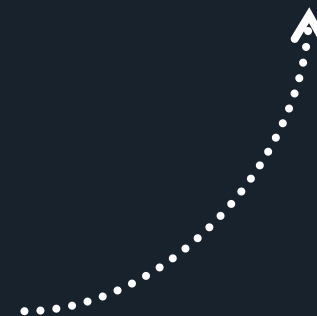


использование нейрокоммуникации в отраслевых задачах

или тренды и перспективы искусственного интеллекта и
нейросетей в России

Тимур Бикбов: аспирант УГНТУ,
был консультантом Deloitte,
получил гранты на 2 млн руб.,
создал ИИ-стартап



**AIR: развиваем маркетплейс нейросетей, создаем
корпоративные нейронки на заказ. Более 300 000
пользователей, контракты
с компаниями из РБК-100**

**Экономический потенциал ИИ в России
к 2028 г. составит 22–36 трлн руб., эффект
к 2028 г. может достичь 4,2–6,9 трлн руб.,
что эквивалентно влиянию на ВВП до 4%. Вклад
генеративного ИИ – 20%**

**Около 70% потенциала ИИ приходится
на шесть отраслей: логистика, банкинг, ритейл,
добыча ископаемых, потребительские товары, ИТ**

**Российский рынок решений на базе ИИ
в 2022 г. оценивается в 30–50 млрд руб.,
к 2028 г. эта цифра может вырасти до 0,3–0,6 трлн руб.
(CAGR – около 50%).**

**Развитие ИИ открывает новые возможности на
рынке труда:**

**повышение продуктивности сотрудников,
высвобождение времени от рутинных задач,
появление новых профессий.**

**Самые популярные бизнес-кейсы применения
генеративного ИИ:
маркетинг и продажи, клиентский сервис,
ИТ и внутрикорпоративные функции.**

**63% лидеров российского бизнеса
ожидают, что развитие искусственного интеллекта
повлияет на их компании
уже на горизонте 1 года**

80%

процессов продуктовой разработки будет предполагать использование генеративного ИИ к 2025 г.

80%

организаций, осуществляющих клиентскую поддержку, внедрят генеративный ИИ для роста продуктивности специалистов поддержки и улучшения клиентского опыта к 2025 г.

60%

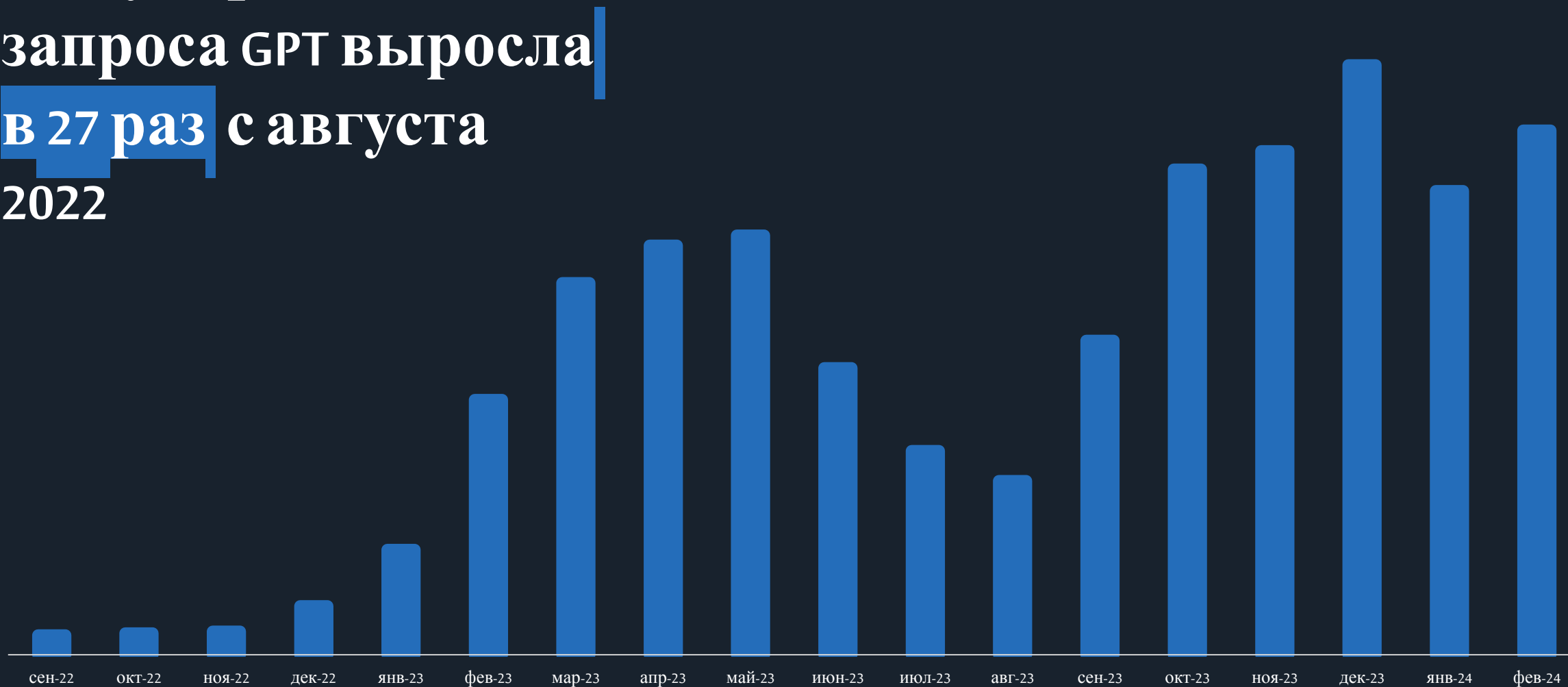
B2B-продаж будет осуществляться при помощи генеративного ИИ к 2028 г.

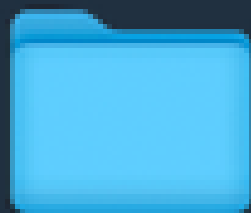
КТО ТАКОЙ «ИИ»

В понятие ИИ входят разные направления:
компьютерное зрение, обработка естественного
языка, поддержка принятия решений,
распознавание речи и т.д.

