыков главным барьером трансформации ¹⁵⁶. В России к 2025 году 51% работодателей внедрили ИИ в программы обучения сотрудников (РБК Компании); доля компаний, использующих ИИ для разработки корпоративных курсов, выросла с 3,1% до 13,5% ¹⁵⁷.

По данным McKinsey State of AI 2024, компании-лидеры создают выделенные AI governance структуры — CoE или AI leadership councils — и инвестируют в массовые role-based upskilling-программы; в качестве стандартов используются ISO/IEC 42001 и NIST AI RMF ¹⁵⁸. Это связывает тему настоящей главы с главой 5: обучение и governance — две стороны одной системы управления ИИ. Формально leadership council отличается от CoE тем, что является не исполнительным подразделением, а руководящим комитетом: он задаёт приоритеты и политику, тогда как исполнение остаётся в бизнес-юнитах. Выбор

153IBM Think (анализ Gartner 2026 predictions); DataGalaxy; Seltech — <https://www.ibm.com/think/insights/gartner-2026-tech-predictions-implications> ; <https://www.datagalaxy.com/en/blog/gartners-top-5-data-analytics-predictions-for-new-wave-data-teams-in-2026/> ; <https://www.seltech.io/blog/gartner-2026-predictions-navigating-era-ai-agents-and-data-integrity> (2025-10-20 ; 2026-06-24)

154IBM Think (анализ Gartner 2026 predictions); DataGalaxy; Seltech — <https://www.ibm.com/think/insights/gartner-2026-tech-predictions-implications> ; <https://www.datagalaxy.com/en/blog/gartners-top-5-data-analytics-predictions-for-new-wave-data-teams-in-2026/> ; <https://www.seltech.io/blog/gartner-2026-predictions-navigating-era-ai-agents-and-data-integrity> (2025-10-20 ; 2026-06-24)

155Genioux Facts (синтез McKinsey State of AI 2025) — <https://blog.geniouxfacts.com/2025/03/g-f23387-rewiring-for-ai-value-key.html> (2025-03)

156Matrix Mag / ESOF (синтезы WEF Future of Jobs 2025) — <https://matrixmag.com/skills-outlook-2025-preparing-the-workforce-for-a-transformative-future/> ; <https://esoftskills.ie/preparing-for-the-workforce-of-the-future-essential-skills-and-strategies/> (2025-01-11 ; 2024-12-22)

157Яндекс Практикум B2B (со ссылкой на РБК Компании) — <https://practicum.yandex.ru/b2b/blog/iskusstvennyj-intellekt-v-obrazovanii/> (2025-10-10)

158RSI Security (синтез McKinsey State of AI 2024) — <https://blog.rsisecurity.com/the-state-of-ai-in-2024-mckinseys-global-survey-recap/> (2025-08-08 (о отчёте 2024))

между CoE и council — по сути выбор между «центр делает и контролирует» и «центр только контролирует»; для компаний со зрелыми юнитами council — более лёгкий и менее конфликтный вариант централизации oversight.

Данные Deloitte (State of Generative AI in the Enterprise, 2024) добавляют количественную рамку: только 5% организаций («superscalers») масштабировали $\geq 70\%$ gen AI пилотов, тогда как 68% масштабировали $\leq 30\%$; отличие superscalers — подход к талантам: широкий доступ сотрудников к инструментам gen AI и аугментация, а не замена людей ¹⁵⁹.

Практический смысл: лидеры не сокращают круг пользователей ИИ до «прошедших отбор», а, наоборот, максимально расширяют его, вкладываясь в обучение. Это прямо перекликается со статистикой shadow AI из главы 5: сотрудники всё равно будут пользоваться ИИ — вопрос лишь в том, произойдёт это в управляемом контуре с обучением или в тени, с личных аккаунтов и без правил. Выбор «аугментация, а не замена» имеет и вторичный эффект: он снижает сопротивление внедрению, поскольку сотрудники воспринимают ИИ как инструмент усиления своей роли, а не как угрозу ей, — что напрямую облегчает работу change-функции и сети чемпионов.

Возвращаясь к правилу BCG 10-20-70: именно 70% усилий, направленные на людей, процессы и организационные изменения — обучение, сети чемпионов, перестройка workflow — отличают лидеров ¹⁶⁰. По синтезам исследования BCG 2024 года, примерно 70% проблем внедрения — людские и процессные, 20% — технологические и лишь 10% связаны с алгоритмами; соответственно, 74% компаний не могут масштабировать ценность ИИ прежде всего из-за организационных, а не технических причин ¹⁶¹.

Бюджетная импликация прямая: компания, тратящая основную часть AI-бюджета на лицензии и модели и почти ничего — на обучение и редизайн процессов, инвестирует с точностью до наоборот.

Отдельный вопрос — измеримость программ обучения. Практика эталонных компаний подсказывает метрики: охват (доля сотрудников, прошедших базовый модуль), глубина (доля power users, способных самостоятельно построить агента или автоматизацию), поведенческие индикаторы (доля работы в корпоративных, а не личных аккаунтах — одновременно метрика shadow AI из главы 5) и бизнес-эффекты (число инициатив, инициированных бизнес-юнитами после обучения). Связка обучения и сети чемпионов делает программу самоподдерживающейся: чемпионы, прошедшие углублённый трек, становятся внутренними тренерами и амбассадорами, снижая зависимость от внешних

159Deloitte Insights Magazine Issue 33 —

https://www2.deloitte.com/content/dam/insights/articles/glob187719_dim33/Deloitte-Insights-Magazine-Issue-33.pdf (2024)

160Clixie / CreativeMarketing (синтезы BCG 2024 “The Leader’s Guide to Transforming with AI”) —

<https://www.clixie.ai/blog/10-20-70-rule-ai> ; <https://creativemarketing.ai/blog/10-20-70-rule-ai-budget-allocation> (2026-05-18 ; 2026-07-16)

161Clixie / CreativeMarketing (синтезы BCG 2024 “The Leader’s Guide to Transforming with AI”) —

<https://www.clixie.ai/blog/10-20-70-rule-ai> ; <https://creativemarketing.ai/blog/10-20-70-rule-ai-budget-allocation> (2026-05-18 ; 2026-07-16)

провайдеров¹⁶²¹⁶³.

Типовые ошибки при построении команд и моделей

Обобщение данных главы позволяет выделить повторяющиеся ошибки. Первая — CoE как «башня из слоновой кости»: центр, который накапливает экспертизу, но не встроен в бизнес-приоритеты; Deloitte прямо требует обратного — CoE должен быть embedded и близок к бизнес-императиву¹⁶⁴. Вторая — переход к hub-and-spoke раньше готовности: без сети чемпионов, платформы и отделённого governance модель коллапсирует обратно в централизацию, в среднем с потерей \$2,3 млн¹⁶⁵. Третья — пропуск ролей AI PM и change manager: команды нанимают инженеров и data-специалистов, но не тех, кто переводит бизнес-цели в требования и доводит внедрение до фактического использования¹⁶⁶. Четвёртая — инвестиционный перекосяк: бюджет тратится на лицензии и модели, тогда как BCG фиксирует, что 70% проблем — людские и процессные¹⁶⁷. Пятая — отсутствие реального спонсорства: формально назначенный ответственный без мандата и доступа к CEO статистически проигрывает конфигурации «CAIO с отчётностью CEO/COO»¹⁶⁸, а организации без активного executive sponsorship в 3,3 раза реже создают ценность выше затрат¹⁶⁹.

4.6. Выводы по главе

Глава показала, что «организационный слой» внедрения ИИ имеет собственную, хорошо измеренную эмпирику: выбор операционной модели, назначение лидера и инвестиции в обучение дают квантифицируемые эффекты (от +36% ROI у hub-and-spoke до 3,3-кратного роста шансов на ценность при executive sponsorship). Ключевые выводы:

1. **Организация важнее технологий.** 70% усилий и проблем внедрения — люди и процессы (правило BCG 10-20-70)¹⁷⁰; бюджет и внимание руководства должны

162Trail, “How to build AI Literacy with AI Trainings Best Practices” — <https://www.trail-ml.com/blog/building-ai-literacy-under-the-ai-act-best-practices> (2025-08-26)

163Writer, “Agentic AI adoption: A leader’s roadmap” — <https://writer.com/blog/responsible-ai-adoption-and-training/> (2025-09-10)

164AI Assembly Lines, “How Do You Build a Lean AI Center of Excellence? A 4-Phase Mid-Market Design” — <https://aiassemblylines.com/post/build-lean-ai-center-excellence-4-phase-mid-market-design> (2026-07-01)

165AI Advisory Practice, “AI Operating Model for Enterprise” — <https://aiadvisorypractice.com/blog/ai-operating-model-enterprise-transformation> (2025-12-02)

166Alice Labs, “AI Operating Model: Structure Your Organization for AI at Scale” — <https://alicelabs.ai/en/insights/ai-operating-model> (2026-05-23)

167Clxie / CreativeMarketing (синтезы BCG 2024 “The Leader’s Guide to Transforming with AI”) — <https://www.clixie.ai/blog/10-20-70-rule-ai> ; <https://creativemarketing.ai/blog/10-20-70-rule-ai-budget-allocation> (2026-05-18 ; 2026-07-16)

168AI Assembly Lines, “Enterprise AI Operating Model” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> (2026-08-31)

169Monetizely, “The Agentic AI Center of Excellence” — <https://www.getmonetizely.com/articles/the-agentic-ai-center-of-excellence-how-should-you-structure-your-organization-for-success> (2025-08-30)

170Clxie / CreativeMarketing (синтезы BCG 2024 “The Leader’s Guide to Transforming with AI”) — <https://www.clixie.ai/blog/10-20-70-rule-ai> ; <https://creativemarketing.ai/blog/10-20-70-rule-ai-budget-allocation> (2026-05-18 ; 2026-07-16)

распределяться соответственно: каждая вложенная в платформу единица должна сопровождаться кратно большими вложениями в обучение и редизайн процессов.

2. **Модель выбирается по зрелости и размеру:** малый бизнес — централизация и команда 1–3 человека; средний — lean CoE 3–5 человек¹⁷¹; крупный — hub-and-spoke с хабом-СоЕ, сетью чемпионов и командой 20+^{172,173}. Переход к hub-and-spoke раньше готовности — статистически проигрышная ставка¹⁷⁴.
3. **Спонсорство и расположение роли имеют измеримый эффект:** активное executive sponsorship утраивает шансы на ценность выше затрат¹⁷⁵; CAIO с отчётностью CEO/COO даёт +10% ROI¹⁷⁶; oversight CEO наиболее коррелирует с EBIT-эффектом¹⁷⁷. Формальное назначение ответственного без реального мандата не работает — Gartner прямо прогнозирует потерю C-level статуса для 75% CDAO, не доказавших значимости для ИИ¹⁷⁸.
4. **Кадровая политика смещается от узких титулов к широким компетенциям:** титул prompt-инженера вымирает¹⁷⁹, а AI PM и change manager остаются самыми дефицитными и самыми влиятельными ролями¹⁸⁰. Приоритет найма/переподготовки — продуктивно-процессные роли, а не узкие технические специальности первой волны.
5. **Обучение — не опция, а регуляторное и конкурентное требование:** ст. 4 EU AI Act прямо требует AI-грамотности¹⁸¹, а Gartner прогнозирует проверку AI-профессиональности в 75% наймов к 2027 году¹⁸². Программы следует строить

171AI Assembly Lines, “How Do You Build a Lean AI Center of Excellence? A 4-Phase Mid-Market Design” — <https://aiassemblylines.com/post/build-lean-ai-center-excellence-4-phase-mid-market-design> (2026-07-01)

172AI Assembly Lines, “Enterprise AI Operating Model” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> (2026-08-31)

173Futureproofing.dev, “AI-Native Team Structure: Roles and Composition” —

<https://www.futureproofing.dev/resources/ai-native-team/ai-native-team-structure> (2026-06-01)

174AI Advisory Practice, “AI Operating Model for Enterprise” — <https://aiadvisorypractice.com/blog/ai-operating-model-enterprise-transformation> (2025-12-02)

175Monetizely, “The Agentic AI Center of Excellence” — <https://www.getmonetizely.com/articles/the-agentic-ai-center-of-excellence-how-should-you-structure-your-organization-for-success> (2025-08-30)

176AI Assembly Lines, “Enterprise AI Operating Model” — <https://aiassemblylines.com/resources/ai-initiatives-operating-model> (2026-08-31)

177McKinsey, “The State of AI” (PDF) —

https://www.mckinsey.com/~media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai/2025/the-state-of-ai-how-organizations-are-rewiring-to-capture-value_final.pdf (2025-03)

178EWSolutions (со ссылкой на Gartner), “Chief Data Officer vs. CIO vs. CDAO” —

<https://www.ewsolutions.com/chief-data-officer-vs-cio-vs-cdao/> (2026-06-16)

179AI Plain English, “Why I Said Goodbye to Prompt Engineering (And What’s Next)” —

<https://ai.plainenglish.io/why-i-said-goodbye-to-prompt-engineering-and-whats-next-8eb18fe82c27> (2025-09-01)

180Alice Labs, “AI Operating Model: Structure Your Organization for AI at Scale” —

<https://alicelabs.ai/en/insights/ai-operating-model> (2026-05-23)

181Alice Labs (обзор официального таймлайна ЕК); Scalewise; Haffa.ai — <https://alicelabs.ai/en/insights/eu-ai-act-timeline-2026> ; <https://scalewise.com/resources/eu-ai-act-timeline-ai-vendors-developers-2025-2026/> (2026-05–2026-08)

182IBM Think (анализ Gartner 2026 predictions); DataGalaxy; Seltech —

<https://www.ibm.com/think/insights/gartner-2026-tech-predictions-implications> ;

<https://www.datagalaxy.com/en/blog/gartners-top-5-data-analytics-predictions-for-new-wave-data-teams->

сегментированно (общий поток / power users / технические команды) и встраивать в онбординг с первого дня¹⁸³¹⁸⁴.

5. Управление ИИ: governance, политики и распределение доступов сотрудников

Если глава 4 отвечала на вопрос «кто внедряет ИИ», то настоящая глава — на вопрос «как этим управлять». Потребность в формализованном governance обусловлена тремя давлениями: регуляторным (EU AI Act перешёл от политических дебатов к действующим штрафам), безопасностным (массовый shadow AI и утечки через генеративные сервисы) и операционным (по мере роста числа ИИ-инструментов и агентов ручное управление доступами перестает масштабироваться). Ниже рассматриваются фреймворки и стандарты (5.1), вызов shadow AI и контроль данных (5.2), технические механизмы распределения доступов в корпоративных LLM-платформах (5.3) и практическая матрица доступов по ролям и размерам компаний (5.4).

5.1. Фреймворки и регуляторные рамки

Рынок управления ИИ в 2024–2026 годах структурировался вокруг трёх опор: добровольной методологии рисков (NIST AI RMF), сертифицируемой системы менеджмента (ISO/IEC 42001) и обязательного регулирования (EU AI Act). Для бизнеса они не взаимозаменяемы: первый отвечает на вопрос «как думать о рисках», второй — «как выстроить и доказать систему управления», третий — «что обязательно по закону». Понимание границ каждой рамки и их пересечений — необходимое условие рационального, без дублирования, построения контура governance.

NIST AI RMF

NIST AI Risk Management Framework (NIST AI 100-1) — добровольный фреймворк, построенный вокруг четырёх функций: Govern (институционализация управления рисками ИИ), Map (идентификация контекста и рисков), Measure (измерение и оценка), Manage (обработка и мониторинг рисков). 26 июля 2024 года NIST опубликовал NIST AI 600-1 — Generative AI Profile, выделяющий 12 категорий рисков, специфичных для генеративного ИИ, включая «confabulation» (галлюцинации), вредоносные и предвзятые результаты, утечки данных, проблемы интеллектуальной собственности и риски информационной безопасности. Профиль разработан в ответ на Executive Order 14110; он не заменяет

in-2026/ ; <https://www.seltech.io/blog/gartner-2026-predictions-navigating-era-ai-agents-and-data-integrity> (2025-10-20 ; 2026-06-24)

183Trail, “How to build AI Literacy with AI Trainings Best Practices” — <https://www.trail-ml.com/blog/building-ai-literacy-under-the-ai-act-best-practices> (2025-08-26)

184Writer, “Agentic AI adoption: A leader’s roadmap” — <https://writer.com/blog/responsible-ai-adoption-and-training/> (2025-09-10)

базовый AI RMF, а применяет те же четыре функции к рискам генеративных систем ¹⁸⁵. Для бизнеса ценность фреймворка — в его добровольности и модульности: компания может внедрять функции поэтапно, начиная с Map/Measure для пилотных use cases, и наращивать Govern по мере масштабирования. При этом добровольность не означает опциональности на практике: AI RMF де-факто стал языком, на котором регуляторы, страховщики и крупные заказчики оценивают зрелость управления ИИ у контрагентов, — и компании, работающие с enterprise-клиентами, всё чаще встречают отсылки к нему в анкетах вендорской проверки.

ISO/IEC 42001

ISO/IEC 42001:2023 (декабрь 2023) — первый международный сертифицируемый стандарт системы менеджмента ИИ (AIMS). Он построен по циклу PDCA и совместим с ISO 27001/9001, что позволяет встраивать AI-менеджмент в уже действующие системы информационной безопасности и качества. Принципиальное отличие от NIST AI RMF — возможность аудита аккредитованным органом и выдачи сертификата: для заказчиков, регуляторов и партнёров сертификат ISO/IEC 42001 служит внешне проверяемым доказательством зрелости governance. 13 марта 2026 года стандарт принят как европейский EN ISO/IEC 42001:2026 без изменений (CEN-CENELEC/JTC 21) ¹⁸⁶, что усилило его роль как практического моста к соответствию EU AI Act. Для компании это упрощает аргументацию перед бордом: инвестиция в AIMS одновременно повышает внутреннюю управляемость, даёт маркетинговый актив при продаже enterprise-заказчикам и сокращает будущий объём работ по регуляторному соответствию (о пересечении с EU AI Act — ниже в подразделе о комбинации фреймворков).

EU AI Act: сроки и штрафы

EU AI Act (Regulation (EU) 2024/1689) вступил в силу 1 августа 2024 года с поэтапным применением ¹⁸⁷:

Дата	Что вступает в силу
1 августа 2024	Вступление в силу регламента
2 февраля 2025	Запрещённые практики (ст. 5) и требование AI-грамотности (ст. 4)
2 августа 2025	Обязательства для GPAI-моделей и штрафной режим
2 августа 2026	Исходная дата общего применения обязательств для высокорисковых систем (Annex III)

Штрафная шкала трёхуровневая: до €35 млн или 7% глобального оборота (Tier 1 — за

185DSALTA / DeepInspect (разбор NIST AI 600-1) — <https://www.dsalta.com/resources/ai-compliance/nist-ai-600-1-generative-ai-profile-12-risks-explained> ; <https://www.deepinspect.ai/blog/nist-genai-profile> (публикация стандарта 26.07.2024; статьи 2026-08)

186KPMG; theiso42001.com; dm11.com.br — <https://kpmg.com/ch/en/insights/artificial-intelligence/iso-iec-42001.html> ; <https://theiso42001.com/> (стандарт — декабрь 2023; EN-версия — 13.03.2026)

187Alice Labs (обзор официального таймлайна ЕК); Scalewise; Haffa.ai — <https://alicelabs.ai/en/insights/eu-ai-act-timeline-2026> ; <https://scalewise.com/resources/eu-ai-act-timeline-ai-vendors-developers-2025-2026/> (2026-05–2026-08)

нарушение запретов ст. 5), €15 млн / 3% (Tier 2 — за нарушение большинства обязательств, включая требования к высокорисковым системам), €7,5 млн / 1,5% (Tier 3 — за предоставление недостоверной информации регуляторам) ¹⁸⁸. Градация воспроизводит логику GDPR: максимальные санкции привязаны к наиболее фундаментальным нарушениям, а потолок в процентах от оборота делает риск значимым для компаний любого масштаба. Для малого и среднего бизнеса ключевое практическое следствие — не размер штрафа, а необходимость с февраля 2025 года документально подтверждать AI-грамотность персонала (ст. 4) и отсутствие запрещённых практик в используемых системах.