Сегодня мы будем говорить об ИИ
в контексте **больших языковых моделей**,
а также моделей диффузии, так как они
демократизировали всю индустрию

Популярность
запроса GPT выросла
в 27 раз с августа
2022



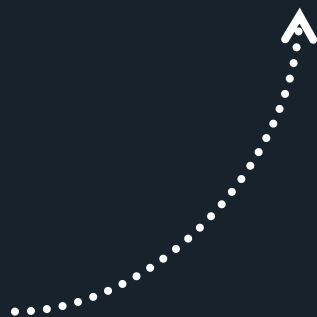


большой файл
с параметрами



маленький файл
для запуска

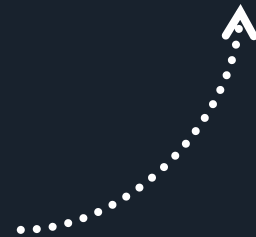
Это нейросеть



Делаем **нейросеть** своими руками:

- 1️⃣ обучаем базовую модель
- 2️⃣ дообучаем модель
- 3️⃣ улучшаем результаты

А это нормальное объяснение
от сооснователя ChatGPT,
разработчика Андрея Карпаты





ChatGPT



Claude



GigaChat



Vicuna



Cohere



LLaMa *



Шедеврум



Midjourney



DALLE



Stable Diffusion



Kandinsky

*использован логотип компании Meta.

Meta признана экстремистской организацией в России. Осуждаем.

Тут можно скачать
сотни **open source** моделей и
арендовать для них
вычислительные мощности в
облаке ☺



Replicate



Hugging Face

Предиктивные модели. Они используются в финансовой функции для выявления мошеннических транзакций, в операционной функции для предиктивной аналитики ремонтов, в коммерции для предсказания спроса.

Оптимизационные модели. Используются в медицине для нахождения лучшего способа проведения операций, в логистике – для оптимизации маршрутов, в закупках – для сокращения расходов.

**Компьютерное зрение. Решения для
распознавания объектов реального мира, к ним
относится распознавание лиц камерой смартфона,
сканирование штрихкодов в магазинах,
распознавание номеров машин камерами на
дорогах, видеоаналитика.**

Обработка естественного языка (NLP).
Технология, позволяющая компьютерным
системам анализировать человеческий язык.
Используется, например, в чат-ботах клиентской
поддержки, в поисковых системах при поиске
текстов.

Речевые технологии. Позволяют работать с человеческой речью, состоят из блоков распознавания и синтеза реч. Зачастую работают в связке с NLP, используются в голосовых ассистентах, таких как Siri.

Рекомендации и персонализация

Эти технологии позволяют сделать пользователю наилучшее предложение исходя из его характеристик, в том числе на основе кластеризации. Используются в маркетинге, рекламе и продажах, преимущественно в цифровых отраслях.

Большие языковые модели (LLM). Применяются для генерации, суммаризации или изменения текста. Спектр задач, для которых могут применяться модели, очень широк, и на сегодняшний день до конца не известны все возможные сценарии использования.

1950–1960 гг.

Появление первых искусственных нейронных сетей

1950 г.

SNARC – первая в мире работающая искусственная нейросеть

1954 г.

Разработка первых программ для игры в шашки и шахматы

1958 г.

Mark I Perceptron – Фрэнк Розенблатт впервые представил алгоритм перцептрона (первой в мире нейронной сети)

1960–1970 гг.

Первые практические применения ИИ

1964 г.

Eliza – первый чат-бот, разработанный исследователем MIT Джозефом Вайзенбаумом и применявшийся для психотерапии. Бот использовал ключевые слова из диалога для выдачи заранее подготовленных фраз. Начало исследований в области NLP

1970–1990 гг.

«Зима» ИИ (отсутствие интереса к технологии из-за неоправданных ожиданий)

1970 г.

Финский математик Сеппо Линнаинмаа впервые описал метод обратного распространения ошибки – одну из ключевых составляющих, сделавшую возможным эффективное обучение нейронной сети

1982 г.

Появление рекуррентных нейронных сетей; они учитывали предварительную информацию, что в будущем сделало возможным моделировать последовательные данные, такие как естественный язык

1986 г.

Появление архитектуры Autoencoder, позволяющей при обучении без учителя находить в неструктурированных данных общие признаки

1990–2010 гг.

Появление моделей и технологий, способных ускорить развитие ИИ

1995 г.

LeNet-5 – первая сверточная нейросеть (Convolutional Neural Network, CNN), позволяющая распознавать изображения

Владимир Вапник представил современный вид SVM (Support Vector Machine) – модель для решения задач классификации и регрессии

1997 г.

Создание архитектуры сетей LSTM (Long Short-term Memory – долгая краткосрочная память) немецкими учеными Зеппом Хохрайтером и Юргеном Шмидхубером. Эта архитектура позволила сетям учитывать долгосрочные зависимости в данных

Компьютерный алгоритм IBM Deep Blue победил чемпиона мира по шахматам Гарри Каспарова

1999 г.

Джером Фридман создал градиентный бустинг – метод, при котором в процессе обучения модели последовательно минимизируют ошибки друг друга

2010–2020 гг.

Реализация потенциала ИИ на основе глубокого обучения

2011 г.

Siri от Apple – первый популярный голосовой помощник, использующий технологии распознавания и синтеза речи, обработки естественного языка для взаимодействия с пользователем

2012 г.

Нейросеть AlexNet на основе технологии CNN победила на конкурсе распознавания изображений, положив начало активному применению глубокого обучения

2013 г.

Word2vec – технология NLP, представляющая слова в виде векторов, отражающих значения слов, что позволяет при поиске учитывать синонимы

2014 г.

Ян Гудфеллоу описал GAN (генеративно-состязательные сети) – такие модели одними из первых стали использоваться для генерации реалистичных изображений

Ноябрь

Amazon запустил своего виртуального помощника – Amazon Alexa

2015 г.

Появление ResNet (Residual Neural Network) – глубокие сверточные нейронные сети с остаточной структурой обучения превосходили человеческий уровень классификации

Декабрь

Основание OpenAI – некоммерческой open-source исследовательской лаборатории для продвижения технологий ИИ (создатели – Илон Маск, Сэм Альтман, Грег Брокман, Илья Суцкевер, Войцех Заремба, Джон Шульман и другие)

2016 г.

AlphaGo – компьютерная программа для игры в го, разработанная Google DeepMind, стала первой программой, победившей профессионального игрока – чемпиона по го Ли Седоля

2017 г.

Исследователи из Google опубликовали статью «Внимание – это все, что вам нужно» (Attention Is All You Need), в которой представляется концепция моделей-трансформеров – прорывной технологии, лежащей в основе ChatGPT

2018 г.