Важная корректировка 2026 года: «Digital Omnibus» (май–июль 2026) перенёс дедлайны по высокорисковым системам — обязательства Annex III с 2 августа 2026 на 2 декабря 2027, Annex I — на 2 августа 2028 (Regulation (EU) 2026/1744, в силе с 27.07.2026); запреты и правила GPAI при этом уже действуют и не затронуты (confidence: medium-high, вторичные источники; проверять по Official Journal) ¹⁸⁹. Для бизнеса перенос не означает паузу: он лишь перераспределяет очередность работ. Инвентаризацию систем, классификацию рисков и обучение по ст. 4 никто не отменял, а штрафной режим для уже действующих обязательств сохраняется в полном объёме.

Регламент имеет экстерриториальное действие по модели GDPR: он применяется к любому провайдеру, выводящему ИИ на рынок ЕС, к деплоерам в ЕС, а также к провайдерам и деплоерам за пределами ЕС, если вывод их ИИ-системы используется в ЕС ¹⁹⁰. Это делает EU AI Act релевантным и для компаний за пределами Европы, работающих с европейскими клиентами. Отдельно стоит отметить обязательства для GPAI-моделей (general-purpose AI), действующие с 2 августа 2025 года: они адресованы прежде всего провайдерам моделей, но влияют и на деплоеров — через требования к документации, передаваемой по цепочке, и через необходимость понимать возможности и ограничения используемых базовых моделей. Практический порядок действий для компаний в скоупе: (1) инвентаризация ИИ-систем и определение их класса риска; (2) проверка на запрещённые практики по ст. 5; (3) для высокорисковых систем — построение системы управления рисками, технической документации и человеческого надзора в сроки с учётом переноса дедлайнов; (4) обучение персонала по ст. 4 — с февраля 2025 года это уже действующее обязательство ^{191 192}.

188 Alice Labs (обзор официального таймлайна ЕК); Scalewise; Haffa.ai — <https://alicelabs.ai/en/insights/eu-ai-act-timeline-2026> ; <https://scalewise.com/resources/eu-ai-act-timeline-ai-vendors-developers-2025-2026/> (2026-05–2026-08)

189 LogicGate; matthewbertram.com; DLA Piper (цит. по anthonywest.co.uk) — <https://www.logicgate.com/blog/the-eu-ai-act-requirements-deadlines-and-how-to-prepare/> ; <https://matthewbertram.com/blog/eu-ai-act-what-us-businesses-need-to-know> (2026-06–2026-07)

190 Alice Labs; CogitX — <https://alicelabs.ai/en/insights/eu-ai-act-timeline-2026> ; <https://cogitx.ai/blog/nist-ai-rmf-vs-iso-42001-vs-eu-ai-act-which-ai-governance-framework-does-your-enterprise-need> (2026-05/2026-07)

191 Alice Labs (обзор официального таймлайна ЕК); Scalewise; Haffa.ai — <https://alicelabs.ai/en/insights/eu-ai-act-timeline-2026> ; <https://scalewise.com/resources/eu-ai-act-timeline-ai-vendors-developers-2025-2026/> (2026-05–2026-08)

192 Alice Labs; CogitX — <https://alicelabs.ai/en/insights/eu-ai-act-timeline-2026> ; <https://cogitx.ai/blog/nist-ai-rmf-vs-iso-42001-vs-eu-ai-act-which-ai-governance-framework-does-your-enterprise-need> (2026-05/2026-07)

Комбинация фреймворков на практике

Сравнительная характеристика трёх основных рамок обобщена в таблице ниже (синтез 193194195196):

Критерий	NIST AI RMF / AI 600-1	ISO/IEC 42001:2023	EU AI Act
Статус	Добровольный фреймворк	Добровольный сертифицируемый стандарт	Обязательное регулирование (ЕС + экстерриториально)
Ядро	4 функции: Govern, Map, Measure, Manage; 12 gen AI-рисков в профиле	Система менеджмента AIMS по циклу PDCA, совместима с ISO 27001/9001	Риск-базированные категории: запреты, GPAI, высокий риск
Верификация	Самооценка	Аудит аккредитованным органом, сертификат	Государственный надзор, штрафы до €35 млн / 7% оборота
Роль в комбинации	Метод анализа рисков	Аудируемый каркас системы управления	Юридическое обязательство для систем в скоупе

Из таблицы следует практическое правило выбора: если вопрос компании — «с чего начать», ответ — NIST AI RMF, поскольку он не требует ни аудита, ни юридического скоупа и масштабируется от одного use case; если вопрос — «как доказать зрелость клиентам и регуляторам», ответ — ISO/IEC 42001; если вопрос — «что мы обязаны», ответ определяет EU AI Act через классификацию систем и экстерриториальный скоуп. Компании малого размера обычно ограничиваются первым, среднего — добавляют оценку скоупа третьего, крупные строят полную комбинацию¹⁹⁷.

193DSALTA / DeepInspect (пазбор NIST AI 600-1) — <https://www.dsalta.com/resources/ai-compliance/nist-ai-600-1-generative-ai-profile-12-risks-explained> ; <https://www.deepinspect.ai/blog/nist-genai-profile> (публикация стандарта 26.07.2024; статьи 2026-08)

194KPMG; theiso42001.com; dm11.com.br — <https://kpmg.com/ch/en/insights/artificial-intelligence/iso-iec-42001.html> ; <https://theiso42001.com/> (стандарт — декабрь 2023; EN-версия — 13.03.2026)

195Alice Labs (обзор официального таймлайна ЕК); Scalewise; Haffa.ai — <https://alicelabs.ai/en/insights/eu-ai-act-timeline-2026> ; <https://scalewise.com/resources/eu-ai-act-timeline-ai-vendors-developers-2025-2026/> (2026-05–2026-08)

196Strac; Orbit/Reconn; CogitX; Trustible — <https://www.strac.io/blog/ai-agent-governance-frameworks> ; <https://orbit.reconn.io/ai-governance-framework-comparison/> ; <https://trustible.ai/post/ai-governance-frameworks-nist-ai-rmf-eu-ai-act-iso-42001-compared/> (2026-07–2026-08)

197Strac; Orbit/Reconn; CogitX; Trustible — <https://www.strac.io/blog/ai-agent-governance-frameworks> ; <https://orbit.reconn.io/ai-governance-framework-comparison/> ; <https://trustible.ai/post/ai-governance-frameworks-nist-ai-rmf-eu-ai-act-iso-42001-compared/> (2026-07–2026-08)

Зрелые программы 2026 года не выбирают один фреймворк, а комбинируют: ISO/IEC 42001 выступает «аудируемым каркасом» (management system), NIST AI RMF используется внутри него как метод анализа рисков, а EU AI Act — как юридическое обязательство для систем в скоупе. По оценкам консультантов, пересечение ISO 42001 с требованиями EU AI Act составляет 40–50% (confidence: medium-high, практические оценки, не официальные цифры) ¹⁹⁸. Это позволяет компании строить единую систему управления ИИ, одновременно сертифицируемую, методологически обоснованную и юридически соответствующую. Экономия от комбинации существенна: без неё компания вынуждена параллельно вести риск-реестр «по NIST», процессы «по ISO» и чек-листы «по AI Act», утраивая документацию и рассинхронизируя контроли. Единый каркас устраняет это дублирование: один контроль обслуживает несколько обязательств, а маппинг требований ведётся один раз и пересматривается по мере изменения регуляторики. Для компаний, не работающих с рынком ЕС, минимальная разумная комбинация — NIST AI RMF как методология рисков плюс внутренние политики; для работающих с ЕС добавление ISO/IEC 42001 закрывает почти половину требований AI Act системным образом ¹⁹⁹.

Практическая последовательность внедрения комбинированного контура выглядит так: сначала инвентаризация ИИ-систем и риск-анализ по методологии NIST AI RMF (функции Map/Measure), затем оформление процессов управления в систему менеджмента по ISO/IEC 42001 (функция Govern, выросшая в AIMS), затем — для систем в скоупе EU AI Act — маппинг требований регулятора на уже построенные контроли. Такой порядок избегает «compliance ради compliance»: каждый следующий слой опирается на предыдущий, а дублирование работ минимизируется за счёт упомянутого 40–50%-ного пересечения стандарта и регулятора.

5.2. Shadow AI как вызов: статистика, кейс Samsung, DLP и GDPR

Shadow AI — главный практический вызов любой системы доступов: пока легальный путь сложнее нелегального, сотрудники выбирают последний. Данные 2025–2026 годов показывают, что явление носит не маргинальный, а системный характер — и потому требует не дисциплинарных, а архитектурных ответов.

Масштаб явления

Shadow AI — использование сотрудниками несогласованных ИИ-инструментов — приобрело массовый характер (confidence: medium, агрегаторы статистики; первичные отчёты требуют верификации) ²⁰⁰. Прежде чем приводить цифры, зафиксируем

198Strac; Orbit/Reconn; CogitX; Trustible — <https://www.strac.io/blog/ai-agent-governance-frameworks> ; <https://orbit.reconn.io/ai-governance-framework-comparison/> ; <https://trustible.ai/post/ai-governance-frameworks-nist-ai-rmf-eu-ai-act-iso-42001-compared/> (2026-07–2026-08)

199Strac; Orbit/Reconn; CogitX; Trustible — <https://www.strac.io/blog/ai-agent-governance-frameworks> ; <https://orbit.reconn.io/ai-governance-framework-comparison/> ; <https://trustible.ai/post/ai-governance-frameworks-nist-ai-rmf-eu-ai-act-iso-42001-compared/> (2026-07–2026-08)

200Adaptive Security (со ссылками на UpGuard, WalkMe, Gartner); AgileBrandGuide (со ссылкой на

определение: под shadow AI понимается любое рабочее использование ИИ-сервиса вне списка одобренных компанией инструментов — от личного аккаунта ChatGPT до «бесплатной» онлайн-нейросети для перевода документов. Именно широта явления и объясняет разброс оценок (59–80%+): разные исследования по-разному считают «использование» и выборки.

- более 80% сотрудников используют несогласованные ИИ-инструменты (UpGuard, 2025);
- 78% признаются в использовании несанкционированных ИИ, и только 7,5% получили полноценное обучение (WalkMe, 2025);
- 59% используют неодобренные ИИ; 75% из них передают потенциально чувствительные данные; 93% руководителей и топ-менеджеров — среди пользователей неодобренных ИИ (Cybernews, 2025);
- 23% работодателей не имеют формальной политики использования ИИ — то есть почти каждая четвёртая организация управляет массовым явлением без каких-либо правил;
- Shadow AI добавляет в среднем \$670 000 к стоимости утечки данных.

Особенно показателен последний структурный факт: shadow AI — не проблема «низов». Когда 93% руководителей и топ-менеджеров сами используют неодобренные инструменты, политика «запретить и контролировать» теряет легитимность на уровне примера. Данные Cyberhaven уточняют типологию утекающих данных: исходный код, РИ клиентов, финансовые данные, юридические документы, HR-данные; 77% сотрудников, использующих ИИ на работе, вставляли корпоративные или клиентские данные в промпты — чаще всего с личных неуправляемых аккаунтов ²⁰¹. Доля \$670 тыс. в стоимости утечки переводит проблему из плоскости комплаенса в плоскость финансового риска, понятного совету директоров.

Отдельно стоит подчеркнуть роль личных аккаунтов как главного канала утечки: в личном аккаунте не действуют ни DPA, ни корпоративные DLP-правила, ни аудит-логи, ни даже настройка «не обучать на моих данных» — исторически потребительские версии использовали диалоги для обучения моделей (с opt-out) ²⁰². С точки зрения риск-менеджмента это означает, что приоритетная цель — не отловить нарушителей, а свести к минимуму долю работы вне управляемого контура: каждый сотрудник, получивший удобный корпоративный доступ, — это закрытый канал утечки.

От списка фактов к стратегии реагирования

Совокупность этих данных задаёт чёткую управленческую рамку. Первый вывод: shadow AI — это не отклонение, а норма поведения при отсутствии легальной альтернативы;

Cybernews) — <https://www.adaptivesecurity.com/blog/shadow-ai-best-practices> ;
<https://www.adaptivesecurity.com/blog/why-do-employees-use-shadow-ai-the-psychology-p pressures-and-structural-gaps-behind-unauthorized> ; <https://agilebrandguide.com/ai-technology-statistics/> (2025–2026)
201DecryptionDigest (на базе Cyberhaven); Themio.ai — <https://www.decryptiondigest.com/blog/generative-ai-data-leakage-enterprise> ; <https://www.themio.ai/en/blog/chatgpt-copilot-gdpr-compliance-2026> (2026-07 / 2026-08)
202DecryptionDigest — <https://www.decryptiondigest.com/blog/generative-ai-data-leakage-enterprise> (2026-07-01)

поэтому целевая установка — не «искоренить», а «перевести в управляемое русло». Второй вывод: поскольку 75% пользователей неодобренных инструментов передают потенциально чувствительные данные, а 77% вставляют корпоративные данные в промпты с личных аккаунтов, приоритет №1 — предоставить корпоративные аккаунты с DPA всем, кто уже пользуется ИИ, до начала любых блокировок. Третий вывод: финансовая квантификация (+\$670 тыс. к стоимости утечки) превращает программу управления shadow AI из ИТ-инициативы в меру с защитимым ROI, что упрощает получение бюджета и спонсорства. Четвёртый вывод касается измерения: компании следует отслеживать долю shadow AI как регулярную метрику — через сетевую телеметрию, опросы и сравнение числа корпоративных аккаунтов с оценкой фактического использования; динамика этой доли — лучший индикатор того, работает ли стратегия «управляемого разрешения» или политика снова ушла в бумагу.

Кейс Samsung

Классический прецедент — Samsung (апрель 2023): три инцидента утечки менее чем за 20 дней — инженеры вставляли в ChatGPT исходный код полупроводникового оборудования, код тестовых программ и транскрипт конфиденциального совещания. В мае 2023 года компания запретила генеративный ИИ на устройствах и в корпоративных сетях. Однако запретительная стратегия оказалась временной: в июне 2026 Samsung официально снял многолетний запрет, разрешив сотрудникам DX Division ChatGPT, Gemini и Claude в рамках программы AI Transformation (AX) ²⁰³.

Хронология кейса поучительна в трёх отношениях. Во-первых, утечки произошли не из-за злого умысла, а из-за отсутствия легального инструмента: инженеры использовали самый доступный сервис для решения рабочих задач. Во-вторых, тотальный запрет на три года отложил, но не отменил внедрение — конкурентное давление в полупроводниковой отрасли потребовало возврата к ген. ИИ. В-третьих, возврат был организован как управляемая программа с несколькими одобренными моделями и рамкой AI Transformation, а не как свободный доступ. Эволюция Samsung иллюстрирует общий вывод: запрет не устраняет спрос, а загоняет использование в тень; управляемое легальное предоставление доступа эффективнее блокировок. Для других компаний кейс задаёт и временной ориентир: три года «запретительной паузы» — это цена отсутствия управляемой альтернативы в 2023 году; компания, сегодня внедряющая корпоративный доступ с DPA, DLP и AUP сразу, проходит тот же путь без многолетней потери производительности и конкурентного отставания.

DLP и GDPR

Ключевые технические и организационные меры контроля утечек ²⁰⁴:

203CIO.com; Authentech; TweakTown (май 2023) — <https://www.cio.com/article/4184660/samsung-which-previously-blocked-chatgpt-is-now-fully-adopting-three-generative-ai-models-and-accelerating-its-ai-initiative.html> ; <https://authentech.ai/blog/shadow-ai/samsung-chatgpt-incident/> (2023-05 / 2026-06-12)

204DecryptionDigest (на базе Cyberhaven); Themio.ai — <https://www.decryptiondigest.com/blog/generative-ai-data-leakage-enterprise> ; <https://www.themio.ai/en/blog/chatgpt-copilot-gdpr-compliance-2026> (2026-07 / 2026-

1. **Корпоративные тарифы с DPA** (data processing agreement) — договорная гарантия, что данные клиента не используются для обучения моделей;
2. **Классификация данных** — сотрудник должен понимать, какие категории данных допустимы для ввода в ИИ-инструменты;
3. **DLP-контроль буфера обмена и веб-форм** — техническая блокировка вставки чувствительных данных в неодобренные сервисы;
4. **Обучение** — только 7,5% сотрудников получили полноценное обучение, при том что это самая прямая мера снижения риска ²⁰⁵.

Порядок внедрения мер имеет значение. Классификация данных — фундамент: без неё и DLP-правила нечего распознавать, и сотрудникам непонятно, что именно запрещено. DLP-контроль в точке вставки (буфер обмена, веб-формы) эффективнее периметровых блокировок, поскольку срабатывает на действиях, а не на домене: один и тот же сервис может быть одобренным (корпоративный аккаунт) и неодобренным (личный). Обучение замыкает контур: оно не заменяет технические меры, но снижает долю случайных нарушений и формирует культуру, при которой сотрудник сам распознаёт чувствительный сценарий. Наконец, корпоративные тарифы с DPA должны предшествовать запретам — иначе запрет просто переводит трафик на личные устройства, где не работает ни один из контролей.

Принципиальный тезис по GDPR: соответствие ChatGPT/Copilot требованиям регламента — свойство конфигурации и использования, а не самого инструмента. Ключевые решения — выбор тарифа (consumer vs enterprise/API с DPA) и то, какие данные сотрудники вводят. DPIA по ст. 35 GDPR требуется при систематическом профилировании, автоматизированных решениях с юридическим эффектом или крупномасштабной обработке специальных категорий (например, скрининг кандидатов), но обычно не нужна для рутинного драфтинга и саммаризации ²⁰⁶. На практике это означает пороговую логику: одна и та же платформа в одной компании может требовать DPIA (HR-скрининг, скоринг клиентов) и не требовать её (подготовка черновиков, саммаризация встреч). Поэтому реестр use cases с указанием категорий данных и наличия решений о людях — не бюрократия, а механизм определения, где именно в компании включается полный регуляторный режим. Enterprise-уровень использования с DPA, письменной политикой и минимизацией данных квалифицируется практиками как защитимая позиция (defensible posture) ²⁰⁷. Риск различается по типу договора: корпоративные тарифы (M365 Copilot, Gemini for Workspace, GitHub Copilot Business) договорно запрещают обучение на данных клиента и предоставляют DPA; потребительские версии исторически использовали диалоги для обучения (с opt-out) ²⁰⁸. Фактические риски — персистентность данных у провайдера, бреши провайдера и нарушения GDPR/HIPAA при передаче регулируемых

08)

205Adaptive Security (со ссылками на UpGuard, WalkMe, Gartner); AgileBrandGuide (со ссылкой на Cybernews) — <https://www.adaptivesecurity.com/blog/shadow-ai-best-practices> ; <https://www.adaptivesecurity.com/blog/why-do-employees-use-shadow-ai-the-psychology-p pressures-and-structural-gaps-behind-unauthorized> ; <https://agilebrandguide.com/ai-technology-statistics/> (2025–2026)

206Themio.ai — <https://www.themio.ai/en/blog/chatgpt-copilot-gdpr-compliance-2026> (2026-08-31)

207Themio.ai — <https://www.themio.ai/en/blog/chatgpt-copilot-gdpr-compliance-2026> (2026-08-31)

208DecryptionDigest — <https://www.decryptiondigest.com/blog/generative-ai-data-leakage-enterprise> (2026-07-01)

данных неавторизованным обработчикам ²⁰⁹.

Практический вывод: разрыв между политикой и поведением закрывается не запретами, а предоставлением легального управляемого доступа плюс встроенными в инструменты политиками. Схема доступов должна проектироваться как «управляемое разрешение», а не «запрет» — это прямой урок кейса Samsung и статистики shadow AI ²¹⁰²¹¹.

5.3. Распределение доступов: RBAC/SCIM/SSO в корпоративных LLM-платформах

Техническая основа распределения доступов к корпоративным LLM-платформам — те же три механизма, что и в классическом IAM: SSO (единая аутентификация через корпоративного провайдера идентичностей), SCIM (автоматический провижининг и депровижининг пользователей и групп), RBAC (ролевая модель разрешений). Новизна 2025–2026 годов — в том, что эти механизмы стали штатными возможностями вендорских ИИ-платформ и распространились на не-человеческих акторов (агентов). Ниже — возможности ключевых платформ и типовые конфигурации по размеру компании.

Возможности платформ

ChatGPT Enterprise: детализация контролей

ChatGPT Enterprise предоставляет полный стек IAM: SAML/OIDC SSO с верификацией домена, SCIM-провижининг (Okta, Entra ID, Google Workspace), кастомный RBAC с гранулярными разрешениями (Canvas, agent mode, Codex, создание GPTs, веб-поиск, память, приложения), IP-allowlist, выбор региона хранения данных, Compliance API с неизменяемыми аудит-логами и интеграцией с SIEM. Тариф Business предлагает SSO, но не включает SCIM, RBAC, data residency и Compliance API — только фиксированные роли Owner/Admin/Member (confidence: high, подтверждено документацией OpenAI) ²¹².

Практический смысл гранулярности разрешений: компания может, например, разрешить рядовым сотрудникам диалоги и поиск, но отключить создание публичных GPTs или агентный режим до завершения оценки рисков — то есть настраивать не только «кто», но и «что именно» доступно каждой роли. Compliance API с экспортом неизменяемых логов в

209DecryptionDigest — <https://www.decryptiondigest.com/blog/generative-ai-data-leakage-enterprise> (2026-07-01)

210Adaptive Security (со ссылками на UpGuard, WalkMe, Gartner); AgileBrandGuide (со ссылкой на Cybernews) — <https://www.adaptivesecurity.com/blog/shadow-ai-best-practices> ; <https://www.adaptivesecurity.com/blog/why-do-employees-use-shadow-ai-the-psychology-p pressures-and-structural-gaps-behind-unauthorized> ; <https://agilebrandguide.com/ai-technology-statistics/> (2025–2026)

211CIO.com; Authentech; TweakTown (май 2023) — <https://www.cio.com/article/4184660/samsung-which-previously-blocked-chatgpt-is-now-fully-adopting-three-generative-ai-models-and-accelerating-its-ai-initiative.html> ; <https://authentech.ai/blog/shadow-ai/samsung-chatgpt-incident/> (2023-05 / 2026-06-12)

212IntuitionLabs (обзор документации OpenAI); OpenAI developers docs; Xenoss — <https://intuitionlabs.ai/articles/chatgpt-enterprise-admin-controls-security> ; <https://developers.openai.com/codex/enterprise/work-admin-faq> ; <https://xenoss.io/blog/gpt-vs-open-source-models-security> (2025-11–2026-04)

SIEM переводит использование ИИ в разряд аудируемых процессов: расследование инцидентов, ответы регуляторам и внутренние проверки опираются на объективный след, а не на опросы сотрудников. IP-allowlist и выбор региона хранения данных закрывают соответственно сетевой периметр и требования к резидентности данных.

Microsoft Copilot Studio / Agent 365 и категория agent identity

Microsoft Copilot Studio / Agent 365: агенты представляются как идентичности в Microsoft Entra и управляются через Conditional Access, role-based и attribute-based access control, DLP-политики, data residency и централизованную плоскость наблюдения (control plane). Microsoft 365 Copilot не обучает модели на данных клиентов и наследует разрешения и метки Purview (confidence: high) ²¹³. Архитектурно это означает, что ИИ-агент получает тот же режим управления, что и сотрудник: его доступ определяется ролью и атрибутами, его действия логируются, его можно заблокировать тем же Conditional Access. Наследование разрешений и меток Purview снимает классическую проблему «ИИ видит больше, чем пользователь»: ассистент оперирует в пределах уже выданных сотруднику прав на документы, а метки чувствительности распространяются на сгенерированный контент.

Возникновение категории «agent identity» — одна из ключевых тем 2026 года: отсутствие персистентной идентичности и аудит-трейла для не-человеческих акторов становится драйвером новых инвестиций в ИИ-governance инфраструктуру ²¹⁴. Фон этого тренда — общая картина рынка: McKinsey фиксирует 88% принятия ИИ при лишь 6% «high performers», а Gartner прогнозирует отмену 40% агентных ИИ-проектов к 2027 году (confidence: medium, вторичная сводка) ²¹⁵. Управление агентами как полноценными идентичностями — институциональный ответ на риск неконтролируемого «agent sprawl»: сценария, при котором подразделения массово создают агентов без реестра, владельцев и отзываемых доступов. Для компаний это означает простое правило: ни один агент не выходит в продакшн без зарегистрированной идентичности, назначенного владельца-человека и набора task-scoped инструментов — по аналогии с тем, что ни один сотрудник не работает без учётной записи ²¹⁶.

Сравнение платформ и выбор стека

Сравнение платформ (2026): ChatGPT Enterprise — SSO+SCIM+RBAC; Microsoft 365 Copilot — Entra ID, Conditional Access, tenant policies; Gemini for Workspace — админ-консоль Workspace, наследует DLP; Claude Enterprise — SSO, ролевые разрешения ²¹⁷.

213Microsoft Learn (официальная документация); BCS365 — <https://learn.microsoft.com/en-us/microsoft-copilot-studio/security-and-governance> ; <https://bcs365.com/insights/microsoft-copilot-security-risks-enterprise> (2026-08)

214anthonywest.co.uk (сводный трекей отчётов: McKinsey, Gartner, CB Insights, MIT SMR, Forrester, DLA Piper, CSA) — <https://anthonywest.co.uk/research/ai-transformation-projects> (2026-07-15)

215anthonywest.co.uk (сводный трекей отчётов: McKinsey, Gartner, CB Insights, MIT SMR, Forrester, DLA Piper, CSA) — <https://anthonywest.co.uk/research/ai-transformation-projects> (2026-07-15)

216arXiv:2512.20535 (ARBITER); arXiv:2509.08646 (Plan-then-Execute) — <https://arxiv.org/html/2512.20535v1> ; <https://arxiv.org/html/2509.08646v1> (2025-09 / 2025-12)

217365DigitalConsulting (сравнительный обзор) — <https://365digitalconsulting.com/article/chatgpt-enterprise-vs->

Выбор платформы обычно привязан к существующему стеку компании: организация на Microsoft 365 получает governance «из коробки» через Entra и Purview; организация на Google Workspace — через админ-консоль Workspace; нейтральные или мультивендорные среды чаще выбирают ChatGPT Enterprise или Claude Enterprise с отдельным IAM-контуром ²¹⁸. Важно, что во всех четырёх случаях enterprise-уровень контроля доступен только на старших тарифах — это создаёт прямую связку «размер бюджета → уровень управляемости».

Мультивендорная реальность большинства компаний (одновременно несколько LLM-платформ, как в кейсе Samsung-2026 с тремя одобренными моделями ²¹⁹) требует единого слоя политики поверх платформ: один список одобренных инструментов, единая классификация данных и согласованные DLP-правила, иначе сотрудник получает разные правила в разных окнах, а аудит — разрозненные логи. Там, где платформа не наследует корпоративный IAM, минимальный стандарт — обязательный SSO и централизованное отключение доступа при увольнении.

Модель доступа по размеру компании

Размер бизнеса определяет архитектуру доступа не только через бюджет, но и через зрелость ИТ-функции: механизмы вроде SCIM требуют действующего IdP и дисциплины управления группами, которые сами по себе являются признаком зрелого IAM ²²⁰:

- **Малые компании** (ChatGPT Business, \$25/польз./мес.) — фиксированные роли Owner/Admin/Member, ручное управление пользователями, без SCIM/RBAC. Управление сводится к дисциплине: своевременное удаление уволенных сотрудников, раздача роли Admin только ответственным, письменная AUP;
- **Крупные предприятия** — Enterprise-тарифы с кастомными ролями, SCIM-синхронизацией групп (автоматический provisioning/deprovisioning при найме/увольнении), spend-контролями и аудитом. Enterprise-тип требуется для любой организации, которой нужны SCIM, аудит-логи и верификация домена ²²¹.

Средний бизнес оказывается в промежуточной зоне: требования к управляемости уже выросли (регуляторы, клиенты-энтерпрайзы, внутренний аудит), а бюджета на enterprise-тарифы может не быть. Практический компромисс — выбор платформы, наследующей существующий IAM-стек (например, Copilot при наличии M365), чтобы получить SSO, Conditional Access и DLP без отдельной интеграции ²²².

microsoft-copilot-vs-google-gemini-vs-claude/ (2026-07-06)
218365DigitalConsulting (сравнительный обзор) — <https://365digitalconsulting.com/article/chatgpt-enterprise-vs-microsoft-copilot-vs-google-gemini-vs-claude/> (2026-07-06)
219CIO.com; Authentech; TweakTown (май 2023) — <https://www.cio.com/article/4184660/samsung-which-previously-blocked-chatgpt-is-now-fully-adopting-three-generative-ai-models-and-accelerating-its-ai-initiative.html> ; <https://authentech.ai/blog/shadow-ai/samsung-chatgpt-incident/> (2023-05 / 2026-06-12)
220Stitchflow (гайд по управлению пользователями ChatGPT); IntuitionLabs — <https://www.stitchflow.com/user-management/chatgpt-openai/manual> (2026-03-04)
221Stitchflow (гайд по управлению пользователями ChatGPT); IntuitionLabs — <https://www.stitchflow.com/user-management/chatgpt-openai/manual> (2026-03-04)
222365DigitalConsulting (сравнительный обзор) — <https://365digitalconsulting.com/article/chatgpt-enterprise-vs-microsoft-copilot-vs-google-gemini-vs-claude/> (2026-07-06)

Академические работы 2025–2026 развивают RBAC для LLM/RAG-систем: ARBITER (arXiv, декабрь 2025) реализует RBAC для RAG со слоистой валидацией ввода/вывода и ролевым поиском (85% accuracy, 89% F1); исследования агентных архитектур рекомендуют комбинировать RBAC (долговременные ролевые ограничения) с task-scoped назначением инструментов (принцип наименьших привилегий) как модель defence-in-depth для LLM-агентов ²²³. Это указывает направление развития: от «роль определяет доступ к чату» к «роль определяет, какие данные видит RAG и какие инструменты получает агент под конкретную задачу».

Жизненный цикл доступа и контроль расходов

Помимо статического распределения ролей, enterprise-управление включает автоматизацию жизненного цикла доступа (joiner–mover–leaver): SCIM-синхронизация с HR/IdP-системами обеспечивает выдачу доступа при найме, пересмотр при смене роли и гарантированный отзыв при увольнении ²²⁴. Для LLM-платформ это критично вдвойне: «забытый» аккаунт — это не только лицензионные потери, но и сохраняющийся канал доступа к корпоративным перепискам с моделью. Второй операционный слой — spend-контроли: лимиты расходов на API и премиальные функции по командам, предотвращающие неконтролируемый рост счетов при массовом adoption ²²⁵. Оба механизма доступны только на Enterprise-тарифах, что ещё раз подчёркивает границу между SMB- и enterprise-моделями доступа.

Acceptable Use Policies

Acceptable Use Policy (AUP) — точка, где governance встречается с повседневным поведением сотрудника. Именно на уровне AUP абстрактные требования фреймворков превращаются в конкретные «можно/нельзя», понятные без юридической подготовки.

Типовая корпоративная AI Acceptable Use Policy (2026) включает восемь разделов, среди которых: интеллектуальная собственность (результаты работы с ИИ принадлежат компании, но требуют человеческого авторства для публикации/регистрации), раскрытие и маркировка ИИ-контента, запрещённые использования — ввод конфиденциальных/персональных данных в неодобренные инструменты, финальные решения о людях без человеческой проверки, обход списка одобренных инструментов ²²⁶. При этом 23% работодателей не имеют никакой формальной политики использования ИИ-инструментов (Cybernews, 2025) — то есть почти четверть компаний управляют AI-рисками исключительно «на словах» ²²⁷ (confidence: medium).

223arXiv:2512.20535 (ARBITER); arXiv:2509.08646 (Plan-then-Execute) — <https://arxiv.org/html/2512.20535v1> ; <https://arxiv.org/html/2509.08646v1> (2025-09 / 2025-12)

224Stitchflow (гайд по управлению пользователями ChatGPT); IntuitionLabs — <https://www.stitchflow.com/user-management/chatgpt-openai/manual> (2026-03-04)

225Stitchflow (гайд по управлению пользователями ChatGPT); IntuitionLabs — <https://www.stitchflow.com/user-management/chatgpt-openai/manual> (2026-03-04)

226GC.ai (шаблон политики) — <https://gc.ai/blog/ai-acceptable-use-policy-template> (2026-08-25)

227AgileBrandGuide (цит. Cybernews) — <https://agilebrandguide.com/ai-technology-statistics/> (2025–2026)

Минимальный состав рабочей AUP для компании любого размера: (1) список одобренных инструментов и тарифов; (2) классификация данных, допустимых к вводу; (3) запрет автоматизированных финальных решений о людях без человеческой проверки; (4) правила маркировки ИИ-контента; (5) ответственность и порядок сообщения об инцидентах. Для среднего и крупного бизнеса добавляются разделы об интеллектуальной собственности, взаимодействии с регуляторами и периодическом пересмотре политики ²²⁸.

Характерный тренд 2026 года — встраивание политик непосредственно в инструменты: ChatGPT Enterprise включает «AI policy modal» — организационную политику использования ИИ, показываемую пользователям при входе и каждые 30 дней ²²⁹. Это закрывает разрыв между формальной политикой и фактическим поведением: напоминание появляется в точке использования, а не в документе на портале. С учётом статистики shadow AI (раздел 5.2) встроенные политики следует рассматривать как более надёжный канал принуждения, чем однократное ознакомление сотрудника с PDF при найме.

5.4. Рекомендуемая схема распределения доступов: практическая матрица

На основе рассмотренных практик (возможности платформ, shadow AI, требования GDPR/EU AI Act) ниже предложена практическая матрица «роль сотрудника × уровень доступа к ИИ-инструментам». Матрица построена на четырёх уровнях доступа: **L1** — чтение/базовые запросы в одобренных инструментах без чувствительных данных; **L2** — полноценная работа в корпоративной платформе с внутренними данными; **L3** — работа с конфиденциальными данными, создание GPTs/агентов, интеграции; **L4** — администрирование платформы, governance, аудит. Матрица носит рекомендательный характер: конкретные разрешения настраиваются в административных консолях платформ (см. 5.3), а пороги «чувствительности» данных определяются корпоративной классификацией. Для каждого размера бизнеса указана не только целевая конфигурация, но и форма её реализации: у малого бизнеса это дисциплина и письменные правила, у среднего — групповые политики IdP, у крупного — автоматизированная синхронизация ролей и аудит.

Роль сотрудника	Малый бизнес (фиксированные роли Owner/Admin/Member)	Средний бизнес (корпоративная платформа + SSO)	Крупный бизнес (кастомный RBAC + SCIM + аудит)
Рядовой сотрудник (операционные функции)	Member; только одобренный инструмент; AUP при онбординге — L1	L1–L2: корпоративный аккаунт, запрет ввода конфиденциальных данных, DLP-	L1–L2 по ролевой группе SCIM; policy modal каждые 30 дней; аудит-логи в SIEM

228GC.ai (шаблон политики) — <https://gc.ai/blog/ai-acceptable-use-policy-template> (2026-08-25)

229IntuitionLabs (по документации OpenAI) — <https://intuitionlabs.ai/articles/chatgpt-enterprise-admin-controls-security> (2026-04-18)

Роль сотрудника	Малый бизнес (фиксированные роли Owner/Admin/Member)	Средний бизнес (корпоративная платформа + SSO)	Крупный бизнес (кастомный RBAC + SCIM + аудит)
Аналитик / специалист по данным	Member с инструкцией по классификации данных — L2	контроль буфера обмена L2–L3: доступ к внутренним данным, отдельные проекты (Projects/GPTs)	L3: кастомная роль (Canvas, память, веб- поиск), data residency, DPIA при чувствительных сценариях
Разработчик / ML- инженер	Member + отдельный API-ключ с квотами — L2–L3	L3: Codex/агентный режим, API с spend- контролем	L3: Codex, agent mode, RAG с ролевым поиском, task-scoped инструменты для агентов
Менеджер / руководитель направления	Member; тренинг по рискам — L1–L2	L2: доступ к корпоративной платформе, запрет финальных решений о людях без проверки	L2: role-based доступ; обязательная человеческая проверка решений о людях (AUP)
HR / Legal (спецкатегории данных)	Member с запретом ввода PII — L1	L1–L2: DPIA при скрининге кандидатов; только enterprise-тариф с DPA	L2: строгий DLP, DPIA обязательна, автоматизированные решения с юридическим эффектом — под контролем
AI Champion / power user	Admin (помимо основной роли) — L3	L3: создание департаментских GPTs/агентов, обучение коллег	L3: создание GPTs и shared projects, наставничество, связка с CoE
IT-администратор ИИ-платформы	Owner — L4	L4: SSO-настройка, управление лицензиями, базовые отчёты	L4: SCIM- провижининг, IP- allowlist, Compliance API, интеграция с SIEM
AI Governance Officer / CISO / DPO	(функция совмещается) — L4- надзор	Надзор: реестр use cases, AUP, DLP- политики	Надзор: кастомный RBAC, аудит-логи, model cards, DPIA, отчётность борду

Комментарии к матрице. Матрица воплощает три правила, повторяющиеся во всех источниках главы: ширина доступа важнее его селективности (практика superscalers²³⁰); чувствительные данные существуют только внутри корпоративного контура с DPA²³¹; решения о людях всегда проходят человеческую проверку²³²²³³. Рядовой сотрудник во всех трёх сегментах остаётся на уровне L1–L2: его риск-профиль закрывается не ограничением доступа, а обучением и техническими барьерами (DLP, policy modal) — это согласуется с практикой superscalers, дающих широкий доступ максимальному числу работников²³⁴. Для разработчиков и аналитиков уровень L3 открывается только при корпоративном тарифе с DPA: работа с конфиденциальными данными в потребительских сервисах недопустима независимо от размера компании²³⁵. Строки HR/Legal отражают специфику GDPR: именно здесь чаще всего возникают сценарии, запускающие обязательную DPIA, — скрининг кандидатов и автоматизированные решения о клиентах²³⁶. Наконец, разделение «администратор платформы» и «governance-надзор» на enterprise-уровне воспроизводит распределённую модель ответственности: CISO владеет контрольным базисом, DPO — оценками влияния, CAIO/CDAO — реестром use cases и риск-тиерами²³⁷.