Алгоритм AlphaZero самообучился игре в шахматы и го на высоком уровне всего лишь за 4 часа (без знания правил и принципов игры)

Октябрь – ноябрь

Google представил BERT (Bidirectional Encoder Representations from Transformers) – двунаправленную модель с 340 млн параметров

2020 г.

AlphaFold, разработанная Google DeepMind, решила проблему фолдинга (сворачивания) белка

Расцвет диффузионных моделей; они представляют собой класс вероятностных генеративных моделей, которые превращают шум в репрезентативную выборку данных. На базе диффузионных моделей работают и текущие решения для генерации изображений: DALL-E 2 (OpenAI), Imagen (Google), Stable Diffusion (Stability AI), Midjourney и YandexART (Яндекс)

2020 г. и далее

Реализация потенциала генеративного ИИ

Несмотря на то что мировую известность технология получила только в 2022 г. с появлением чат-бота ChatGPT на базе модели GPT-3.5, генеративный ИИ – закономерный результат развития технологии глубокого обучения (Deep Learning) и исследований в области обработки естественного языка (NLP), продолжавшихся с 1960-х гг.

2020 г.

Май

GPT-3 (175 млрд параметров) – языковая модель, которая генерирует текст при помощи предварительно обученного алгоритма и выполняет языковые задачи почти неотличимо от человека

2021 г.

Январь

OpenAI анонсировала DALL·E – модель ИИ, предназначенную для создания и редактирования уникальных произведений искусства и фотореалистичных изображений на основе текста

Май

Google представила «языковую модель для диалоговых приложений» (LaMDA) – крупную языковую модель, которая умеет отвечать в разговоре подобно человеку

Июль

Яндекс одним из первых в мире запустил нейросетевой перевод видео (комбинацию нескольких технологий, в частности распознавания речи, синтеза речи, биометрии и машинного перевода) в дополнение к уже существовавшему переводу картинок и текста – пользователи получили возможность смотреть видео не просто со сгенерированными субтитрами, а автоматически переведенными и озвученными (разными голосами у мужских и женских персонажей) прямо в браузере¹⁰

2022 г.

Август

Stability AI выпустила Stable Diffusion – open-source-инструмент на основе ИИ, который генерирует изображения по текстовым описаниям

Яндекс запустил потоковый голосовой перевод трансляций в Яндекс Браузере

Ноябрь

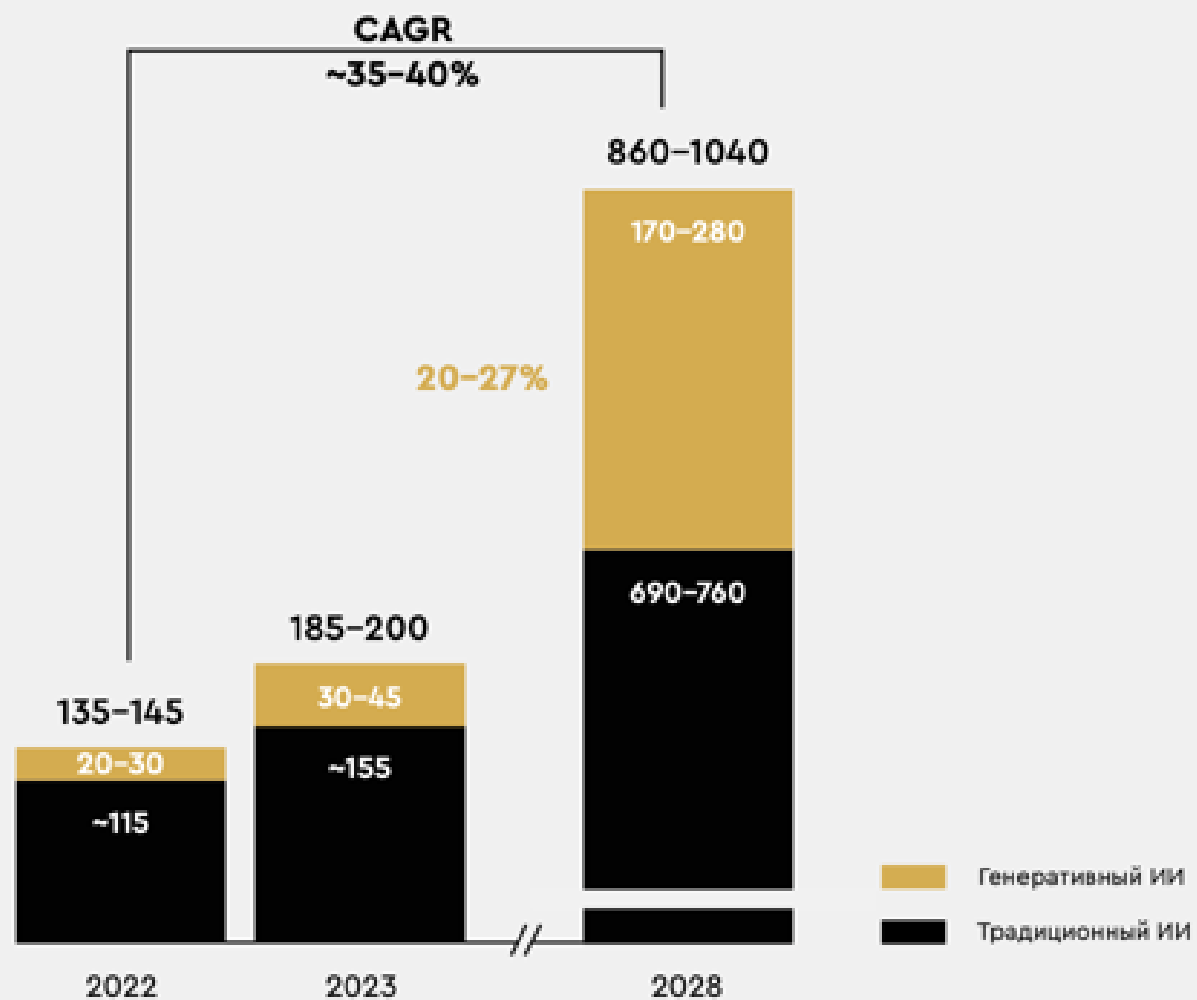
OpenAI запустил ChatGPT – практичный инструмент, который генерирует тексты, подобные человеческим ответам на вводимые запросы («напиши письмо на руководство о необходимости выделить команде еще одну неделю на завершение верстки исследования») и учитывающие контекст («статья касается генеративного ИИ и включает множество графиков»). Он становится первым активно используемым инструментом на базе технологии: 100 млн пользователей за 2 месяца

После запуска ChatGPT в большую игру по разработке собственных базовых моделей активно включились другие игроки – как в США, так и в других странах, в том числе в России

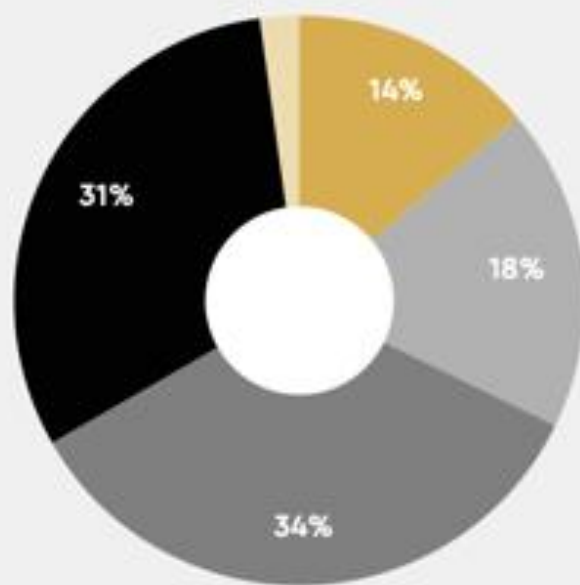
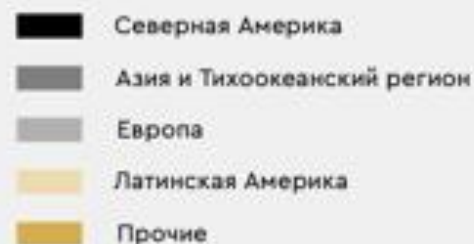
Декабрь

Cohere выпустил свою языковую модель Claude

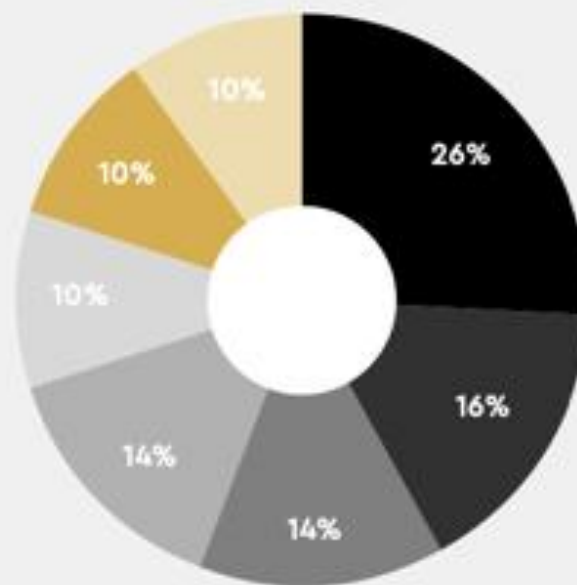
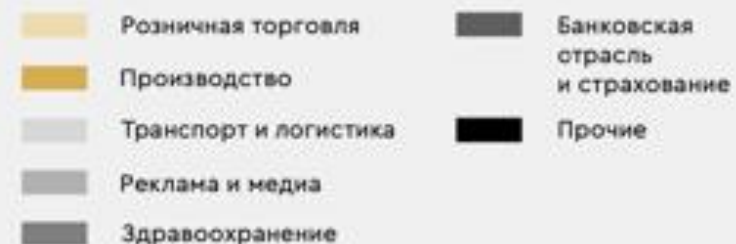
Мировой рынок решений на базе ИИ, млрд долл. США в год



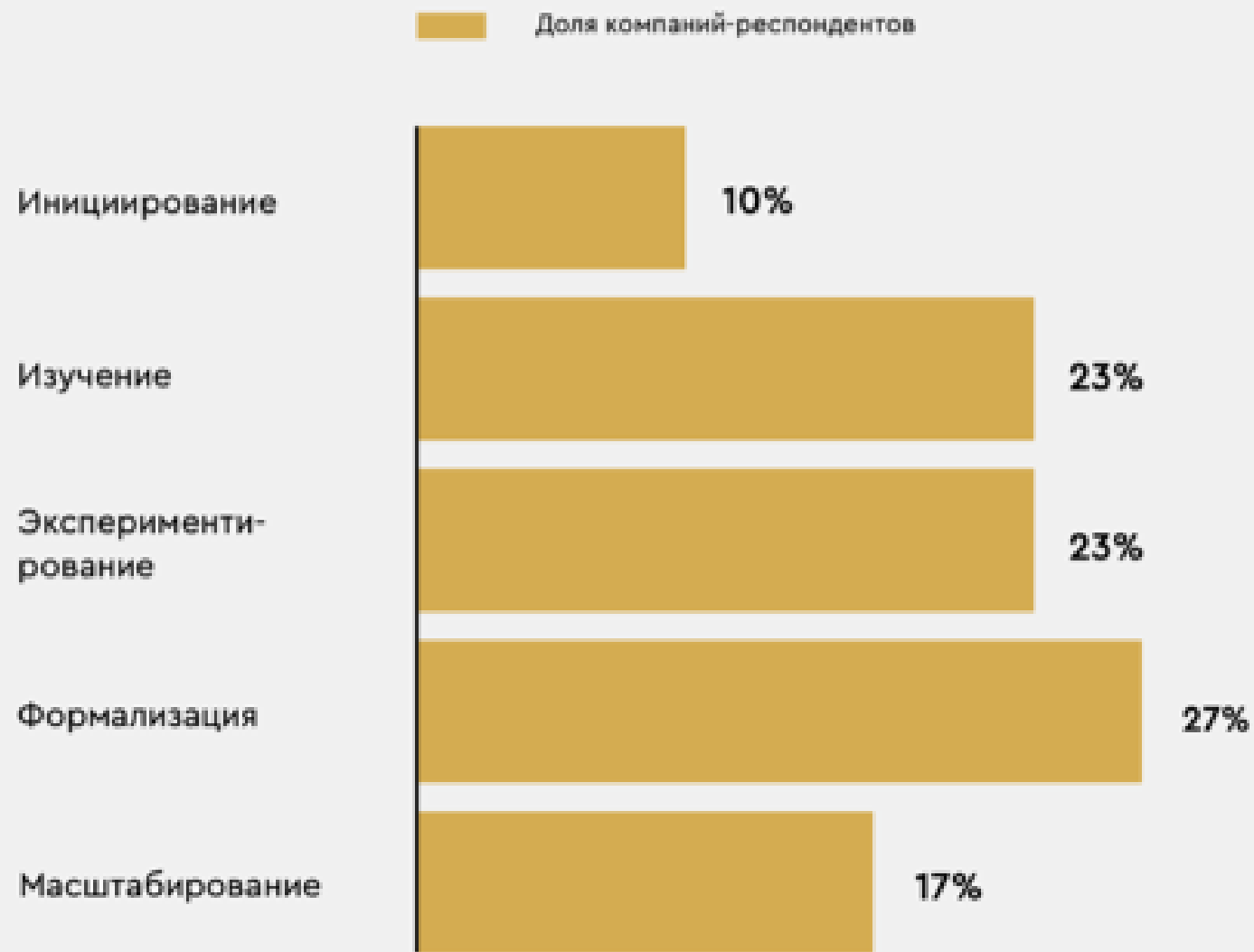
Рынок решений на базе ИИ по регионам к 2028 г.