Сценарии применения по размеру компании. В малой компании матрица реализуется без какой-либо технической инфраструктуры: Owner (владелец бизнеса или IT-ответственный) ведёт список пользователей вручную, Admin — доверенный сотрудник, все остальные — Member с одинаковыми правами; единственный реальный рычаг дифференциации — письменная AUP и выборочная выдача API-ключей разработчикам²³⁸. Главные риски этого сегмента — «забытые» аккаунты уволенных и использование личных аккаунтов в обход корпоративного; оба закрываются дисциплиной offboarding и доступностью корпоративного тарифа каждому сотруднику, а не техническими средствами.

В средней компании добавляется SSO и появляется возможность групповых политик: доступ L3 (создание GPTs, работа с внутренними данными) выдаётся группам «аналитики» и «разработка» через группы IdP, а DLP-правила ограничивают вставку чувствительных данных для всех остальных. На этом этапе критично не перестроить сложность: если платформа не поддерживает кастомный RBAC, не следует эмулировать

230Deloitte Insights Magazine Issue 33 —

https://www2.deloitte.com/content/dam/insights/articles/glob187719_dim33/Deloitte-Insights-Magazine-Issue-33.pdf (2024)

231DecryptionDigest — <https://www.decryptiondigest.com/blog/generative-ai-data-leakage-enterprise> (2026-07-01)

232GC.ai (шаблон политики) — <https://gc.ai/blog/ai-acceptable-use-policy-template> (2026-08-25)

233Themio.ai — <https://www.themio.ai/en/blog/chatgpt-copilot-gdpr-compliance-2026> (2026-08-31)

234Deloitte Insights Magazine Issue 33 —

https://www2.deloitte.com/content/dam/insights/articles/glob187719_dim33/Deloitte-Insights-Magazine-Issue-33.pdf (2024)

235DecryptionDigest — <https://www.decryptiondigest.com/blog/generative-ai-data-leakage-enterprise> (2026-07-01)

236Themio.ai — <https://www.themio.ai/en/blog/chatgpt-copilot-gdpr-compliance-2026> (2026-08-31)

237Knostic, “Primer: AI Governance Roles and Responsibilities”; Diligent, “AI governance: A guide for boards” — <https://www.knostic.ai/blog/ai-governance-roles-stakeholders> ; <https://www.diligent.com/resources/blog/ai-governance> (2025-10-15 ; 2026-06-12)

238Stitchflow (гайд по управлению пользователями ChatGPT); IntuitionLabs —

<https://www.stitchflow.com/user-management/chatgpt-openai/manual> (2026-03-04)

его десятками ручных исключений — это не масштабируется и создаёт ложное ощущение контроля.

В крупной компании матрица превращается в конфигурацию SCIM-групп и кастомных ролей: каждая строка таблицы соответствует ролевой группе в IdP, синхронизируемой с платформой; изменение роли сотрудника в HR-системе автоматически меняет его доступ; все действия L3–L4 пишутся в неизменяемые аудит-логи и выгружаются в SIEM²³⁹²⁴⁰. Для агентных сценариев каждая строка матрицы дополнительно получает «теневую» колонку: какие агенты и с какими идентичностями Entra могут действовать от имени этой роли и с каким набором task-scoped инструментов²⁴¹²⁴².

Ключевые принципы схемы. Во-первых, доступ проектируется как «управляемое разрешение», а не запрет: массовый shadow AI (59–80% сотрудников) делает формальные запреты бессмысленными без предоставления легального управляемого доступа²⁴³. Во-вторых, жизненный цикл доступа автоматизируется: на enterprise-уровне SCIM-синхронизация групп обеспечивает автоматический provisioning/deprovisioning при найме и увольнении²⁴⁴. В-третьих, политики встраиваются в инструменты (policy modal, DLP-контроль форм и буфера обмена), а не существуют только на бумаге²⁴⁵. В-четвёртых, для агентных сценариев 2026 года принцип наименьших привилегий реализуется комбинацией долговременного RBAC и task-scoped назначения инструментов, а сами агенты получают персистентные идентичности (agent identity) с аудит-трейлом²⁴⁶²⁴⁷. Наконец, все автоматизированные решения о людях требуют человеческой проверки — это и типовый пункт AUP²⁴⁸, и требование DPIA-практики по GDPR²⁴⁹.

239IntuitionLabs (обзор документации OpenAI); OpenAI developers docs; Xenoss —

<https://intuitionlabs.ai/articles/chatgpt-enterprise-admin-controls-security> ;

<https://developers.openai.com/codex/enterprise/work-admin-faq> ; <https://xenoss.io/blog/gpt-vs-open-source-models-security> (2025-11–2026-04)

240Stitchflow (гайд по управлению пользователями ChatGPT); IntuitionLabs —

<https://www.stitchflow.com/user-management/chatgpt-openai/manual> (2026-03-04)

241Microsoft Learn (официальная документация); BCS365 — [https://learn.microsoft.com/en-us/microsoft-](https://learn.microsoft.com/en-us/microsoft-copilot-studio/security-and-governance)

[copilot-studio/security-and-governance](https://learn.microsoft.com/en-us/microsoft-copilot-studio/security-and-governance) ; <https://bcs365.com/insights/microsoft-copilot-security-risks-enterprise> (2026-08)

242arXiv:2512.20535 (ARBITER); arXiv:2509.08646 (Plan-then-Execute) — <https://arxiv.org/html/2512.20535v1> ;

<https://arxiv.org/html/2509.08646v1> (2025-09 / 2025-12)

243Adaptive Security (со ссылками на UpGuard, WalkMe, Gartner); AgileBrandGuide (со ссылкой на

Cybernews) — <https://www.adaptivesecurity.com/blog/shadow-ai-best-practices> ;

<https://www.adaptivesecurity.com/blog/why-do-employees-use-shadow-ai-the-psychology-pressures-and-structural-gaps-behind-unauthorized> ; <https://agilebrandguide.com/ai-technology-statistics/> (2025–2026)

244Stitchflow (гайд по управлению пользователями ChatGPT); IntuitionLabs —

<https://www.stitchflow.com/user-management/chatgpt-openai/manual> (2026-03-04)

245IntuitionLabs (по документации OpenAI) — <https://intuitionlabs.ai/articles/chatgpt-enterprise-admin-controls-security> (2026-04-18)

246arXiv:2512.20535 (ARBITER); arXiv:2509.08646 (Plan-then-Execute) — <https://arxiv.org/html/2512.20535v1> ;

<https://arxiv.org/html/2509.08646v1> (2025-09 / 2025-12)

247Microsoft Learn (официальная документация); BCS365 — [https://learn.microsoft.com/en-us/microsoft-](https://learn.microsoft.com/en-us/microsoft-copilot-studio/security-and-governance)

[copilot-studio/security-and-governance](https://learn.microsoft.com/en-us/microsoft-copilot-studio/security-and-governance) ; <https://bcs365.com/insights/microsoft-copilot-security-risks-enterprise> (2026-08)

248GC.ai (шаблон политики) — <https://gc.ai/blog/ai-acceptable-use-policy-template> (2026-08-25)

249Themio.ai — <https://www.themio.ai/en/blog/chatgpt-copilot-gdpr-compliance-2026> (2026-08-31)

5.5. Выводы по главе: минимальный контур governance по размеру бизнеса

Глава 5 показала, что управление ИИ в 2024–2026 годах складывается из трёх взаимозависимых слоёв: нормативного (фреймворки и регуляторика, задающие внешние требования), поведенческого (shadow AI и AUP, определяющие фактическое использование) и технического (RBAC/SCIM/SSO, agent identity, DLP, обеспечивающие принуждение). Ни один из слоёв не работает отдельно: политика без технических контролей игнорируется, контроли без политики не имеют легитимности, а оба без учёта регуляторики создают юридический риск. Собранные данные позволяют сформулировать дифференцированный минимум управления ИИ. Общая логика: каждый следующий уровень зрелости добавляет не «ещё один документ», а новый технический или институциональный механизм принуждения — от письменной политики к SSO, от SSO к SCIM и аудиту, от аудита к управлению идентичностями агентов.

- **Малый бизнес.** Достаточно корпоративного тарифа уровня Business с фиксированными ролями Owner/Admin/Member, одностраничной AUP, явного списка одобренных инструментов и запрета ввода конфиденциальных данных в личные аккаунты²⁵⁰²⁵¹. Формальные фреймворки преждевременны; метод риск-анализа NIST AI RMF может применяться упрощённо²⁵². Критично не «экономить» на двух вещах: корпоративном тарифе для каждого активного пользователя (иначе личные аккаунты гарантированно станут каналом утечки²⁵³²⁵⁴) и дисциплине отзыва доступа при увольнении — в отсутствие SCIM это ручной, но обязательный процесс.
- **Средний бизнес.** Добавляются SSO, классификация данных, DLP-контроль буфера обмена и форм, обучение с разделением на общий и ролевой модули, назначение ответственного за governance (функция может совмещаться)²⁵⁵²⁵⁶. Если компания работает с рынком ЕС, уже на этом этапе целесообразна оценка попадания под EU AI Act с учётом его экстерриториального действия²⁵⁷. Характерная ошибка сегмента — «платёжеспособный хаос»: бюджет на несколько ИИ-платформ есть, а

250Stitchflow (гайд по управлению пользователями ChatGPT); IntuitionLabs —

<https://www.stitchflow.com/user-management/chatgpt-openai/manual> (2026-03-04)

251GC.ai (шаблон политики) — <https://gc.ai/blog/ai-acceptable-use-policy-template> (2026-08-25)

252DSALTA / DeepInspect (разбор NIST AI 600-1) — <https://www.dsalta.com/resources/ai-compliance/nist-ai-600-1-generative-ai-profile-12-risks-explained> ; <https://www.deepinspect.ai/blog/nist-genai-profile> (публикация стандарта 26.07.2024; статьи 2026-08)

253DecryptionDigest (на базе Cyberhaven); Themio.ai — <https://www.decryptiondigest.com/blog/generative-ai-data-leakage-enterprise> ; <https://www.themio.ai/en/blog/chatgpt-copilot-gdpr-compliance-2026> (2026-07 / 2026-08)

254DecryptionDigest — <https://www.decryptiondigest.com/blog/generative-ai-data-leakage-enterprise> (2026-07-01)

255DecryptionDigest (на базе Cyberhaven); Themio.ai — <https://www.decryptiondigest.com/blog/generative-ai-data-leakage-enterprise> ; <https://www.themio.ai/en/blog/chatgpt-copilot-gdpr-compliance-2026> (2026-07 / 2026-08)

256Trail, “How to build AI Literacy with AI Trainings Best Practices” — <https://www.trail-ml.com/blog/building-ai-literacy-under-the-ai-act-best-practices> (2025-08-26)

257Alice Labs; CogitX — <https://alicelabs.ai/en/insights/eu-ai-act-timeline-2026> ; <https://cogitx.ai/blog/nist-ai-rmf-vs-iso-42001-vs-eu-ai-act-which-ai-governance-framework-does-your-enterprise-need> (2026-05/2026-07)

единой политики и владельца governance нет; лечится не сокращением числа инструментов, а введением единого реестра use cases и одного ответственного за политику.

- **Крупный бизнес.** Enterprise-тарифы с кастомным RBAC, SCIM, аудит-логами в SIEM и верификацией домена ²⁵⁸; комбинация ISO/IEC 42001 (аудируемый каркас) + NIST AI RMF (метод рисков) + EU AI Act (юридические обязательства), закрывающая 40–50% требований регулятора системным образом ²⁵⁹; распределённая модель ответственности (борд — CDAO/CAIO — CISO — DPO) ²⁶⁰; управление идентичностями агентов как новый обязательный слой ^{261,262}.

Общий вектор 2024–2026 годов: governance перестаёт быть документом и становится инженерной дисциплиной — политики встроены в платформы, доступы синхронизированы с HR-системами, а каждый агент имеет идентичность и аудит-трейл. Компании, которые строят этот контур одновременно с внедрением инструментов, избегают и сценария «Samsung-2023» (утечки и запрет), и сценария «shadow AI» (неконтролируемое использование в обход политик). Обратная сторона вектора — рост требований к компетенциям тех, кто этот контур эксплуатирует: AI Governance Officer, CISO и DPO должны говорить на общем языке рисков ИИ, а борд — получать регулярную отчётность о реестре use cases, инцидентах и соответствии регуляторным срокам ^{263,264}. Governance, таким образом, замыкает контур, начатый в главе 4: организационная структура без распределения доступов порождает shadow AI, а доступы без оргструктуры — некому контролировать.

В качестве практической отправной точки независимо от размера компании можно рекомендовать последовательность «первых шагов»: (1) инвентаризация фактического использования ИИ — включая неформальный опрос, поскольку официальная картина систематически занижает shadow AI ²⁶⁵; (2) принятие базовой AUP и публикация списка

258Stitchflow (гайд по управлению пользователями ChatGPT); IntuitionLabs —

<https://www.stitchflow.com/user-management/chatgpt-openai/manual> (2026-03-04)

259Strac; Orbit/Reconn; CogitX; Trustible — <https://www.strac.io/blog/ai-agent-governance-frameworks> ; <https://orbit.reconn.io/ai-governance-framework-comparison/> ; <https://trustible.ai/post/ai-governance-frameworks-nist-ai-rmf-eu-ai-act-iso-42001-compared/> (2026-07–2026-08)

260Knostic, “Primer: AI Governance Roles and Responsibilities”; Diligent, “AI governance: A guide for boards” — <https://www.knostic.ai/blog/ai-governance-roles-stakeholders> ; <https://www.diligent.com/resources/blog/ai-governance> (2025-10-15 ; 2026-06-12)

261Microsoft Learn (официальная документация); BCS365 — <https://learn.microsoft.com/en-us/microsoft-copilot-studio/security-and-governance> ; <https://bcs365.com/insights/microsoft-copilot-security-risks-enterprise> (2026-08)

262anthonywest.co.uk (сводный трекеп отчётов: McKinsey, Gartner, CB Insights, MIT SMR, Forrester, DLA Piper, CSA) — <https://anthonywest.co.uk/research/ai-transformation-projects> (2026-07-15)

263Taylor Root, “AI Governance Officer job description” — <https://www.taylorroot.com/resources/job-description-templates/ai-governance-officer-job-description/> (2026-04-28)

264Knostic, “Primer: AI Governance Roles and Responsibilities”; Diligent, “AI governance: A guide for boards” — <https://www.knostic.ai/blog/ai-governance-roles-stakeholders> ; <https://www.diligent.com/resources/blog/ai-governance> (2025-10-15 ; 2026-06-12)

265Adaptive Security (со ссылками на UpGuard, WalkMe, Gartner); AgileBrandGuide (со ссылкой на Cybernews) — <https://www.adaptivesecurity.com/blog/shadow-ai-best-practices> ; <https://www.adaptivesecurity.com/blog/why-do-employees-use-shadow-ai-the-psychology-pressure-and-structural-gaps-behind-unauthorized> ; <https://agilebrandguide.com/ai-technology-statistics/> (2025–2026)

одобренных инструментов ²⁶⁶; (3) перевод активных пользователей на корпоративные тарифы с DPA ²⁶⁷; (4) назначение владельца governance и запуск реестра use cases ²⁶⁸; (5) только затем — выбор фреймворка (NIST AI RMF / ISO 42001) и масштабирование технических контролей ^{269,270}. Такой порядок сначала закрывает самый массовый и самый дорогой риск (утечки через теневое использование) и лишь потом наращивает формальную зрелость — в соответствии с главным уроком собранных данных: доступ проектируется как управляемое разрешение, а не как запрет.

Оговорки об уверенности в данных глав 4–5

Значительная часть количественных оценок в этих главах получена из отраслевых и консалтинговых источников разной степени авторитетности. Наивысшую уверенность (high) имеют: данные McKinsey State of AI 2025 (централизация governance, oversight CEO, 88% принятия), Gartner CDAO Survey 2025, IBM IBV по роли CAIO, характеристики тарифов ChatGPT Enterprise (подтверждены документацией OpenAI), возможности Microsoft Copilot Studio/Entra (официальная документация Microsoft), хронология EU AI Act и кейс Samsung. Средняя уверенность (medium) — у статистики shadow AI из агрегаторов (UpGuard, WalkMe, Cybernews — первичные отчёты стоит верифицировать), у оценки Deloitte «+73% к результатам масштабирования» (вторичный источник), у цифры «71% откатов hub-and-spoke» (единичный источник), у практических оценок пересечения ISO 42001 и EU AI Act (40–50%) и у данных о сокращении роли prompt-инженера. Пониженная уверенность (low-medium) — у ориентира выбора операционной модели по численности сотрудников, заимствованного из смежной области специализированных команд. Даты «Digital Omnibus» по переносу дедлайнов EU AI Act основаны на вторичных источниках и перед использованием в юридических целях должны быть сверены с Official Journal ЕС. Наконец, рекомендованные матрица доступов (5.4) и «минимальный контур» (5.5) — синтез авторов отчёта на основе приведённых фактов, а не прямая цитата какого-либо одного источника; их следует адаптировать под отраслевую регуляторику и существующий ИТ-стек конкретной компании.

Рассмотрев этапы, команды и механизмы управления, перейдём к третьему столпу внедрения — программному стеку: от того, какие инструменты выбрала компания, во многом зависят стоимость, скорость и управляемость всех описанных выше этапов.

266GC.ai (шаблон политики) — <https://gc.ai/blog/ai-acceptable-use-policy-template> (2026-08-25)

267DecryptionDigest — <https://www.decryptiondigest.com/blog/generative-ai-data-leakage-enterprise> (2026-07-01)

268Knostic, “Primer: AI Governance Roles and Responsibilities”; Diligent, “AI governance: A guide for boards” — <https://www.knostic.ai/blog/ai-governance-roles-stakeholders> ; <https://www.diligent.com/resources/blog/ai-governance> (2025-10-15 ; 2026-06-12)

269DSALTA / DeepInspect (разбор NIST AI 600-1) — <https://www.dsalta.com/resources/ai-compliance/nist-ai-600-1-generative-ai-profile-12-risks-explained> ; <https://www.deepinspect.ai/blog/nist-genai-profile> (публикация стандарта 26.07.2024; статьи 2026-08)

270KPMG; theiso42001.com; dm11.com.br — <https://kpmg.com/ch/en/insights/artificial-intelligence/iso-iec-42001.html> ; <https://theiso42001.com/> (стандарт — декабрь 2023; EN-версия — 13.03.2026)

6. Программные инструменты для управления ИИ и ИИ-процессами

Выбор программного стека — второй по важности (после организационной модели) фактор успеха внедрения ИИ. Рынок за 2024–2026 гг. структурировался вокруг двух конкурирующих подходов: самостоятельная сборка стека на open source и покупка готовых коммерческих платформ. При этом доминирует стратегия buy: 76% корпоративных use-кейсов ИИ покупаются, а не строятся внутри (в 2024 г. — 53%)²⁷¹, а вендорские внедрения, по данным MIT NANDA, успешны примерно в 67% случаев против ~33% у in-house разработок. Тем не менее open source остаётся стратегически важным: по оценке Linux Foundation (июнь 2025), две трети организаций считают open source AI дешевле в развёртывании, почти половина называет экономию главной причиной выбора, а без open source компании тратили бы на софт в 3,5 раза больше²⁷².