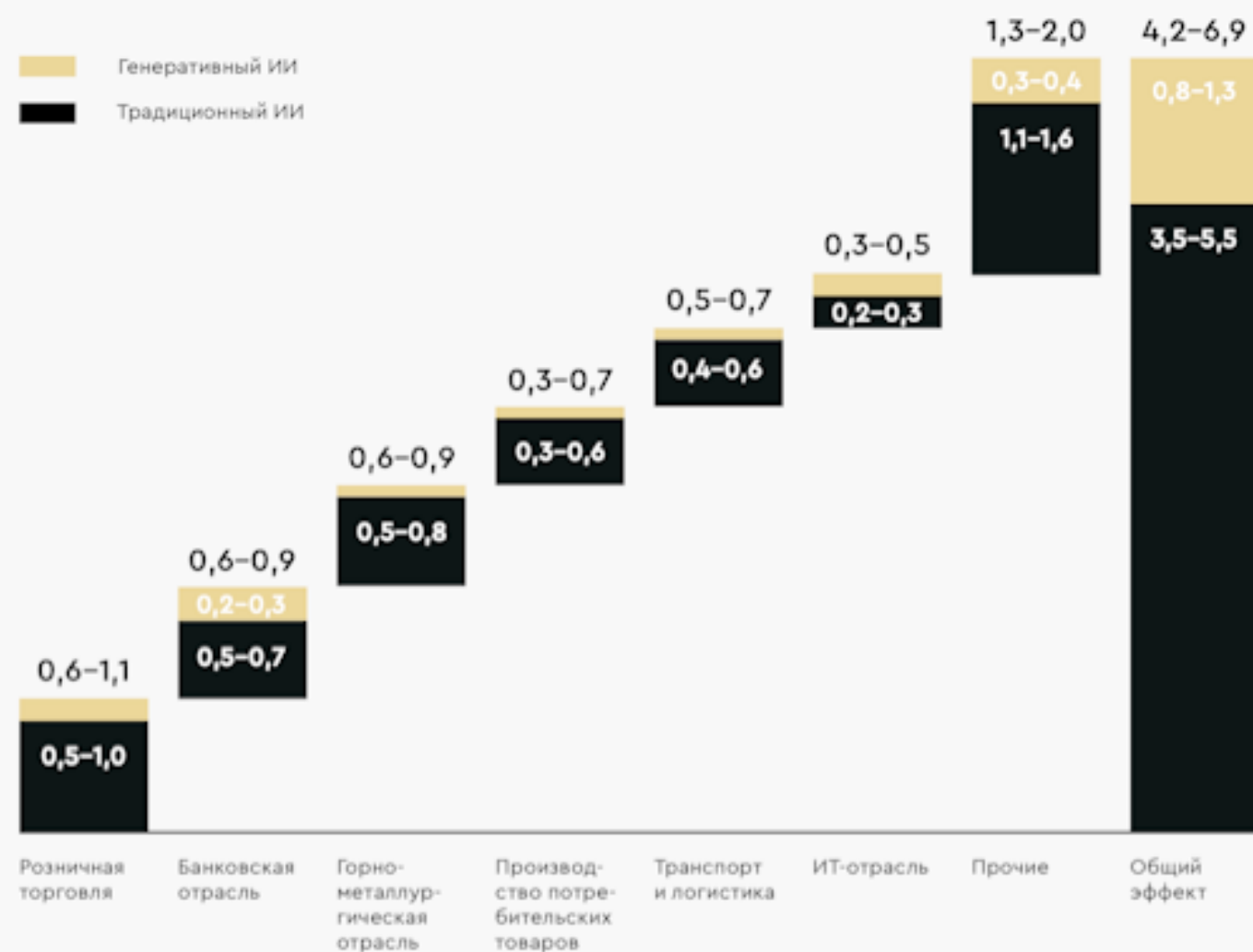
Рынок решений на базе традиционного ИИ по отраслям применения к 2028 г.



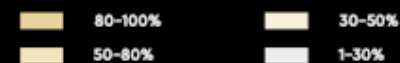
Стадии внедрения ИИ в России



Ожидаемый финансовый эффект от внедрения ИИ для экономики России, трлн руб. в год



Доля компаний, внедряющих ИИ в бизнес-функции,
от общего числа опрошенных компаний по индустриям