6.1 Open source

Open source стек для ИИ за рассматриваемый период вырос из набора инструментов для ML-инженеров в полноценную альтернативу коммерческим платформам, покрывающую все слои: от версионирования данных до агентной оркестрации. Ниже стек разобран послойно, с акцентом на то, что реально используется в продакшне, и с типичными рисками каждого слоя.

MLOps: MLflow, Kubeflow, Airflow, DVC

MLOps-слой отвечает за управление жизненным циклом моделей: трекинг экспериментов, версионирование данных, оркестрацию пайплайнов и развёртывание. Для компаний, выходящих за пределы «вызовов API + промпты», именно этот слой превращает разрозненные эксперименты в воспроизводимый процесс.

Классический MLOps-слой в 2024–2026 гг. консолидировался вокруг четырёх инструментов. Систематический обзор 41 исследования (arXiv, 2026) выделяет MLflow, DVC и Kubeflow Pipelines как самые упоминаемые open-source решения, рядом с которыми стоит проприетарный AWS SageMaker — это отражает базовый компромисс «комьюнити-инструменты vs управляемый сервис»²⁷³.

MLflow — де-факто отраслевой стандарт трекинга экспериментов: vendor-neutral, лицензия Apache 2.0, 20K+ GitHub stars, 30M+ загрузок в месяц, production-использование

271Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

272Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

273Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

в Fortune 500. В новых версиях добавлены LLM-функции — prompt logging, evaluation, tracing, AI Gateway²⁷⁴. **Apache Airflow** — индустриальный стандарт оркестрации данных: 77K+ организаций на ноябрь 2024 (рост с 25K в 2020), 3000+ контрибьюторов (больше всех среди проектов Apache Software Foundation); 21,4% пользователей уже применяют его для MLOps и 8,2% — для GenAI-сценариев. Масштаб внедрений иллюстрируют Uber (450 тыс. запусков пайплайнов в день), Stripe (150 тыс. задач) и LinkedIn (10K+ параллельных DAG)²⁷⁵. **Kubeflow** остаётся выбором Kubernetes-зрелых средних и крупных организаций, но с оговоркой: «The learning curve is steep. You need real Kubernetes skills on the team»²⁷⁶; для снижения порога появились управляемые дистрибутивы (Canonical Charmed Kubeflow, Arrikto, AWS SageMaker Kubeflow Components). **DVC** занял нишу Git-native версионирования данных и типично используется в связке с MLflow: у MLflow нет встроенного версионирования данных, у DVC — реестра моделей²⁷⁷.

Ключевая оговорка: компании, не желающие зависеть от облачного провайдера, действительно собирают in-house пайплайны из open source (MLflow + Kubeflow + DVC), но такой вариант «demands far more integration effort and long-term maintenance» и подходит только зрелым командам²⁷⁸.

LLMOps и оркестрация LLM: LangChain/LangGraph, LlamaIndex, Langfuse + LiteLLM

С переходом от классического ML к генеративному ИИ поверх MLOps надстроился новый слой: фреймворки построения LLM-приложений, трассировка и оценка качества (observability) и шлюзы доступа к множеству провайдеров моделей.

LangChain — крупнейший LLM-фреймворк (~90–140K GitHub stars на срезах 2025–2026, ~2M PyPI-загрузок/мес в начале 2025); **LlamaIndex** (34–50K stars) сильнее в RAG и индексации данных — его production-пользователи включают Salesforce, KPMG, Carlyle и Databricks. Production-стандартом агентной оркестрации стал **LangGraph** (граф состояний, чекпоинты, human-in-the-loop; версия 1.0 GA в октябре 2025), используемый Klarna, Replit, Elastic, LinkedIn и Uber²⁷⁹.

274Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

275Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

276Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

277Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

278Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

279Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph,

В слое наблюдаемости и шлюзов сложился популярный self-hosted «OSS LLM Ops Stack» (официальный запуск 14 февраля 2025): **Langfuse** — лидер open-source LLM-observability (20K+ GitHub stars, 23,1M+ SDK-установок/мес; используется 19 из Fortune 50 и 63 из Fortune 500 — Intuit, Twilio, 7-Eleven, Merck; рост корпоративных клиентов с 200 в январе 2025 до 2200+ к марту 2026), а **LiteLLM** — AI-шлюз с единым OpenAI-совместимым API к 100+ провайдерам (роутинг, fallback, cost tracking, кэш, гардрейлы) ²⁸⁰. Важные риски: LiteLLM по тестам конкурента (TensorZero) деградирует при ~1K QPS (оценка заинтересованной стороны, требует независимой проверки), а Langfuse в январе 2026 приобретён ClickHouse — ядро остаётся под MIT, но смена владельца ставит вопрос о приоритетах roadmap ²⁸¹.

Low-code платформы: Dify, n8n — лицензионные нюансы

Low-code слой критичен для малого и среднего бизнеса: он позволяет собирать LLM-приложения (чат-боты, RAG-ассистенты, агентные workflow) без полноценной ML-команды. Однако именно здесь сосредоточены главные лицензионные риски open source стека.

Среди визуальных билдеров LLM-приложений в категории «production-ready LLM SaaS» с заметным отрывом лидирует **Dify** (~154K GitHub stars в 2026): автоматический API gateway, RBAC, SSO, аудит-логи. Обратная сторона — самый сложный self-hosted деплой (Docker Compose: API, Celery, PostgreSQL, Redis, Nginx; минимум 4–8 GB RAM). **Flowise** (~51K stars) был самым быстрым стартом для чат-ботов, но в августе 2026 репозиторий переведён в read-only архив. **n8n** (~196–203K stars) формально не является open source — Sustainable Use License ²⁸².

Здесь кроется «лицензионная ловушка», критичная для юридической проверки при внедрении: из популярных low-code AI-платформ строго open source (в смысле OSI) только Langflow (MIT); Dify — Apache 2.0 с дополнительными ограничениями (запрет мультитенантного использования и удаления брендинга), n8n — fair-code лицензия, блокирующая построение продукта для клиентов и мультитенантность ²⁸³. Для компаний, планирующих коммерческие продукты на базе этих платформ, проверка лицензий обязательна.

Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

280Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

281Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

282Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

283Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

Векторные базы данных: pgvector, Qdrant, Weaviate, Milvus

Векторная БД — обязательный компонент RAG-архитектур, через которые компании подключают модели к собственным данным. Выбор здесь за последние два года радикально упростился.

Реальная практика 2024–2026 гг.: **pgvector** поглотил **mid-scale сегмент** — к концу 2024 команды массово отказались от отдельных векторных БД ради устранения лишней инфраструктурной зависимости. Эмпирическое правило: до ~10 млн векторов правильная стартовая точка — PostgreSQL с pgvector; выше ~100 млн векторов экономически оправдан self-hosted Milvus или управляемый Zilliz²⁸⁴. Ниши специализированных БД: **Qdrant** — производительность/стоимость на масштабе (p50 — единицы мс) и on-prem; **Weaviate** — нативный гибридный поиск (vector + BM25), 1M+ Docker pulls/мес; **Milvus** (Apache 2.0, LF AI & Data, ~25K stars) — миллиардные масштабы с GPU-индексами²⁸⁵. Гибридный поиск стал стандартом production RAG: ~90% новых развёртываний используют hybrid search, поскольку чистый векторный поиск «возвращает правильную окрестность, но неправильный ответ»²⁸⁶. Контекст: векторный поиск коммодитизируется (нативную поддержку добавили SQL Server 2025, Oracle, MongoDB, AWS OpenSearch), а Gartner оценивает, что к 2026 г. более 30% предприятий будут использовать векторные БД для аугментации AI-моделей²⁸⁷.

Локальный инференс: Ollama, llama.cpp, vLLM

Инференс-слой определяет, где и как выполняются модели: на ноутбуке разработчика, в защищённом контуре компании или в высоконагруженном продакшне. За 2024–2026 гг. инструменты локального и серверного инференса полностью разделили ниши.

Сложилось чёткое разделение труда. **llama.cpp** (100K+ GitHub stars к апрелю 2026 — один из самых быстрорастущих AI-проектов в истории) и **Ollama** (110K–172K+ stars, «Docker для LLM») — разработка, edge, air-gapped и privacy-сценарии; >60% квантованных моделей на Hugging Face выпускаются в формате GGUF llama.cpp²⁸⁸. **vLLM** — production high-throughput сервинг (GitHub stars ×2,3 за 2024, до 32,6K; используется в Amazon Rufus

284Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

285Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

286Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

287Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

288Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

и LinkedIn AI; партнёрства с IBM, AWS, NVIDIA)²⁸⁹. Типичный гибрид: разработка на llama.cpp/Ollama локально, продакшн на vLLM. Главный корпоративный драйвер локального инференса — приватность: европейские компании с медицинскими и юридическими данными разворачивают Ollama за файрволом, «no GDPR-sensitive data leaves your servers, fulfilling data residency requirements by design»²⁹⁰. Для российских компаний это же свойство закрывает требования ФЗ-152 (см. российскую специфику ниже).

Открытые модели: смена лидера

Слой открытых моделей — фундамент open source стека: именно доступность весов frontier-качества делает self-hosting экономически осмысленным. За 2024–2026 гг. здесь произошла смена лидерства, имеющая прямые последствия для выбора моделей компаниями.

Важнейший сдвиг 2025 г.: **китайские open-weight модели обогнали американские**. Qwen (Alibaba) обошла Llama по кумулятивным загрузкам Hugging Face в сентябре–октябре 2025 (325,4М vs 323,7М) и к марту 2026 достигла 942,1М против 476,0М у Llama; доля Qwen в новых файнтюнах выросла с 1% (январь 2024) до 69% (февраль 2026), тогда как пик Llama (44% в августе 2024) снизился до 11%²⁹¹. При этом Llama к марту 2025 превзошла 1 млрд кумулятивных загрузок и остаётся «Linux среди AI-моделей» для компаний, желающих кастомизировать и self-host (кейс Spotify)²⁹². **DeepSeek-V3/R1** (декабрь 2024 – январь 2025, веса открыты) стал переломным моментом: frontier-уровень рассуждений при низкой себестоимости спровоцировал мировую волну self-hosting²⁹³. **Mistral** — европейский open-weight флагман для enterprise: 1031 «high-value customer» (июль 2025), 100+ корпоративных клиентов, 60% выручки из Европы; ценность для enterprise — on-prem деплой, аудируемость, глубокий файнтюнинг²⁹⁴. По вторичной агрегации (низкая надёжность методологии), ~80% американских AI-стартапов

289Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

290Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

291Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

292Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

293Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

294Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

используют китайские open-source модели при разработке продуктов ²⁹⁵.

Как читать метрики open source

Популярность проекта (stars, загрузки) — индикатор экосистемы и доступности экспертизы на рынке труда, но не гарантия пригодности: коммерческая жизнеспособность сопровождающей компании, лицензия и активность мейнтейнеров важнее «звёзд» при выборе стратегического компонента.

При сравнении инструментов по GitHub stars и загрузкам важно учитывать методологические оговорки: приведённые цифры — срезы на разные даты 2025–2026 гг., поэтому в тексте даны диапазоны; оценка производительности LiteLLM (~отказ при 1K QPS) исходит от конкурента (TensorZero) и требует независимой проверки; отчёт Astronomer по Airflow подготовлен коммерчески заинтересованным вендором, хотя и основан на опросе 5000+ практиков и объективных метриках загрузок; тезис «80% американских AI-стартапов используют китайские модели» — вторичная агрегация без раскрытой методологии ²⁹⁶. Эти оговорки не отменяют выводов, но должны учитываться при принятии решений о закупке.

Для практического ориентирования ниже — сводная таблица «что реально используют» по слоям стека ²⁹⁷:

Слой	Open source стандарт де-факто	Альтернативы / ниши	Production-пользователи
Experiment tracking	MLflow (30M+ загрузок/мес)	W&B (proprietary), ClearML	Fortune 500
Оркестрация данных	Airflow (77K+ организаций)	Prefect, Dagster, Kestra	Uber, Stripe, LinkedIn, PayPal
K8s ML-платформа	Kubeflow	Managed: Canonical, Arrikto, Vertex	K8s-шопы mid/large
Версионирование данных	DVC	LakeFS, Pachyderm	Git-native команды
LLM-фреймворк	LangChain/ LangGraph (90–140K stars)	LlamaIndex (RAG), Haystack	Klarna, Uber, LinkedIn, Salesforce, KPMG
LLM-observability	Langfuse (19 из Fortune 50)	LangSmith, Helicone, Arize	Intuit, Twilio, Merck, 7-Eleven
LLM-gateway	LiteLLM (100+	Portkey, OpenRouter,	OSS LLM Ops Stack c

²⁹⁵Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

²⁹⁶Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

²⁹⁷Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

Слой	Open source стандарт де-факто провайдеров)	Альтернативы / ниши	Production-пользователи
Low-code LLM	Dify (~154K stars)	TensorZero Flowise (архив 08/2026), n8n, Langflow	Langfuse Enterprise IT, СМБ
Векторная БД	pgvector (mid-scale), Qdrant , Milvus , Weaviate (hybrid)	Pinecone (managed), Chroma (прототипы)	—
Локальный инференс	llama.cpp (100K+ stars) + Ollama (110K+ stars)	MLX (Apple), GPT4All, Jan	privacy/GDPR-кейсы
Серверный инференс	vLLM (×2,3 stars за 2024)	TensorRT-LLM, TGI, SGLang	Amazon Rufus, LinkedIn
Открытые модели	Qwen (лидер с 09/2025), Llama , DeepSeek , Mistral	gpt-oss, Kimi, MiniMax, Gemma	Spotify, 1031 клиент Mistral
Агенты	LangGraph (продакшн), CrewAI (прототипы)	OpenAI Agents SDK, Google ADK, MS Agent Framework	Uber, LinkedIn, AppFolio

Платформы для агентов

Отдельно стоит слой агентных фреймворков, оформившийся к 2025–2026 гг. Продакшн-стандарт — **LangGraph**: граф состояний, чекпоинты, human-in-the-loop, релиз 1.0 GA в октябре 2025; используется в Uber, LinkedIn, Klarna, Replit, Elastic и AppFolio. **CrewAI** (~46–55K stars, \$18М инвестиций) — самый низкий порог входа для ролевых мультиагентных прототипов. Microsoft консолидировал AutoGen и Semantic Kernel в единый **Agent Framework** (GA апрель 2026); **Google ADK** (20K stars, 3,3М загрузок/мес) — выбор для GCP-шопов ²⁹⁸. Сложившийся практический паттерн: сквозной прототип за неделю на CrewAI или AutoGen Studio, затем переписывание на LangGraph при доказанной ценности продукта ²⁹⁹. При выходе в enterprise обратите внимание на коммерческие слои поверх open source: CrewAI Enterprise — \$60K–\$120K/год (SOC2, SSO, PII-masking, on-prem; LLM-затраты отдельно, ~2–3× от платформенного сбора), LangSmith — от \$39/seat/мес при узком бесплатном типе (5K traces/мес) ³⁰⁰. Это важно для ТКО-расчётов: «бесплатный» фреймворк в продакшне почти всегда означает платный слой наблюдаемости и безопасности.

298Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

299Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

300Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

Плюсы и минусы open source: синтез

Обобщение по всем слоям стека показывает: open source в ИИ — это не «бесплатная версия» коммерческих продуктов, а альтернативная модель владения с иным распределением затрат и рисков.

Плюсы: экономия (⅓ организаций считают open source AI дешевле; без open source траты на софт были бы в 3,5 раза выше — Linux Foundation); контроль данных и data residency (GDPR, ФЗ-152); отсутствие vendor lock-in и per-token платежей; кастомизация и файнтюнинг; аудируемость; permissive licenses — 42,3% рынка open-source AI-моделей, причём enterprise-сегмент составляет 68,9% этого рынка (2024) ³⁰¹.

Структурно плюсы open source концентрируются на входе (стоимость лицензий, контроль данных, отсутствие lock-in), а минусы — на эксплуатации (интеграция, сопровождение, отсутствие SLA, кадровые требования). Поэтому экономическое сравнение корректно вести по полной стоимости владения на горизонте 2–3 лет, а не по стартовым затратам.

Минусы: затраты на интеграцию и сопровождение in-house стека; крутые кривые обучения (Kubeflow, LangGraph); отсутствие встроенных enterprise-функций (шифрование, SLA, гарантии); расходы на GPU-инфраструктуру; лицензионные ограничения (nbn SUL, Dify brand/multi-tenant clauses, Llama Community License ≠ OSI open source); риск смены roadmap после M&A — показательны сделки Langfuse → ClickHouse, Langflow → DataStax/IBM, Portkey → Palo Alto ³⁰².

С точки зрения размера бизнеса open source наименее доступен малому бизнесу (нет компетенций для интеграции и сопровождения), оправдан точно в среднем (pgvector, Dify, Ollama для privacy-сценариев) и стратегически значим для крупного (data residency, экономика масштаба, независимость от вендоров) — что согласуется с распределением путей внедрения из раздела 6.3 ³⁰³³⁰⁴.

Российская специфика 2025 г. (по анализу вакансий, confidence medium): платформенные лидеры — Сбер (GigaChat) и Яндекс (YandexGPT); банки и ритейл в 60%+ случаев используют российские модели из-за ФЗ-152 и on-prem требований; стартапы и ИТ-вендоры — Mistral, LLaMA, Qwen, DeepSeek, Saiga. Унифицированный стек: LangChain + Qdrant + Airflow + MLflow/DVC; RAG встречается в 70%+ вакансий ³⁰⁵.

301Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

302Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

303Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

304Кросс-мерные инсайты исследования

305Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

6.2 Закрытые (коммерческие) решения

Коммерческий сегмент структурируется в четыре слоя: облачные AI-платформы гиперскейлеров, корпоративные LLM-продукты «из коробки», платформы автоматизации с агентными функциями и коммерческий LLMOps. Общая черта 2024–2026 гг. — быстрая эволюция ценообразования и упаковки: вендоры несколько раз за период пересматривали модели монетизации (Salesforce — трижды за 18 месяцев, ServiceNow полностью переписал модель в 2026, Google дважды за полтора года менял упаковку Gemini), поэтому все цены ниже актуальны на даты источников и требуют проверки перед закупкой ³⁰⁶.

Облачные платформы: Azure AI Foundry, AWS Bedrock, Google Vertex AI

Облачные платформы гиперскейлеров — основной «вход» крупного и среднего бизнеса в генеративный ИИ: они дают доступ к множеству моделей через единый API, корпоративные механизмы безопасности и биллинга, а также инфраструктуру RAG и файнтьюнинга.

Все три гиперскейлера используют потоковую модель оплаты с механизмами резервирования ёмкости для предсказуемых высоких объёмов.

AWS Bedrock: помодельная оплата on-demand плюс Provisioned Throughput (почасовая оплата за «модельные юниты» с обязательствами 1/6 месяцев; пример: Claude 2.0 ~\$35/час за юнит при 6-месячном контракте) и Reserved tier (запущен в ноябре 2025: резервирование токенов/мин на 1 или 3 месяца с фиксированной месячной ценой) ³⁰⁷. Примерные цены Standard tier (за 1 млн токенов, вход/выход): Amazon Nova Micro \$0,035/\$0,14; Claude Haiku 4.5 \$1,00/\$5,00; Claude Sonnet 4.6 \$3,00/\$15,00; Claude Opus 4.8 \$5,00/\$25,00; batch/flex — примерно вдвое дешевле ³⁰⁸.

Azure AI Foundry: потребительская модель; GPT-4o input ~\$2,50 за 1 млн токенов (середина 2025); для высоких объёмов — provisioned throughput units (PTU). Главное отличие — широчайший каталог: 1700+ моделей против ~40 у Bedrock. Полная стоимость включает Azure AI Search (RAG-индексация) и вычисления для fine-tuning ³⁰⁹. В каталоге доступны модели, продаваемые напрямую Azure (линейка GPT-5.x, o-серия), с

306Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

307Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

308Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

309Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

региональной доступностью и типами деплоймента standard/provisioned/batch ³¹⁰.

Google Vertex AI: помодельная потоковая оплата; сторонние модели (например, Claude Haiku 4.5) — по ценам Anthropic (\$1,00/\$5,00 за 1M in/out); бесплатного тарифа нет ³¹¹. В октябре 2025 Google переименовал Agentspace в **Gemini Enterprise** (permission-aware поиск, no-code Agent Designer, агенты Deep Research/NotebookLM); прежние add-on'ы Gemini Business/Enterprise (\$20/\$30 за пользователя/мес) упразднены в марте 2025 и включены в базовые планы Workspace с ростом цен на \$2–4 ³¹².

Корпоративные LLM-продукты: Copilot, ChatGPT, Claude, Glean

Подписочные LLM-продукты — самый быстрый путь внедрения для компаний любого размера: нулевая интеграция, немедленный доступ к frontier-моделям, встроенные в рабочие инструменты сценарии. Платой являются per-seat расходы, зависимость от вендора и передача данных провайдеру.

Microsoft 365 Copilot (Enterprise): \$30/польз./мес при годовой оплате поверх квалифицирующей базовой лицензии (E3/E5/Business Standard/Premium); Copilot Chat бесплатен. Для SMB (≤300 мест) Copilot Business с 1 декабря 2025 снижен с \$30 до \$21/польз./мес (промо \$18 до 30 сентября 2026) ³¹³. Важно считать полную стоимость места: с 1 июля 2026 E3 вырос с \$36 до \$39, E5 с \$57 до \$60, т.е. all-in — \$69 (E3) / \$90 (E5) за пользователя в месяц ³¹⁴. Проникновение: 30+ млн платных мест (Q4 FY26, июль 2026; рост с 15 млн в Q2 FY26), что составляет ~6,5% от ~464 млн коммерческих мест M365; по независимым опросам, еженедельное активное использование — лишь 20–30% купленных мест (confidence medium) ³¹⁵.

ChatGPT Business (бывший Team, переименован в августе 2025): \$20/польз./мес при годовой оплате (\$25 ежемесячно), минимум 2 места. **ChatGPT Enterprise:** ~\$60/польз./мес, минимум ~150 мест, только годовой контракт, публичного прайса нет (confidence medium-

310Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

311Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

312Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

313Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

314Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

315Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

high)³¹⁶. OpenAI сообщила о 5 млн платных бизнес-пользователей на релизе GPT-5 (август 2025; BNY, Figma, Intercom, Lowe's, Morgan Stanley, T-Mobile), позже — ~9 млн (2026, вторичная оценка)³¹⁷.

Claude Enterprise (Anthropic): посадочная модель — seat fee за доступ (включает Claude Code и Cowork), а все токены оплачиваются отдельно по стандартным API-тарифам; годовая оплата; расширенные security/compliance-функции поверх Team-плана³¹⁸.

Glean (корпоративный поиск и агенты): закрытое ценообразование без публичного прайса; по данным конкурентов и агрегаторов (confidence medium): ~\$50–75/польз./мес, минимум ~100 мест (~\$60K/год ACV), платный POC до \$70K, support fee ~10% ARR, агентские возможности — +~\$15/польз./мес; полная TCO средних и крупных внедрений \$350K–480K/год. Средний контракт (SpendHound): ~\$142K (SMB) и ~\$260K (Enterprise), при этом цены выросли год к году на +104% (SMB) и +47% (Enterprise)³¹⁹.

Сравнение ключевых корпоративных LLM-продуктов (цены на даты источников; часть данных buyer-reported, confidence medium)³²⁰:

Продукт	Цена	Минимум / условия	Модель оплаты	Особенности
Microsoft 365 Copilot Enterprise	\$30/польз./мес (all-in \$69–90 с базовой лицензией)	годовая оплата, базовая лицензия E3/E5	per-seat	30+ млн платных мест, активность 20–30% мест
Microsoft 365 Copilot Business (SMB)	\$21/польз./мес (промо \$18 до 30.09.2026)	≤300 мест	per-seat	снижен с \$30 01.12.2025
ChatGPT Business	\$20/польз./мес (годовая), \$25 ежемесячно	мин. 2 места	per-seat	бывший Team, переименован 08/2025
ChatGPT Enterprise	~\$60/польз./мес	мин. ~150 мест, годовой контракт	per-seat	публичного прайса нет; ~5–9 млн бизнес-

316Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

317Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

318Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

319Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

320Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

Продукт	Цена	Минимум / условия	Модель оплаты	Особенности
Claude Enterprise	seat fee + токены по API-тарифам	годовая оплата	гибридная	пользователей OpenAI
Glean	~\$50–75/польз./мес; TCO \$350K–480K/год	~100 мест, ~\$60K/год ACV; РОС до \$70K	per-seat + fees	включает Claude Code и Cowork
Salesforce Agentforce	\$2/диалог; Flex Credits \$0.10/действие; \$125–550/польз./мес	раздельные модели несовместимы в одном org	usage / per-user / flat (AELA)	рост цен YoY +104% (SMB) / +47% (Enterprise)
ServiceNow Now Assist	~\$15–25/fulfiller/мес; с 07/2026 — «assists»	квотное ценообразование	per-user → потребительская	3 модели ценообразования за 18 мес.
UiPath	Basic \$25/мес; SMB ~\$2,2K/мес; Enterprise ~\$35,9K/мес	Unified Pricing с 05/2025	потребительский кредит (Platform Unit)	агентский workflow = 25–150 assists
W&B	Pro ~\$60/мес (≤10 мест); Enterprise \$315–400/место/мес	buyer-reported (Vendr)	per-seat	объединяет AI/Robot/App/API/Agent Units
Arize AX	Free \$0; Pro от \$50/мес; Enterprise ~\$60K/год (медиана)	25K spans/мес на Free	per-usage/годовой	конкурирует с self-hosted MLflow
Datadog LLM Observability	Free ≤40K spans; Pro от \$160/мес; +\$3.50 за 10K spans	биллинг только по LLM-spans	per-usage	open-source ядро Phoenix
				новое ценообразование с 01.05.2026

Выбор корпоративного LLM-продукта: практические замечания

При сравнении подписочных продуктов важно смотреть не на номинальную цену, а на три параметра. Первый — **полная стоимость места**: Copilot требует базовой лицензии (all-in \$69–90), Claude Enterprise добавляет API-токены поверх seat fee, Glean включает support fee ~10% ARR ³²¹. Второй — **фактическое использование**: проникновение M365 Copilot

321Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock,

~6,5% базы при еженедельной активности 20–30% купленных мест означает, что массовая закупка мест «впрок» почти гарантированно создаёт неиспользуемые лицензии; рациональнее волновое развёртывание с замером активности³²². Третий — **функции доступа**: именно enterprise-тарифы дают кастомный RBAC, SCIM, аудит-логи и agent identity, а малый бизнес получает фиксированные роли — это определяет пригодность продукта для регулируемых отраслей³²³. Динамика цен Glean (+104% YoY у SMB, +47% у Enterprise) показывает и риск продления: вендоры в условиях дефицита альтернатив резко поднимают цены на второй год, поэтому мультiletные соглашения с фиксацией цены (модель AELA у Salesforce) становятся инструментом управления TCO³²⁴.

Платформы автоматизации: Salesforce Agentforce, ServiceNow, UiPath

Третий слой коммерческого рынка — платформы, монетизирующие не доступ к модели, а выполненную работу: роботизация процессов, ITSM, CRM-агенты. Именно здесь в 2025–2026 гг. происходили самые агрессивные эксперименты с ценообразованием, и именно здесь цена ошибки в выборе модели лицензирования максимальна.

Salesforce Agentforce за 18 месяцев сменил три параллельные модели ценообразования: \$2 за диалог (GA октябрь 2024), Flex Credits (\$500 за 100К кредитов; \$0,10 за стандартное действие = 20 кредитов; май 2025), per-user add-on \$125/польз./мес (Sales/Service Cloud), Industries add-on \$150, Agentforce 1 Edition от \$550/польз./мес; бесплатный Foundations. Flex Credits и Conversations несовместимы в одном org; break-even — 20 действий на диалог³²⁵. Исходная модель \$2/диалог критиковалась за непредсказуемость и отсекала SMB и некоммерческие организации, что привело к Flex Credits, а затем к AELA (Agentic Enterprise License Agreement, Dreamforce октябрь 2025, flat fee, мультiletный)³²⁶. Показательно: из 5000 сделок Agentforce за первые два квартала оплаченными были лишь 3000³²⁷.

Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

322Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

323Кросс-мерные инсайты исследования

324Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

325Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

326Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

327Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

ServiceNow Now Assist: квотное ценообразование; standalone-лицензия ~\$15–25 за fulfiller/мес поверх базовых модулей; с июля 2026 — переход на потребительскую модель «assists»: суммаризация инцидента = 1 assist, агентский workflow = 25–150 assists, с пулом на уровне аккаунта ³²⁸.

UiPath: с мая 2025 новые клиенты покупают по Unified Pricing — единый потребительский кредит «Platform Unit» (объединяет AI Units, Robot Units, App Units, API Calls, Agent Units). Basic — \$25/мес (2 робота), средний SMB-контракт ~\$2,2K/мес, средний enterprise-контракт ~\$35,9K/мес ³²⁹.

Коммерческий LLMOps: W&B, Arize, Datadog

Коммерческий LLMOps конкурирует с open source связкой MLflow + Langfuse, продавая управляемость, SLA и интеграцию с существующим мониторингом. Ценовой разрыв существенен и должен осознанно закладываться в TCO: например, Enterprise-тариф W&B (\$315–400/место/мес) на команду из 10 человек даёт \$38–48K/год против близкой к нулю лицензионной стоимости self-hosted MLflow — но с затратами на развёртывание и сопровождение последнего ³³⁰³³¹.

Weights & Biases: Pro ~\$60/мес (до 10 мест), Enterprise — \$315–400 за место/мес по buyer-reported контрактам Vendr (confidence medium) ³³². **Arize:** open-source Phoenix бесплатен (OTel/OpenInference); коммерческая Arize AX — Free (\$0, 25K spans/мес), Pro от \$50/мес, Enterprise — медиана ~\$60K/год (диапазон ~\$60K–98K); компания подняла \$70M Series C в феврале 2025 (среди инвесторов — Datadog и PagerDuty) ³³³. **Datadog LLM/Agent Observability:** Free до 40K LLM spans/мес; Pro от \$160/мес (100K spans включено), дополнительные spans \$3,50 за 10K; биллинг только по LLM-spans; новое ценообразование с 1 мая 2026 ³³⁴. Известная жалоба при сравнении TCO — «тихий» ingestion-premium:

328Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

329Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

330Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

331Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

332Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

333Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

334Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

автоматическая активация биллинга при наличии GenAI-атрибутов в OTel-spans без явной фильтрации³³⁵.

Модели лицензирования: обобщение

Для закупочных команд практический вывод двойной. Во-первых, условия коммерческих AI-контрактов в 2025–2026 гг. нестабильны и пересматриваются вендорами по несколько раз в год — это создаёт окно для переговоров (скидки на мультiletных соглашениях, фиксация цен, пилотные кредиты), но требует ежегодного пересмотра контрактной стратегии. Во-вторых, потребительские модели (за действие, за токен, за assist) переносят риск неэффективности на покупателя: без LLM-observability и бюджетных лимитов на уровне шлюза счёт за агентные workflow может вырасти непредсказуемо — как показывает пример «тихого» ingestion-premium у Datadog³³⁶.

По рынку прослеживаются четыре модели: (1) **per-seat подписка** (M365 Copilot \$30, ChatGPT Business \$20) — предсказуема, но страдает от неиспользуемых мест; (2) **потокенная/usage-based** (Bedrock, Vertex, токены Claude Enterprise поверх seat fee) — масштабируется с ценностью, но требует cost-контроля; (3) **потребительские кредиты/действия** (UiPath Platform Unit, Agentforce Flex Credits, ServiceNow assists) — отраслевой тренд 2025–2026 на оплату за «действие агента»; (4) **гибридные/flat-fee enterprise-соглашения** (Agentforce AELA) — ответ на критику непредсказуемости. Эксперименты Salesforce с тремя моделями за 18 месяцев показывают: рынок ещё не сошёлся на едином стандарте, и условия ценообразования — легитимный предмет переговоров.

6.3 Open source vs закрытые решения: как выбирают компании разного размера

Противопоставление «open source vs закрытые» на практике не бинарно: одна и та же компания обычно комбинирует оба пути по слоям стека. Тем не менее центр тяжести выбора системно смещается в зависимости от размера и зрелости бизнеса, что подтверждается количественно.

Ключевая рыночная закономерность (Menlo Ventures, State of Generative AI in the Enterprise 2025): корпоративные расходы на GenAI выросли с \$1,7B (2023) до \$37B (2025; +3,2× за 2024), при этом **76% use-кейсов покупаются, а не строятся** (53% в 2024), а конверсия в production составляет 47% против 25% у традиционного SaaS³³⁷. В приложениях стартапы

335Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

336Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

337Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

захватывают 63% рынка, но на инфраструктурном уровне инкумбенты (Databricks, Snowflake, MongoDB, Datadog) удерживают 56% — предприятия строят ИИ на уже доверенных платформах ³³⁸.

Одновременно в enterprise растёт предпочтение закрытых моделей: по опросу CIO a16z (2026), **более трети предприятий предпочитают closed source** (скорость улучшения качества, дефицит ИИ-тантов, data security); **ТСО открытых и закрытых моделей сближается**; enterprise уходят от файнтьюнинга к prompt engineering и роутингу между моделями, тогда как стартапы активно файнтьюнят open source ³³⁹. Это разрешает кажущийся конфликт «LF: open дешевле» vs «a16z: enterprise выбирает closed»: выбор зависит от зрелости компании и требований к данным, а экономический зазор между путями сужается.

Внедрение по размеру компании (Q1 2026, вторичная агрегация, confidence medium): 83% компаний 5000+ сотрудников имеют развёрнутый ИИ; mid-market — 64%; SMB (50–499) — 42%; малый бизнес (<50) — 18%. Средний SMB тратит ~\$18K/год на ИИ-инструменты; 61% SMB называют цену главным барьером; 74% SMB используют ИИ косвенно через встроенные функции SaaS ³⁴⁰.

Критерий	Open source стек	Закрытые платформы
Типовой покупатель	Зрелые enterprise, стартапы, Китай; компании с требованиями data residency	SMB (через встроенный SaaS-ИИ), mid-market, enterprise без глубокой ИИ-экспертизы
Стартовые затраты	Низкие лицензионные, но высокие интеграционные («far more integration effort and long-term maintenance» ³⁴¹)	От \$20–30/польз./мес (Copilot, ChatGPT Business); enterprise-контракты \$60K–480K/год (Arize, Glean) ³⁴²
ТСО	⅓ организаций считают open дешевле; без open source траты ×3,5 (Linux Foundation) ³⁴³	ТСО сближается с open source (a16z) ³⁴⁴ ; переплата за скорость и SLA

338Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

339Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

340Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

341Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

342Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

Критерий	Open source стек	Закрытые платформы
Данные и compliance	Полный контроль, data residency by design (GDPR, ФЗ-152) ³⁴⁵	Зависит от вендора; enterprise-тарифы дают аудит, но данные уходят провайдеру
Персонал	Нужны реальные K8s/ML-компетенции (Kubeflow, LangGraph) ³⁴⁶	Достаточно бизнес-пользователей и prompt engineering ³⁴⁷
SLA и поддержка	Отсутствуют; риск смены roadmap после M&A (Langfuse → ClickHouse) ³⁴⁸	Контрактные SLA, поддержка, enterprise security-функции
Гибкость	Файнтюнинг, кастомизация, отсутствие lock-in и per-token платежей ³⁴⁹	Ограничена API и roadmap вендора; lock-in
Скорость внедрения	Недели–месяцы на сборку	Дни; конверсия в production 47% ³⁵⁰

Стоит подчеркнуть, что цифры adoption по размеру бизнеса сильно зависят от определения «использования»: опросы (например, US Chamber — 58% малых компаний в США в 2025) дают кратно более высокие значения, чем государственная статистика (Eurostat — 17% малых против 55% крупных в ЕС); оба ряда корректны, но измеряют разное — любое эпизодическое использование против регулярного внедрения ³⁵¹. Поэтому при бенчмаркинге своей компании нужно сравниваться с данными той же методологии.

343Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

344Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

345Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

346Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

347Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

348Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

349Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

350Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

351Кросс-верификация результатов исследования

Практическая эвристика по размеру бизнеса: малый бизнес почти всегда выигрывает от встроенного ИИ в уже оплаченный SaaS (74% SMB идут этим путём) и подписок уровня ChatGPT Business / Copilot Business (\$18–21); средний бизнес комбинирует SaaS-подписки с точечным open source (Dify для внутренних ассистентов, pgvector в существующем PostgreSQL); крупный бизнес с регуляторными ограничениями и масштабом строит гибрид: закрытые frontier-модели через облачные платформы для сложных задач + self-hosted открытые модели (Qwen, DeepSeek, Mistral) на vLLM для массовых и чувствительных сценариев, с обязательным LLM-шлюзом (LiteLLM) и наблюдаемостью (Langfuse или коммерческий аналог) ³⁵²³⁵³.

Итоговая закономерность раздела: выбор «open vs closed» в 2026 г. — это не идеологическое, а инженерно-экономическое решение, принимаемое послойно. Компании, делающие его осознанно (по критериям данных, компетенций и ТСО на горизонте 2–3 лет), получают преимущество обеих моделей; компании, выбирающие «всё у одного вендора» или «всё self-hosted» по умолчанию, переплачивают либо деньгами, либо рисками ³⁵⁴³⁵⁵³⁵⁶.

Отдельно отметим матрицу выбора пути (инсайт 3): путь «встроенный ИИ в SaaS» оптимален при ограниченных ИИ-компетенциях, допустимости передачи данных провайдеру и стандартных сценариях; путь «самостоятельный стек на open source» — при наличии ML/K8s-экспертизы, жёстких требованиях data residency и масштабе, где per-token платежи становятся значимыми; гибрид — норма для крупных компаний: закрытые frontier-модели для сложных рассуждений, открытые self-hosted — для массовых и чувствительных нагрузок ³⁵⁷³⁵⁸.

7. Китайский опыт внедрения ИИ

Китай за 2024–2026 гг. продемонстрировал альтернативную модель массового внедрения ИИ — «государственное форсирование + ценовая война + открытые модели» — и по глубине развёртывания обогнал глобальное среднее. Глава построена по пяти блокам:

352Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

353Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

354Кросс-мерные инсайты исследования

355Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

356Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

357Кросс-мерные инсайты исследования

358Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

макростатистика, госполитика и регулирование, ценовая динамика и открытые модели, корпоративные платформы, рынок; завершает её разбор переносимых и специфичных уроков.