	Клиентский сервис	Маркетинг и продажи	Производство	Цепочки поставок	Финансы	ИТ	Исследования и разработки	Управление персоналом	Стратегия	Внутренние коммуникации
Банковская отрасль и страхование	80–100%	50–80%	30–50%	1–30%	80–100%	50–80%	30–50%	50–80%	50–80%	30–50%
ИТ и технологии	50–80%	80–100%	50–80%	50–80%	30–50%	80–100%	50–80%	30–50%	80–100%	30–50%
Розничная торговля	80–100%	80–100%	80–100%	50–80%	50–80%	30–50%	30–50%	30–50%	80–100%	30–50%
Металлы и горная добыча	30–50%	50–80%	80–100%	50–80%	50–80%	30–50%	80–100%	50–80%	30–50%	30–50%
Электроэнергетика	50–80%	30–50%	50–80%	30–50%	30–50%	30–50%	30–50%	50–80%	30–50%	80–100%
Телекоммуникации	80–100%	50–80%	80–100%	80–100%	50–80%	80–100%	80–100%	30–50%	30–50%	30–50%
Сельское хозяйство	80–100%	50–80%	80–100%	50–80%	30–50%	80–100%	50–80%	80–100%	80–100%	80–100%
Электронная коммерция	80–100%	80–100%	80–100%	50–80%	50–80%	50–80%	80–100%	80–100%	80–100%	80–100%
Нефть и газ	30–50%	30–50%	80–100%	50–80%	50–80%	50–80%	50–80%	80–100%	80–100%	80–100%
Транспорт и логистика	50–80%	50–80%	50–80%	80–100%	30–50%	30–50%	30–50%	80–100%	80–100%	80–100%
Химия и нефтехимия	30–50%	50–80%	80–100%	30–50%	30–50%	30–50%	30–50%	80–100%	80–100%	80–100%
Потребительские товары	80–100%	30–50%	30–50%	30–50%	30–50%	80–100%	30–50%	30–50%	80–100%	80–100%
Машиностроение	30–50%	80–100%	80–100%	80–100%	30–50%	30–50%	30–50%	80–100%	80–100%	80–100%
Автомобильная промышленность	50–80%	50–80%	80–100%	50–80%	80–100%	80–100%	50–80%	80–100%	80–100%	80–100%
Медиа	50–80%	50–80%	80–100%	80–100%	80–100%	30–50%	80–100%	80–100%	80–100%	80–100%
Итого	55%	52%	46%	35%	31%	29%	27%	16%	6%	1%

К 2028 г. совокупная выручка компаний от продажи ИИ-решений для B2B может вырасти до 0,3–0,6 трлн рублей.

Около 80% респондентов опроса «Яков и партнеры» используют решения YandexGPT для решения задач по работе с текстами.

Финансирование GenAI в России

76%

опрошенных компаний
тратят на генеративный ИИ
не более 1% годового
бюджета на ИТ/
цифровизацию

15%

опрошенных компаний
тратят на генеративный ИИ
от 1 до 3% годового
бюджета на ИТ/
цифровизацию

6%

опрошенных компаний
тратят на генеративный ИИ
от 3 до 5% годового
бюджета на ИТ/
цифровизацию

0%

опрошенных компаний
тратят на генеративный ИИ
от 5 до 10% годового
бюджета на ИТ/
цифровизацию

3%

опрошенных компаний
тратят на генеративный ИИ
более 10% годового
бюджета на ИТ/
цифровизацию

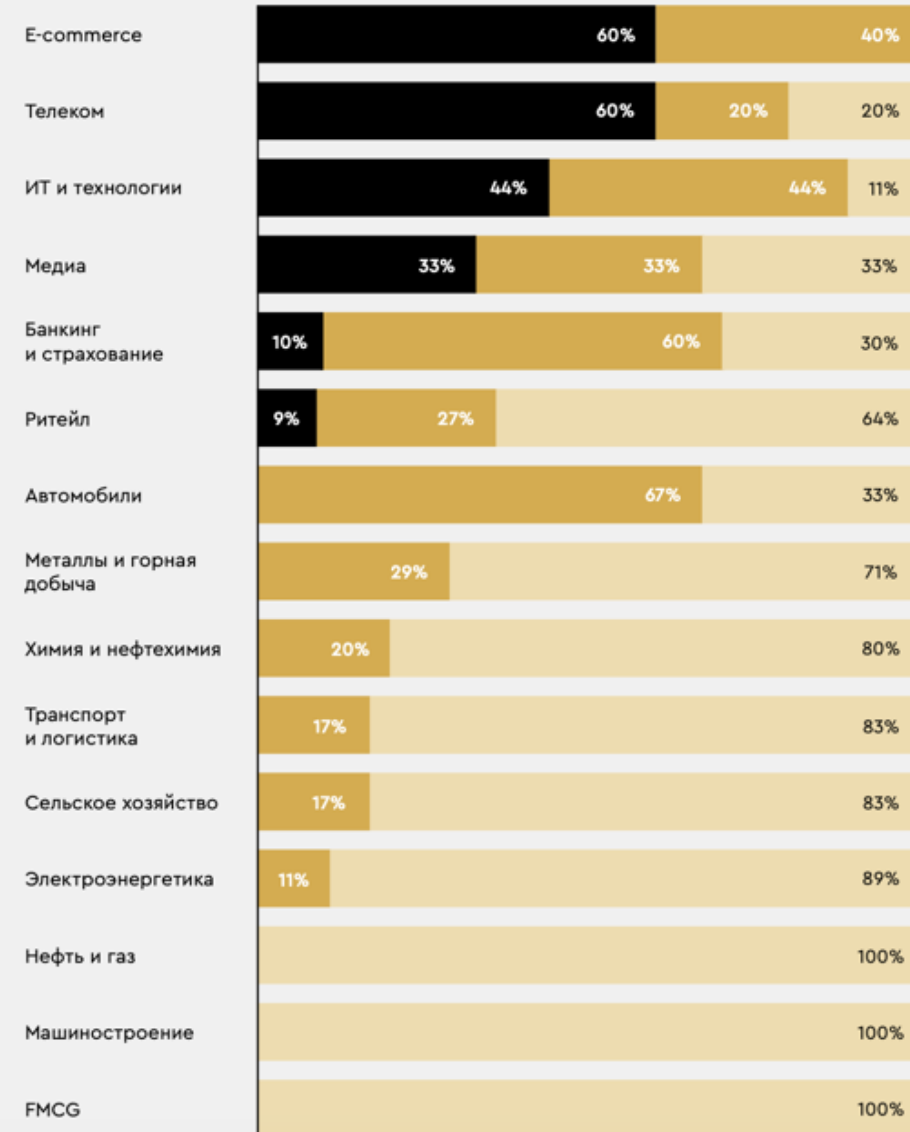
	Технология	Инфраструктура	Бизнес-продукт	Заказная разработка
Продукт	Базовая модель (Foundation Model)	Инфраструктурные решения для работы с данными и базовыми моделями	Коробочные продукты под универсальные юз-кейсы, не требующие кастомизации	Решения под ключ для узкоспециализированных юз-кейсов
Модель деплоя	Доступ через API / Amazon Bedrock и другие платформы	SaaS-платформа с инструментами (Fine Tuning, RLHF, Prompt Engineering и т. д.)	(Облачная) платформа / приложение разработчика	On-premise (собственная) инфраструктура / платформа по выбору заказчика
Примеры	 OpenAI  Inflection  cohere  AI21labs  ADEPT  ANTHROPIC	 mosaic ^{ML}  CoreWeave  aporia  FIXIE  Humanloop  scale	 algolia  runway  glean  Minty  peppercontent  Harvey.	 ML6  RisingMax  InData Labs  SLT OUA  LeewayHertz  scale

Количество юз-кейсов с использованием генеративного ИИ в ключевых функциях



Как отрасли относятся к новым технологиям GenAI?