7.1 Макростатистика

Китай — единственная крупная экономика, где adoption ИИ измеряется государственной статистикой (CNNIC, MIIT) и форсируется государственной политикой, что делает кейс методологически уникальным и требует осторожности при прямых сравнениях с рыночными экономиками.

По данным CNNIC (57-й отчёт, февраль 2026), к декабрю 2025 г. в Китае насчитывалось **602 млн пользователей генеративного ИИ** (+141,7% к концу 2024 г.), проникновение genAI достигло 42,8% (+25,2 п.п. за год)³⁵⁹. За первое полугодие 2025 г. база выросла с ~257 млн до 515 млн (+106,6% за полгода); более 90% пользователей предпочитают отечественные модели³⁶⁰. В стране более 6000 ИИ-предприятий; масштаб core-ИИ-индустрии в 2025 г. ожидался >1,2 трлн юаней (~\$171,4 млрд); AI-приложения покрывают сталь, цветные металлы, энергетику и телеком (R&D, контроль качества, клиентский сервис)³⁶¹.

По данным McKinsey Greater China (2025), **83% китайских компаний регулярно используют genAI хотя бы в одной функции** — выше глобального среднего, а **45% достигли масштабного или полного развёртывания против 38% в мире**³⁶². МСБ-сегмент вовлечён массово: за Q1 2025 в Китае зарегистрировано 254 000 новых МСП, специализирующихся на ИИ-софте (всего в стране >60 млн МСП — 99,9% всех предприятий). Пример эффекта: ИИ снизил брак на производственной линии компании новой энергетики с 3‰ до 0,5‰³⁶³. Это принципиально отличается от западной картины, где малый бизнес системно отстаёт (18% при <50 сотрудниках): в Китае барьер для МСБ снят готовыми AI-функциями в DingTalk/WeCom и почти бесплатными API, тогда как на Западе малый бизнес упирается в цену (61% SMB называют её главным барьером) и дефицит компетенций³⁶⁴³⁶⁵.

359Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

360Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

361Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

362Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

363Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

364Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

365Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

7.2 Государственная политика AI+ и регулирование САС

Китайская государственная рамка уникальна тем, что одновременно форсирует спрос (директивное внедрение) и лицензирует предложение (обязательная регистрация сервисов).

Инициатива «AI+» впервые вписана в Отчёт о работе правительства в 2024 г.; в 2025 г. Госсовет КНР издал «Мнение об углубленном осуществлении действия “AI+”» (2025-11) с шестью ключевыми направлениями и восемью опорными системами; 15-й пятилетний план («15-й») предусматривает полномасштабное внедрение AI+ ³⁶⁶. SASAC (Комитет по контролю над госимуществом) провела совещание по углублению кампании AI+ для центральных госпредприятий (ЦГП) — фактически сигнал обязательного внедрения ИИ в госсекторе (confidence medium) ³⁶⁷.

Параллельно действует жёсткий регуляторный контур: по «Временным мерам управления генеративными ИИ-сервисами» (САС, 2023) каждый genAI-сервис обязан пройти фейлинг (ЦГП). К концу 2025 г. накоплено 748 genAI-сервисов прошли федеральный фейлинг и 435 приложений — регистрацию; к 30 апреля 2026 г. — уже **868 и 530** соответственно; только за 2025 г. добавлено 446 сервисов ³⁶⁸. Модель сочетает форсирование спроса (AI+ в госсекторе) с лицензированием предложения (фейлинг САС) — уникальная конфигурация «ускорение с контролем».

7.3 Ценовая война API и экосистема открытых моделей

Если госполитика создала спрос, то предложение сделала массовым ценовая война и открытость весов. Эти два фактора взаимоусиливались: открытые модели давали бесплатную альтернативу API, заставляя облачных вендоров демпинговать, а дешёвые API снижали стимул к self-hosting у малых компаний.

По данным CAICT (Китайская академия ИКТ), **средняя цена API китайских больших моделей упала более чем на 90% с 2023 г.**; ведущие вендоры снижали цены более 10 раз; DeepSeek V4 Pro в июне 2026 снизил цену до ¼ от исходной ³⁶⁹. Точка воспламенения — май 2024, когда ByteDance снизил цены Doubao с «1000» до «100» (падение на 99,3%), что «взорвало» спрос; после выхода DeepSeek-R1 Volcano Engine первым выпустил DeepSeek-R1 API ³⁷⁰.

Экосистема открытых моделей достигла глобального лидерства: китайские open-weight

366Свод источников по китайскому рынку ИИ (CNNIC, MIIT, САС, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

367Свод источников по китайскому рынку ИИ (CNNIC, MIIT, САС, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

368Свод источников по китайскому рынку ИИ (CNNIC, MIIT, САС, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

369Свод источников по китайскому рынку ИИ (CNNIC, MIIT, САС, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

370Свод источников по китайскому рынку ИИ (CNNIC, MIIT, САС, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

модели обогнули американские по кумулятивным загрузкам Hugging Face в августе 2025 г. (разрыв вырос до >400 млн); **Qwen — самая широко используемая открытая семья моделей в мире**: >1 млрд кумулятивных загрузок HF к 21 января 2026 (Alibaba), 942,1 млн к 27 марта 2026 — первое место с большим отрывом³⁷¹. После релиза DeepSeek-R1 (январь 2025) китайские компании и госорганы интегрировали модель массово: Шэньчжэнь развернул DeepSeek-приложения в облаке «для всех госорганов города», Чанша использует DeepSeek в smart-city; открытость и дешевизна модели «dramatically lowers costs for deployment»³⁷².

7.4 Корпоративные платформы: дистрибуция через super-app'ы

Ключевая особенность китайской модели — ИИ дистрибутируется бизнесу через уже доминирующие корпоративные платформы, снимая барьер компетенций: компании не собирают стек, а активируют AI-функции внутри привычной среды совместной работы. Это китайский эквивалент стратегий Microsoft (Copilot внутри M365) и Google (Gemini внутри Workspace), но с существенно большим проникновением базовых платформ.

Alibaba / DingTalk. Alibaba сочетает все три роли: разработчик лидирующей открытой модели (Qwen), облачный провайдер (№1 AI Cloud) и владелец главной корпоративной платформы (DingTalk). AI-продуктовая выручка Alibaba Cloud растёт трёхзначными темпами 11 кварталов подряд; AI-продукты = 30% внешней выручки облака; годовой run-rate AI-выручки >35,8 млрд юаней (~\$5,2 млрд); клиентская база Model Studio выросла в 8 раз за год³⁷³. DingTalk: 700 млн пользователей и **25 млн организаций** (конец 2023), 120 тыс. платящих предприятий, 700 тыс. предприятий реально используют DingTalk AI (январь 2024); с апреля 2024 действует AI-ассистент-маркет с целью **10 млн AI-ассистентов за 3 года**; к марту 2025 — >2,2 млн AI-предприятий-пользователей³⁷⁴. DingTalk интегрирован с Tongyi Qianwen с апреля 2023 (к августу 2023 — 17 продуктовых линий и 55 сценариев); CEO Alibaba Эдди Ву называет DingTalk одним из двух главных AI-входов в B2B³⁷⁵. Alibaba объявила, что инвестиции в ИИ за следующие 3 года превысят суммарные инвестиции прошлых 10 лет (февраль 2025)³⁷⁶.

Tencent / WeCom. Стратегия Tencent — интеграция ИИ в существующие денежные потоки (реклама, игры) и корпоративную дистрибуцию через WeCom. WeCom (微信): >10 млн

371Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

372Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

373Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

374Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

375Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

376Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

предприятий и организаций, ~180 млн активных пользователей, **80% топ-500 китайских компаний** (конец 2021). Модель Hunyuan интегрирована в WeCom в апреле 2024; в феврале 2025 добавлены Smart Robot и Smart Spreadsheet с серым доступом к DeepSeek-R1³⁷⁷. В Q3 2025 выручка Tencent выросла на 15% г/г (¥192,9 млрд) во многом за счёт AI-таргетинга рекламы и кодинговых инструментов³⁷⁸. С июня 2026 Tencent внутренне тестирует enterprise-агента «Dayuan» в WeCom, использующего данные чатов, документов, встреч, почты и календаря пользователя³⁷⁹.

Baidu / ERNIE / Qianfan. Baidu — пионер китайского genAI (ERNIE Bot, 2023) с сильной позицией в enterprise-облаке и агентной платформе. ERNIE Bot: >200 млн MAU (январь 2026), >450 млн зарегистрированных пользователей, >1,5 млрд запросов API в день; >150 000 enterprise-клиентов через Baidu AI Cloud³⁸⁰. Выручка AI Cloud +34% г/г; годовая выручка AI-приложений >10 млрд юаней; **>1,3 млн агентов на платформе Qianfan**; PaddlePaddle — 10,7 млн разработчиков, 235 000 предприятий и институтов, ~860 000 моделей (часть данных — агрегаторы, confidence medium)³⁸¹.

ByteDance / Doubao / Volcano Engine / Feishu. ByteDance иллюстрирует самую агрессивную стратегию: демпинг токенов, массовое потребительское приложение и enterprise-облако усиливают друг друга. Дневное потребление токенов Doubao превысило 120 трлн (март 2026) и 180 трлн (июнь 2026) — рост ~1000–1500× с запуска в мае 2024; число enterprise-клиентов с кумулятивным использованием >1 трлн токенов выросло со 100 (конец 2025) до 200³⁸². По IDC, **Volcano Engine — №1 рынка MaaS публичного облака Китая: 46,4% (2024) → 49,5% (2025)**; далее Alibaba Cloud (27–28%) и Baidu Cloud. Весь рынок MaaS вырос с 114 трлн токенов (2024) до 1944 трлн (2025, ×16 за год); прогноз IDC на 2026 — 40 000 трлн (ещё ×20)³⁸³. Feishu/Lark: клиенты — Xiaomi, Li Auto, Nio, Xpeng; **>90% новых клиентов Feishu с начала 2026 г. также покупают AI-возможности**; в июле 2026 Feishu объединён с Doubao, GTM — с Volcano Engine. Doubao — крупнейшее AI-приложение Китая: 345–382 млн MAU (QuestMobile)³⁸⁴. Volcano Engine обслуживает автопром (Changan, GAC Aion, SAIC, Mercedes-Benz, BMW Brilliance, Tesla), финансы (□□□□, □□□□ и др.), >50% вузов «проекта 985», ведущие смартфон-бренды³⁸⁵.

377Свод источников по китайскому рынку ИИ (CNNIC, МИИТ, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

378Свод источников по китайскому рынку ИИ (CNNIC, МИИТ, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

379Свод источников по китайскому рынку ИИ (CNNIC, МИИТ, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

380Свод источников по китайскому рынку ИИ (CNNIC, МИИТ, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

381Свод источников по китайскому рынку ИИ (CNNIC, МИИТ, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

382Свод источников по китайскому рынку ИИ (CNNIC, МИИТ, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

383Свод источников по китайскому рынку ИИ (CNNIC, МИИТ, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

384Свод источников по китайскому рынку ИИ (CNNIC, МИИТ, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

385Свод источников по китайскому рынку ИИ (CNNIC, МИИТ, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

Huawei / Pangu. В отличие от интернет-гигантов, Huawei строит стратегию вокруг промышленных вертикалей и собственного кремния, что делает её предложение устойчивым к экспортным ограничениям. Huawei Pangu занял нишу «industrial deployment orientation»: промышленные модели для дефектоскопии, оптимизации энергопотребления, предиктивного обслуживания; медицинские, финансовые и госверсии; связка Ascend + Pangu снижает порог внедрения для традиционных отраслей (умные шахты, заводы, финриск, городское управление) ³⁸⁶. Pangu Ultra MoE (718 млрд параметров, май 2025) — первая frontier-модель, обученная целиком на отечественном кремнии (Ascend 910B/910C, MindSpore); go-to-market — правительство, ГОП, банки, крупные производители через Huawei Cloud ModelArts ³⁸⁷.

7.5 Механика ускорения: почему модель сработала

Китайский кейс важен не только цифрами, но и механикой. Три усилителя adoption сработали одновременно. Первый — **спросовый**: директивная кампания AI+ сделала внедрение ИИ обязательным KPI для центральных госпредприятий и госорганов (Шэньчжэнь развернул DeepSeek-приложения «для всех госорганов города»), создав гарантированный рынок первого спроса ³⁸⁸. Второй — **ценовой**: война цен API (>90% с 2023 г., -99,3% у Doubao в мае 2024) превратила токены в почти бесплатную коммодити; рынок MaaS ответил ростом потребления ×16 за год, а число «трлн-токенных» enterprise-клиентов Volcano Engine удвоилось за полгода ³⁸⁹. Третий — **компетенционный**: платформы DingTalk, WeCom, Qianfan и Model Studio упаковали ИИ в готовые сценарии, не требующие собственной ML-команды, — клиентская база Model Studio выросла в 8 раз за год, а >90% новых клиентов Feishu сразу покупают AI-возможности ³⁹⁰. По сути, Китай государственно оркестровал тот же «встроенный в SaaS» путь, который на Западе формируется рынком, — и достиг 45% масштабных внедрений против 38% глобально ³⁹¹.

7.6 Рынок и конкуренция

Итоговая рыночная структура китайского корпоративного ИИ — олигополия облачных платформ с быстрой сменой позиций: лидерство в AI Cloud у Alibaba, в MaaS — у Volcano Engine, в корпоративной дистрибуции — у DingTalk и WeCom, в промышленном сегменте

386Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

387Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

388Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

389Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

390Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

391Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

392Кросс-мерные инсайты исследования

— у Huawei.

Распределение сил иллюстрирует, что китайский рынок не монополизирован одним игроком, а конкурентен сразу по нескольким осям (облако, токены, платформы, промышленные решения), что поддерживает ценовое давление и скорость инноваций. По Omdia, рынок AI Cloud Китая 2025 — ~56,7 млрд юаней; лидеры H1 2025: Alibaba Cloud (35,8%), Volcano Engine (14,8%), Huawei Cloud (13,1%), Tencent Cloud (7%), Baidu Cloud (6,1%); Tencent Cloud впервые достигла прибыльности в 2025 ³⁹³. 2026 год назван «годом агентов»: Tencent WorkBuddy (встроен в WeCom, мультимодельный — Hunyuan/DeepSeek/GLM/Kimi/MiniMax) — крупнейший AI-офис Китая (20,97 млн PC-визитов в июне 2026, в 2,6 раза больше ByteDance Trae); ByteDance запустил Doubao Work; Alibaba консолидирует офисные агенты на DingTalk ³⁹⁴.

Дополнительный контекст глобального сравнения: по McKinsey State of AI 2026, 88% организаций мира регулярно используют AI, 44% масштабируют (годом ранее — 38%); у крупных компаний (>\$1 млрд выручки) масштабирование агентов выросло с 27% до 40% ³⁹⁵. Китай (45% масштабных внедрений по McKinsey China) опережает это среднее не за счёт уникальных технологий — модели и платформы глобально сопоставимы, — а за счёт описанной выше механики ускорения ³⁹⁶.

7.7 Уроки: что переносимо, а что специфично для Китая

Разбор уроков важен для нерелевантных Китаю экономик по двум причинам: во-первых, переносимые механизмы (цена, дистрибуция, открытые модели, вертикализация) уже работают и на Западе; во-вторых, понимание непереносимых условий защищает от некритичного копирования китайских практик в иной институциональной среде.

Китайский опыт не копируется буквально, но декомпозируется на переносимые механизмы и специфичные условия.

Переносимо:

1. **Цена как главный разблокировщик спроса.** Падение цен API >90% и релиз дешёвых открытых моделей (DeepSeek) увеличили рынок MaaS в 16 раз за год — глобально применимый механизм: снижение стоимости токена напрямую конвертируется в adoption ³⁹⁷.
2. **Дистрибуция через существующие рабочие платформы.** Модель «ИИ внутри DingTalk/WeCom» — китайский аналог пути «встроенный ИИ в SaaS», который на

393Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

394Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

395Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

396Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

397Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

Западе реализуют Microsoft и Google; 74% SMB используют ИИ именно косвенно через SaaS³⁹⁸.

3. **Открытые модели как инфраструктура self-hosting.** Доминирование Qwen и волна self-hosting DeepSeek показывают жизнеспособность open source пути при достаточной дешевизне моделей³⁹⁹⁴⁰⁰.
4. **Вертикализация для традиционных отраслей** (Pangu для ГОП, промышленности, финансов) — паттерн, воспроизводимый в любой стране с крупным госсектором⁴⁰¹.

Специфично для Китая (непереносимо напрямую):

1. **Государственное форсирование** — директивная кампания AI+ для госпредприятий и обязательный фейлинг САС (868 сервисов) не имеют аналогов в рыночных экономиках⁴⁰².
2. **Замкнутая национальная экосистема:** >90% пользователей предпочитают отечественные модели — результат и регулирования, и недоступности западных сервисов⁴⁰³.
3. **Масштаб super-app дистрибуции** (700 млн пользователей DingTalk, 25 млн организаций) вне Китая невоспроизводим.
4. **Ценовая война такой глубины** (–99,3% у Doubao) возможна только при субсидировании платформенными гигантами и избыточных мощностях⁴⁰⁴.

Общий вывод по главе: китайский опыт подтверждает универсальный тезис отчёта — adoption ускоряется снятием барьеров стоимости и компетенций, а не качеством моделей как таковым; различие в том, что в Китае эти барьеры сняло государство и платформенная конкуренция, а на Западе — рыночная дистрибуция SaaS⁴⁰⁵⁴⁰⁶.

8. Выводы и практические рекомендации

Настоящая глава синтезирует результаты предыдущих разделов отчёта и

398Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

399Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

400Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

401Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

402Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

403Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

404Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

405Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

406Кросс-мерные инсайты исследования

исследовательских измерений (dim01–dim08) в систему выводов и действий. Базовая рамка: технология ИИ в 2024–2026 гг. стала доступной и дешёвой — frontier-модели продаются за токены, открытые модели (Qwen, DeepSeek) бесплатны для self-hosting, а цены на API в Китае упали более чем на 90% — однако большинство компаний по-прежнему не извлекает измеримой ценности. Разница между лидерами и большинством определяется организацией, управлением и дисциплиной внедрения, а не доступом к моделям.

8.1 Синтез ключевых инсайтов

Шесть кросс-мерных инсайтов ниже получены пересечением независимых исследовательских измерений (модели зрелости, оргмодели, governance, open source, коммерческие платформы, SMB, кейсы, Китай) и потому устойчивее любого одиночного источника ⁴⁰⁷⁴⁰⁸.

Шесть кросс-мерных инсайтов исследования формируют каркас всех практических выводов ⁴⁰⁹:

Инсайт 1. «Бутылочное горлышко» — не технология, а переход пилот → масштаб, и он организационный (confidence: high). Технологии доступны и дешевеют (ценовая война API в Китае, сближение TCO open/closed), но 60–95% компаний не извлекают ценности: Gartner оценивает долю заброшенных после PoC GenAI-проектов в ≥30% (фактически ~50%), MIT NANDA — 95% без измеримого P&L-эффекта, BCG — 74% без существенной ценности (расхождения объясняются разными определениями). Сильнейшие корреляты успеха — спонсорство CEO, редизайн процессов, оргмодель; правило BCG 10-20-70 (70% усилий — люди и процессы). Следствие: программу внедрения нужно строить вокруг этапов и команд, а не вокруг выбора модели; каскад провалов почти всегда одинаков — пилот без метрики, отсутствие спонсора, процесс не перестроен, масштабирование не наступает ⁴¹⁰⁴¹¹.

Каждый инсайт ниже снабжён следствием для практики; уровни уверенности (confidence) проставлены по результатам кросс-верификации ⁴¹².

Инсайт 2. Размер бизнеса определяет не «использовать ли ИИ», а архитектуру доступа и governance (confidence: high). Один и тот же инструмент даёт малому бизнесу 3 фиксированные роли, а enterprise — кастомный RBAC, SCIM, аудит и agent identity; SMB идут через встроенный SaaS-ИИ (74%), enterprise — через платформы с разделением доступа. Рекомендации по доступам должны быть сегментированы по размеру компании ⁴¹³.

407Кросс-мерные инсайты исследования

408Кросс-верификация результатов исследования

409Кросс-мерные инсайты исследования

410Кросс-мерные инсайты исследования

411Кросс-верификация результатов исследования

412Кросс-верификация результатов исследования

413Кросс-мерные инсайты исследования

Инсайт 3. Два конкурирующих пути внедрения: «встроенный ИИ в SaaS» vs «самостоятельный стек на open source» (confidence: high). 76% use-кейсов покупаются, вендорские внедрения успешны в ~67% случаев против ~33% in-house; при этом Китай демонстрирует массовый успех self-hosted open source благодаря дешевизне DeepSeek/Qwen. Развилка определяется компетенциями команды и требованиями к данным, а не размером бюджета как таковым ⁴¹⁴⁴¹⁵⁴¹⁶.

Инсайт 3 дополнение. Развилка путей не статична: компании мигрируют между ними по мере роста зрелости — типичная траектория «встроенный SaaS-ИИ → подписки на LLM-продукты → low-code на open source → гибридный self-hosted стек», причём каждый переход должен оправдываться требованиями к данным или экономикой масштаба, а не модой ⁴¹⁷⁴¹⁸.

Инсайт 4. Китайская модель «государственное форсирование + ценовая война + открытые модели» ускоряет adoption быстрее рыночной модели Запада: 45% масштабных внедрений против 38% глобально (confidence: high). Барьер стоимости и компетенций снят ценовой войной (>90%) и готовыми платформами (DingTalk, WeCom) — это тот же SaaS-путь, но государственно оркестрованный ⁴¹⁹⁴²⁰.

Инсайт 5. Shadow AI делает формальные политики доступа бессмысленными без технического принуждения (confidence: high). 59–90% сотрудников используют несанкционированный ИИ, 77% вставляют корпоративные данные в промпты; разрыв между политикой и поведением закрывается не запретами (кейс Samsung: запрет → разворот политики), а предоставлением легального управляемого доступа с политиками, встроенными в инструменты. Схема доступов должна проектироваться как «управляемое разрешение», а не «запрет» ⁴²¹.

Инсайт 5 дополнение. Практическая схема «управляемого разрешения» включает: выдачу корпоративных лицензий вместо личных аккаунтов, LLM-шлюз с логированием и маскированием ПИ, встроенные в инструменты политики (DLP на уровне промптов) и обучение — программы AI literacy закрывают разрыв между политикой и поведением лучше запретов ⁴²².

414Кросс-мерные инсайты исследования

415Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

416Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

417Кросс-мерные инсайты исследования

418Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

419Кросс-мерные инсайты исследования

420Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

421Кросс-мерные инсайты исследования

422Кросс-мерные инсайты исследования

Инсайт 5 имеет прямое следствие для бюджетирования: поскольку shadow AI составляет до ~40% расходов на приложения, легализация доступа часто не увеличивает, а перераспределяет существующие траты в управляемый контур ⁴²³⁴²⁴.

Инсайт 6. Агентный ИИ — фронтир 2026, но и главный риск-зон: >40% агентных проектов будут отменены к 2027 (Gartner), при этом agent identity становится новой категорией governance (confidence: medium-high). Все измерения конвергируют к агентам (LangGraph как продакшн-стандарт, DingTalk с целью 10 млн агентов, >1,3 млн агентов на Qianfan), но кейс Walmart — консолидация в 4 супер-агента после agent sprawl — показывает типовую ошибку: расползание агентов без оркестрации. Рекомендация: агентная архитектура с централизованным реестром и идентичностями агентов ⁴²⁵⁴²⁶⁴²⁷.

Сводка уровней уверенности: инсайты 1–5 подтверждены несколькими независимыми измерениями (high confidence), инсайт 6 — medium-high; расхождения в оценках провалов (95% MIT / 30–50% Gartner / 74% BCG) и adoption SMB (58% опросы / 17% Eurostat) разрешаются различиями определений и приводятся с оговорками ⁴²⁸.

Методологические оговорки к рекомендациям

Рекомендации главы опираются на источники с разным уровнем надёжности. Высокая уверенность (high confidence): разрыв «adoption vs impact», доли провалов пилотов, правило 10-20-70, доминирование hub-and-spoke, преимущество buy, массовость shadow AI, лидерство китайских открытых моделей, ценовая война API ⁴²⁹. Средняя уверенность: масштабы JPMorgan, экономия Klarna, проникновение Copilot ~6,5%, доля CAIO 26% → 76%, китайские 83%/45% McKinsey ⁴³⁰. С оговорками использованы buyer-reported цены (Glean, ServiceNow, W&B), тезис про 80% US-стартапов и статистика shadow AI из агрегаторов ⁴³¹. Расхождения в оценках провалов (95% / 30–50% / 74%) отражают разные определения — «нет P&L-эффекта», «заброшен после PoC», «нет существенной ценности» — и в тексте приводятся совместно ⁴³².

423Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

424Кросс-мерные инсайты исследования

425Кросс-мерные инсайты исследования

426Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

427Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

428Кросс-верификация результатов исследования

429Кросс-верификация результатов исследования

430Кросс-верификация результатов исследования

431Кросс-верификация результатов исследования

432Кросс-верификация результатов исследования

8.2 Дорожная карта по размеру бизнеса

Все три карты построены на единой логике этапов, к которой сходятся модели зрелости из разных измерений: (1) готовность — спонсор, политика доступа, базовая грамотность; (2) пилоты с метриками на 2–3 процессах; (3) редизайн процессов и масштабирование; (4) governance и оргмодель под нагрузкой; (5) агентный слой с реестром и идентичностями⁴³³⁴³⁴. Различие между размерами бизнеса — в глубине каждого этапа и в точке входа в стек.

Дорожные карты ниже учитывают инсайт 2 (архитектура доступа зависит от размера) и инсайт 3 (развилка buy vs open source по компетенциям и данным). Для всех размеров общий принцип — итеративность: пилот с метрикой → редизайн процесса → масштабирование → governance, а не «большой взрыв»⁴³⁵.

Малый бизнес (<50 сотрудников; текущее внедрение ~18% по строгим замерам): 1. Начать с ИИ, уже встроенного в оплаченный SaaS (74% SMB идут этим путём) — нулевые интеграционные затраты⁴³⁶. 2. Добавить 1–2 подписки уровня ChatGPT Business (\$20) или Copilot Business (\$18–21) на ключевых сотрудников⁴³⁷. 3. Ввести минимальную письменную политику: какие данные нельзя вносить в промпты (77% сотрудников делают это без политики)⁴³⁸. 4. Измерять результат деньгами (72% руководителей уже формально измеряют ROI; трое из четырёх видят положительную отдачу — Wharton/GBK 2025)⁴³⁹. 5. Не строить собственный стек: средний бюджет SMB ~\$18K/год не покрывает интеграцию open source⁴⁴⁰.

Горизонт: 3–6 месяцев до первых измеримых эффектов; ключевой риск — переплата за неиспользуемые места и утечки через личные аккаунты (shadow AI).

Средний бизнес (50–499; внедрение 42–64%): 1. Назначить ответственного за ИИ (роль AI PM — ключевой коррелят успеха) и спонсора на уровне CEO⁴⁴¹. 2. Выбрать 2–3 процесса с измеримой ценностью и провести редизайн процесса, а не «наложение» ИИ

433Кросс-мерные инсайты исследования

434Кросс-верификация результатов исследования

435Кросс-мерные инсайты исследования

436Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

437Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

438Кросс-мерные инсайты исследования

439Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

440Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

441Кросс-мерные инсайты исследования

(правило 10-20-70)⁴⁴²⁴⁴³. 3. Купить платформенные подписки для массовых сценариев; для кастомных — low-code (Dify self-hosted, с проверкой лицензии) и pgvector в существующем PostgreSQL вместо отдельной векторной БД⁴⁴⁴. 4. Легализовать shadow AI: предоставить управляемый доступ вместо запретов⁴⁴⁵. 5. Пилотировать агентов только на процессах с human-in-the-loop; прототип на CrewAI, продакшн — на LangGraph⁴⁴⁶.

Горизонт: 6–12 месяцев до масштабирования 2–3 процессов; ключевой риск — застревание в пилотах (именно на этом этапе проваливаются 60–95% инициатив) и лицензионные сюрпризы open source компонентов⁴⁴⁷⁴⁴⁸.

Крупный бизнес (1000+; внедрение 83%): 1. Оргмодель hub-and-spoke: центр компетенций + встроенные команды в функциях (Deloitte: +73% по метрикам, IBM: +36% ROI)⁴⁴⁹. 2. Гибридный стек: закрытые frontier-модели через Azure AI Foundry / Bedrock / Vertex для сложных задач + self-hosted Qwen/DeepSeek/Mistral на vLLM для массовых и чувствительных сценариев; обязательные LLM-шлюз (LiteLLM) и наблюдаемость (Langfuse или Datadog/Arize)⁴⁵⁰⁴⁵¹. 3. Governance по enterprise-тарифам: кастомный RBAC, аудит, agent identity; техническое принуждение политик вместо декларативных запретов⁴⁵². 4. Агентная архитектура с централизованным реестром агентов — во избежание agent sprawl по образцу Walmart⁴⁵³. 5. Переговорная позиция по лицензированию: рынок не сошёлся на единой модели (per-seat / потоковая / кредиты / flat fee), условия гибки⁴⁵⁴.

Горизонт: 12–24 месяца до системного масштабирования; ключевые риски — agent sprawl, непредсказуемые usage-счета (кейс Datadog ingestion-premium) и зависимость от roadmap

442Кросс-мерные инсайты исследования

443Кросс-верификация результатов исследования

444Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

445Кросс-мерные инсайты исследования

446Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

447Кросс-мерные инсайты исследования

448Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

449Кросс-верификация результатов исследования

450Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

451Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

452Кросс-мерные инсайты исследования

453Кросс-мерные инсайты исследования

454Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

вендоров как коммерческих, так и open source ⁴⁵⁵⁴⁵⁶⁴⁵⁷.

8.3 Типовые ошибки

Анализ провалов (Gartner, MIT NANDA, BCG) и антикейсов (Samsung, Walmart, модель \$2/диалог у Salesforce) позволяет сформулировать повторяющиеся ошибки, избежать которые дешевле, чем исправить ⁴⁵⁸⁴⁵⁹⁴⁶⁰. Ошибки ниже повторяются независимо от размера бизнеса и страны.

1. **Начинать с выбора модели вместо редизайна процесса** — 60–95% провалов имеют организационную, а не технологическую природу ⁴⁶¹.
2. **Строить in-house там, где можно купить**: buy успешен в ~67% случаев против ~33% build ⁴⁶².
3. **Игнорировать лицензии open source**: n8n (SUL), Dify (мультиитенантность/брендинг), Llama Community License ≠ OSI — юридическая проверка обязательна ⁴⁶³.
4. **Разворачивать отдельную векторную БД на mid-scale**: до ~10 млн векторов достаточно pgvector ⁴⁶⁴.
5. **Запрещать shadow AI вместо легализации** — запреты проваливаются (Samsung) ⁴⁶⁵.
6. **Agent sprawl**: запускать агентов без реестра, оркестрации и идентичностей — путь Walmart, закончившийся консолидацией; Gartner прогнозирует отмену >40% агентных проектов к 2027 ⁴⁶⁶.
7. **Недооценивать полную стоимость подписок**: место M365 Copilot — это \$69–90 all-in, а не \$30; Claude Enterprise тарифицирует токены поверх seat fee ⁴⁶⁷.

455Кросс-мерные инсайты исследования

456Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

457Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

458Кросс-верификация результатов исследования

459Кросс-мерные инсайты исследования

460Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

461Кросс-мерные инсайты исследования

462Кросс-верификация результатов исследования

463Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

464Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

465Кросс-мерные инсайты исследования

466Кросс-мерные инсайты исследования

467Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock,

8. **Закупать места «впрок» без замера использования:** при еженедельной активности 20–30% купленных мест Copilot волновое развёртывание экономит до 70% бюджета подписки ⁴⁶⁸.
9. **Файнтюнить там, где достаточно prompt engineering:** enterprise-практика сместилась к prompt engineering и роутингу между моделями; файнтюнинг оправдан точно (a16z) ⁴⁶⁹.
10. **Строить отдельную ИИ-инфраструктуру вместо использования доверенных платформ:** 56% инфраструктурных расходов остаются у инкумбентов — предприятия осознанно строят ИИ поверх уже проверенных Databricks, Snowflake, MongoDB, Datadog ⁴⁷⁰.
11. **Игнорировать M&A-риски open source зависимостей:** приобретения Langfuse (ClickHouse), Langflow (DataStax/IBM) и Portkey (Palo Alto) показывают, что roadmap критичного OSS-компонента может измениться после сделки — нужен план форка или замены ⁴⁷¹.
12. **Отсутствие спонсорства CEO и измерений:** отсутствие измеримого P&L-эффекта — главный фактор «95% провалов» MIT NANDA; 72% успешных компаний формально измеряют ROI ^{472,473}.

8.4 Фронтир: агентный ИИ и agent identity

Завершая отчёт, необходимо зафиксировать направление, которое определит повестку 2026–2027 гг.: переход от генеративного ИИ «вопрос–ответ» к агентному ИИ, выполняющему многошаговые действия в корпоративных системах. Для бизнеса это качественный сдвиг: ценность (автоматизация процессов целиком) и риски (действия с необратимыми последствиями, распоряжение данными и деньгами) возрастают одновременно, поэтому governance должен опережать развёртывание, а не догонять его.

Агентный ИИ — точка конвергенции всех трендов отчёта: продакшн-стандарт LangGraph 1.0 (Uber, LinkedIn, Klarna) ⁴⁷⁴, потребительские модели оплаты «за действие агента»

Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

468Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

469Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

470Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

471Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

472Кросс-мерные инсайты исследования

473Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

474Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph,

(Agentforce Flex Credits, ServiceNow assists, UiPath Agent Units) ⁴⁷⁵, «год агентов» в Китае (WorkBuddy, цель DingTalk — 10 млн агентов, 1,3 млн агентов на Qianfan) ⁴⁷⁶. Одновременно это главный риск-зон: Gartner прогнозирует отмену >40% агентных проектов к 2027 г., а Walmart уже прошёл цикл «sprawl → консолидация в 4 супер-агента» ⁴⁷⁷. Новая категория governance — **agent identity**: идентичность, права и аудит каждого агента как полноценного «цифрового сотрудника» становится обязательным элементом enterprise-архитектуры доступа наряду с человеческими RBAC ⁴⁷⁸.

Корпоративные прецеденты фронта подтверждают обе стороны медали. JPMorgan инвестирует ~\$2 млрд в ИИ и развернул LLM Suite для ~250 тыс. сотрудников — пример централизованного доступа вместо распыления ⁴⁷⁹. Klarna зафиксировала \$60 млн экономии на ИИ, но вернула людей в гибридную модель — напоминание, что полная автоматизация без human-in-the-loop деградирует по качеству ⁴⁸⁰. Наконец, организационный слой: доля организаций с ролью Chief AI Officer выросла с 26% (2025) до 76% (2026, IBM), а ISO/IEC 42001 принят как европейский стандарт менеджмента ИИ (март 2026) — governance агентов и моделей формализуется на уровне стандартов ⁴⁸¹.

Приоритеты для агентных инициатив 2026: (1) начинать с одного сквозного агента на процессе с чёткой метрикой, а не с платформы «агентов для всего»; (2) вести централизованный реестр агентов с владельцем, score и идентичностью (agent identity) у каждого; (3) проектировать human-in-the-loop для действий с финансовыми и юридическими последствиями (LangGraph поддерживает это нативно) ⁴⁸²; (4) заранее оценивать потребительскую модель оплаты за действия агента — break-even у Agentforce Flex Credits наступает на 20 действиях на диалог ⁴⁸³; (5) относиться к прогнозу Gartner (>40% отмен к 2027) как к сигналу дисциплинирующему: отменять слабые агентные проекты рано и без сожаления ⁴⁸⁴.

Среднесрочный прогноз по стеку: слой агентных идентичностей (agent identity) будет встроен в IAM-платформы и enterprise-тарифы LLM-продуктов, а «assists/credits/действия»

Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

475Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

476Свод источников по китайскому рынку ИИ (CNNIC, MIIT, CAC, CAICT, McKinsey Greater China, IDC, Omdia, QuestMobile; Alibaba/DingTalk, Tencent/WeCom, Baidu, ByteDance/Volcano Engine, Huawei и др.)

477Кросс-мерные инсайты исследования

478Кросс-мерные инсайты исследования

479Кросс-верификация результатов исследования

480Кросс-верификация результатов исследования

481Кросс-верификация результатов исследования

482Свод источников по open source стеку ИИ (MLflow, Airflow, Kubeflow, DVC, LangChain/LangGraph, Langfuse, LiteLLM, Dify, pgvector/Qdrant/Milvus, Ollama/llama.cpp/vLLM, Qwen/Llama/DeepSeek/Mistral и др.)

483Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

484Кросс-мерные инсайты исследования

— потребительские модели оплаты — вытеснят простые per-seat подписки в сценариях, где ценность измеряется выполненными действиями, а не доступом ⁴⁸⁵⁴⁸⁶. Компаниям, проектирующим агентную архитектуру сегодня, следует закладывать реестр агентов, разделение прав и аудит с первого дня — модернизация задним числом (кейс Walmart) обходится дороже ⁴⁸⁷.

Итоговая формула 2024–2026: технология перестала быть ограничением — ограничением стали организация, управление доступом и дисциплина перехода от пилота к масштабу. Компании, которые инвестируют 70% усилий в людей и процессы, выбирают путь (buy vs open source) под свои компетенции и требования к данным, легализуют shadow AI и вводят агентов под централизованным реестром, системно попадают в меньшинство, извлекающее из ИИ измеримую ценность ⁴⁸⁸⁴⁸⁹.

Практический первый шаг не зависит от размера бизнеса: инвентаризировать текущее (включая теневое) использование ИИ, назначить владельца программы и выбрать один процесс с чёткой денежной метрикой для первого управляемого пилота. Всё остальное — стек, платформа, модель — вторично и выводится из этого выбора ⁴⁹⁰.

Источники

485Кросс-мерные инсайты исследования

486Свод источников по коммерческим платформам и ценообразованию (Azure AI Foundry, AWS Bedrock, Google Vertex AI, M365 Copilot, ChatGPT, Claude, Glean, Agentforce, ServiceNow, UiPath, W&B, Arize, Datadog; Menlo Ventures, a16z и др.)

487Кросс-мерные инсайты исследования

488Кросс-мерные инсайты исследования

489Кросс-верификация результатов исследования

490Кросс-мерные инсайты исследования