■ Пионеры ■ Экспериментаторы ■ Последователи



Стадии внедрения ИИ в компаниях России



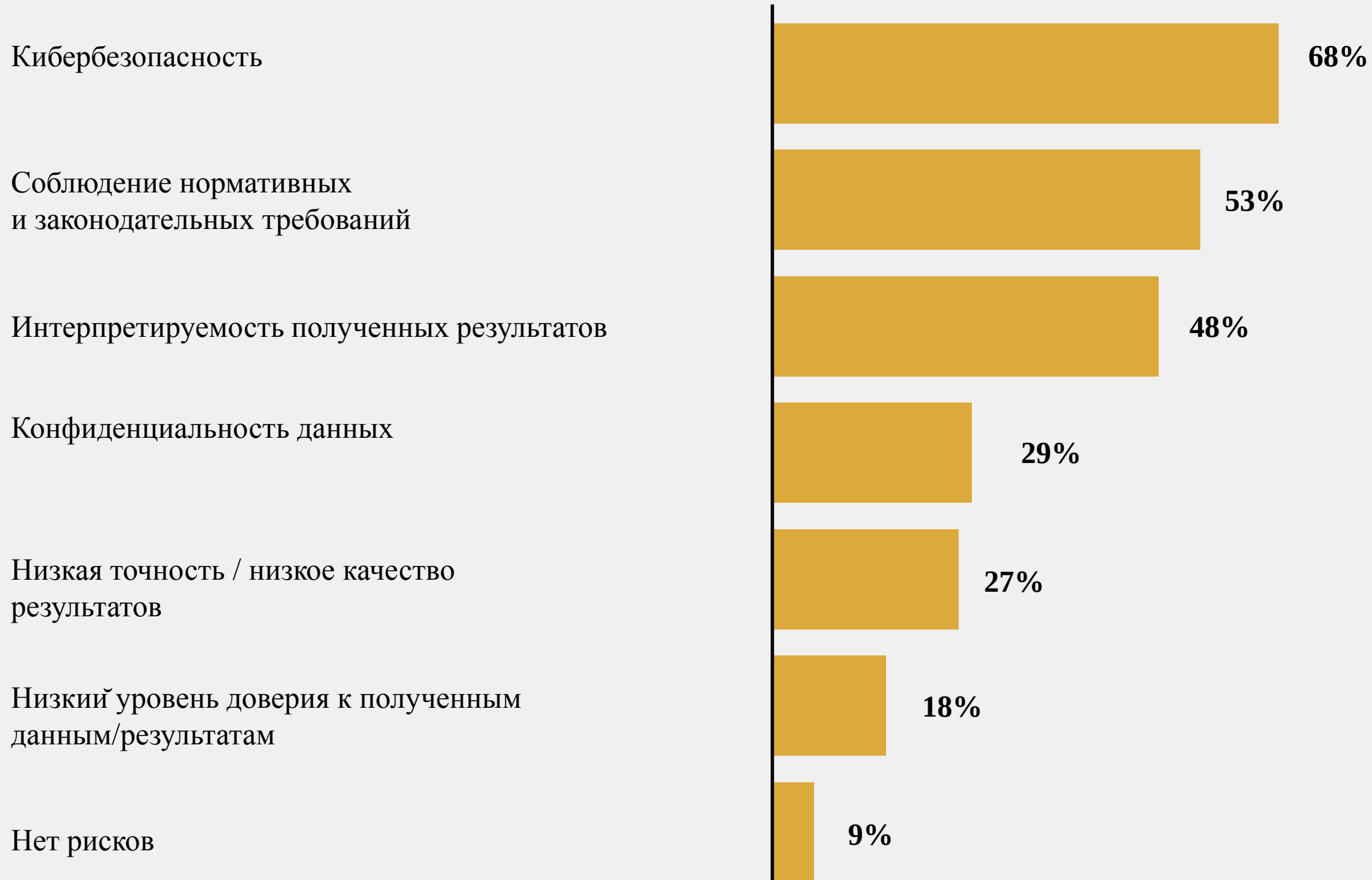
Наиболее интересные юз-кейсы: доля компаний (в %), спонтанно отметивших юз-кейс в качестве приоритетного

[illegible]

Доля опрошенных компаний (в %), которые:

Потенциально заинтересованы в пилоте	14/15	
Готовы рассматривать интеграцию через облако	9/15	
Хотят увидеть «доказанный практический эффект»	8/15	
Готовы сформировать рабочую группу для выбора приоритетных кейсов	8/15	
Работают над лонг-листом юз-кейсов	6/15	

Функции	Ассистент написания кода	Продажи и поддержка	Общекорпоративные функции и управление персоналом
Продуктовое наполнение	Завершение 1–2 строк кода (inline-подсказки)	Генеративные подсказки оператору поддержки	Умный корпоративный портал: внутренний поисковик, систематизация, суммаризация материалов в базе знаний
	Генерация кода по сигнатуре функции / контексту файла	Автоматические ответы пользователю	Генерация тренингов на основе внутренних/внешних материалов («помощник методиста»)
	Написание кода по текстовому комментарию	Суммаризация обращений	Анализ интервью и формирование рекомендации по кандидату (массовый наем)
	Написание документации к коду	Контроль качества	Суммаризация обратной связи по кандидату для передачи следующему интервьюеру, заполнение ATS (точечный наем)
	Проверка кода	Тренажер для операторов с персонализированным фидбэком	
	Оптимизация/рефакторинг кода		
	Создание тестов для проверки качества		
Потенциальный эффект	–10–15%	–30–40%	5–10%
	Сокращение рабочего времени разработчиков за счет ускорения:	Сокращение АНТ за счет экономии времени на:	Сокращение затрат на функцию управления персоналом за счет:
	Написания кода	Чтение и осмысление истории взаимодействия	Генерации готовых тренингов
	Написания тестов	Поиск макросов	Написания саммари интервью (точечный наем)
	Рефакторинга кода	Написание/редактирование ответа	Анализа интервью (массовый наем)
	Создания документации		
	+0,1–0,2	+0,1–0,2	x5–7
	Прирост eNPS за счет сокращения объема рутинных задач	Прирост CSAT за счет создания персонализированных ответов пользователям	Увеличение скорости поиска информации внутри корпоративных баз знаний



Поиск и привлечение специалистов с ИИ-компетенциями

99%

Поиск решений/инструментов на основе ИИ на рынке или их разработка

78%

Высокая финансовая стоимость

57%

Оценка потенциального эффекта от использования ИИ и его влияния на КПЭ

51%

Создание необходимой платформы данных для работы с ИИ

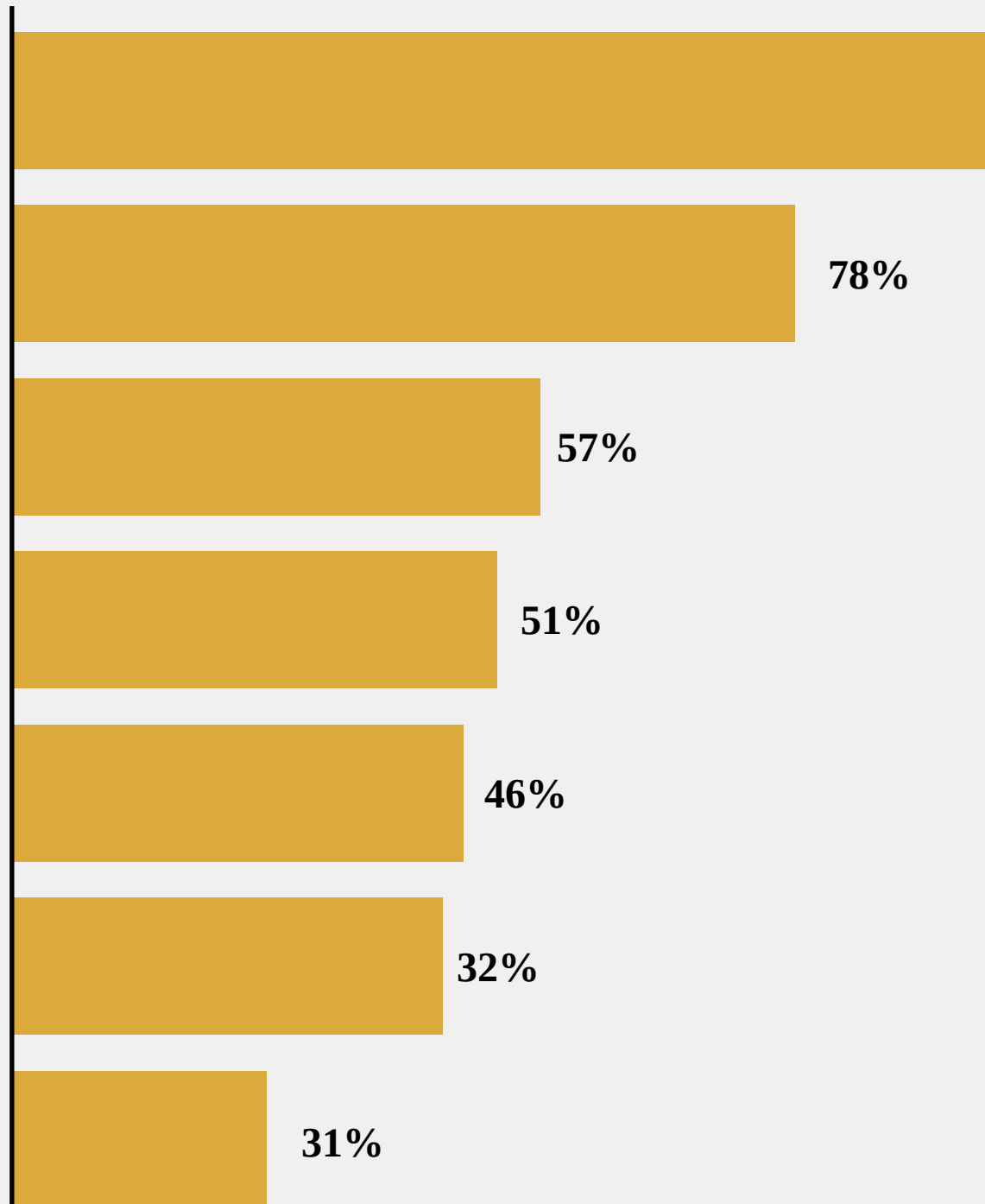
46%

Изменение привычных бизнес-процессов для эффективного использования ИИ

32%

Масштабирование ИИ внутри организации

31%



Плохой промпт	Хороший промпт	Комментарии
Напиши что-нибудь про банки	Напиши вступление к научной статье о банковских системах в разных странах	Указали стиль и уточнили контекст
Перескажи данную статью	Сократи эту статью до 250 слов	Указали желаемый размер текста
Как стать продуктивным в учебе?	Назови десять лучших советов по повышению продуктивности для студентов технических направлений	Указали желаемую структуру ответа и уточнили про учебу
Что такое искусственный интеллект?	Дай определение, виды и примеры применения искусственного интеллекта	Вместо абстрактной формулировки декомпозировали на подзадачи
Суммаризируй видео	Твоя задача – суммаризировать видео. Сделай его короткий пересказ, разбив ровно на три главы. Дай каждой главе название и короткий пересказ в виде тезисов – так чтобы прочтение заняло не более 30 секунд. В конце каждой главы оставь тайминг ее начала	Уточнили формат, описали необходимый подход словами

Концепция уровней регулирования ИИ

1 уровень

Общее регулирование

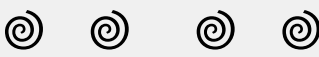
основанное на соблюдении правил, норм, этики

2 уровень

Дополнительные меры

Возможные архитипы

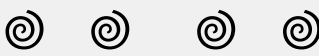
Отсутствие
дополнительного
регулирования



Ограничения в
отдельных отраслях
и саморегуляция



Ограничения
в зависимости
от риск-профиля



Учет алгоритмов ИИ
и проверка контента
для обучения



Для более чем

60%

сотрудников

развитие генеративного ИИ позволит высвободить
время за счет автоматизации рутинных задач